LÍLIAN MARIA DE OLIVEIRA

CLASSIFICAÇÃO DE DADOS SENSORIAIS DE CAFÉS ESPECIAIS COM RESPOSTA MULTICLASSE VIA ALGORITMO BOOSTING E BAGGING

LAVRAS - MG

2016

LÍLIAN MARIA DE OLIVEIRA

CLASSIFICAÇÃO DE DADOS SENSORIAIS DE CAFÉS ESPECIAIS COM RESPOSTA MULTICLASSE VIA ALGORITMO BOOSTING E BAGGING

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

Orientador
Dr. Fortunato Silva de Menezes

Coorientador
Dr. Marcelo Ângelo Cirillo

LAVRAS - MG
2016

Ficha catalográfica elaborada pelo Sistema de Geração de Ficha Catalográfica da Biblioteca Universitária da UFLA, com dados informados pelo(a) próprio(a) autor(a).

Oliveira, Lílian Maria de.
Classificação de dados sensoriais de cafés especiais com resposta multiclasse via Algoritmo Boosting e Bagging / Lílian Maria de Oliveira. – Lavras : UFLA, 2016.
85 p. : il.

Dissertação (mestrado acadêmico)–Universidade Federal de Lavras, 2016.
Orientador(a): Fortunato Silva de Menezes.
Bibliografia.

1. Métodos de classificação. 2. Qualidade de cafés. 3. Análise Discriminante. I. Universidade Federal de Lavras. II. Título.

LÍLIAN MARIA DE OLIVEIRA

CLASSIFICAÇÃO DE DADOS SENSORIAIS DE CAFÉS ESPECIAIS COM RESPOSTA MULTICLASSE VIA ALGORITMO BOOSTING E BAGGING

Dissertação apresentada à Universidade Federal de Lavras, como parte das exigências do Programa de Pós-graduação em Estatística e Experimentação Agropecuária, área de concentração em Estatística e Experimentação Agropecuária, para a obtenção do título de Mestre.

APROVADA em 25 de fevereiro de 2016.

Dr. Carla Regina Guimarães Brighenti	UFSJ
Dr. João Domingos Scalon	UFLA
Dr. Marcelo Ângelo Cirillo	UFLA

Orientador
Dr. Fortunato Silva de Menezes

Coorientador
Dr. Marcelo Ângelo Cirillo

LAVRAS - MG
2016

Aos meus pais com todo meu amor e gratidão, por tudo que fizeram por mim ao longo da minha vida. Espero ser merecedora do esforço, incentivo e dedicação quanto à minha formação. Amo vocês!

À minha irmã por sempre estar ao meu lado nas decisões mais difíceis.

Aos meus amigos.

DEDICO.

AGRADECIMENTOS

A Deus, por me guiar em todos os caminhos, iluminando os meus passos e me presenteando com uma família que me deu todo suporte para vencer as barreiras da distância.

Aos meus pais por serem a base da minha vida. Se há algo que fez toda a diferença na minha formação e na vida, é o amor que recebi de vocês.

À minha irmã, pela atenção, companheirismo, amizade e por me convencer que sempre haverá uma porta destrancada só esperando por mim para abri-la.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pela concessão da bolsa de estudos.

Aos professores do Programa de Pós-Graduação em Estatística e Experimentação Agropecuária. Em especial ao professor Renato Ribeiro de Lima por não me deixar desistir nos momentos mais difíceis e possibilitar a minha formação.

Às secretárias Nádia e Josi, por serem sempre prestativas, atenciosas, competentes, dedicadas e eficientes.

Ao Professor Tiago Martins Pereira, que desde a graduação sempre foi prestativo e dedicado.

Aos meus amigos de Ouro Preto que mesmo de longe estavam na torcida.

Aos amigos do mestrado, pela nossa união, característica principal de nossa turma. Em especial, Mariana, Renata, Marcel, Carlos, Henrique, Sidcleide, Débora e Ricardo.

Às minhas amigas irmãs Janaína e Kelly, jamais esquecerei tudo o que fizeram por mim, e por serem acolhedoras e otimistas. Amizade para a vida toda!

Às amigas da República "As Mercenárias"Cris e Camila, pela paciência, carinho, compreensão em nossa vivência e momentos de descontração.

À Carolina Bicalho, pela amizade, preocupação e principalmente compa-

nheirismo.

Ao Ismael, pela preocupação de pai e por me considerar como filha.

Aos alunos do doutorado que de alguma forma sempre estavam dispostos a ajudar. Em especial a Jackelya Araújo da Silva e Gilberto Rodrigues Liska.

À Cris Nogueira, que me fez aprender a gostar de Lavras!

Aos meus orientadores e aos professores João Domingos Scalon e Carla Regina Guimarães Brighenti por terem aceito o convite para serem examinadores na banca e pelas considerações apresentadas para contribuição deste trabalho.

E por fim, a todos que torceram e que, diretamente ou indiretamente, contribuíram pelo meu sucesso.

"As melhores coisas da minha vida foram as lições que aprendi com as coisas ruins que me aconteceram."

Jerry Seinfeld

RESUMO

Os métodos automáticos de classificação têm sido desenvolvidos na área de Aprendizado de Máquina com o intuito de facilitar a categorização de dados. Dentre os métodos mais bem sucedidos destacam-se o Boosting e o Bagging. O Bagging funciona combinando classificadores ajustados em amostras *bootstrap* dos dados e o Boosting funciona aplicando-se sequencialmente um algoritmo de classificação a versões reponderadas do conjunto de dados de treinamento, dando maior peso às observações classificadas erroneamente no passo anterior. Esses classificadores se caracterizam por produzirem resultados satisfatórios, baixo custo computacional e vantagem da simplicidade de implementação. Dadas essas características, surge um interesse em verificar o desempenho desses métodos automáticos comparados com os métodos clássicos de classificação existentes na Estatística, a Análise Discriminante Linear e Quadrática. Com o propósito de comparar essas técnicas, utilizou-se as taxas de erro de classificação dos modelos. Para melhorar a confiança da utilização dos métodos Boosting e Bagging em problemas mais complexos de classificação, um estudo foi realizado aplicando essas técnicas em dados reais e simulados que eram compostos por mais que duas categorias na variável resposta. Nesta dissertação, para estimular a implementação do Boosting e Bagging, realizou-se uma aplicação na Análise Sensorial. Concluiu-se que os métodos automáticos tiverem um bom desempenho de classificação, proporcionando taxas de erro menores que as Análises Discriminante Linear e Quadrática nas aplicações testadas.

Palavras-chave: Métodos de classificação. Qualidade de cafés. Análise Discriminante.

ABSTRACT

Automatic classification methods have been developed in machine learning area in order to facilitate the categorization of data. Among the most successful methods include the Boosting and Bagging. The Bagging works by combining classifiers adjusted in bootstrap samples of the data and the Boosting works by applying sequentially an algorithm to rank the reweighted versions of the set of training data, giving greater weight to the observations misclassified in the previous step. These classifiers are characterized by providing satisfactory results, low computational cost and benefit of implementation simplicity. Given these characteristics, comes an interest in checking the performance of these automated methods compared with traditional existing classification methods in Statistics, Linear Discriminate Analysis and Quadratic. In order to compare these techniques it was used misclassification rates and accuracy of the models. To improve confidence in the use of Boosting and Bagging methods in more complex problems of classification, a study was carried out by applying these techniques in real and simulated data composed of more than two categories in the response variable. In this dissertation, to encourage the implementation of Boosting and Bagging was held in an application Sensory Analysis. We conclude that automatic methods have a good classification performance by providing lower error rates than Discriminant Linear analysis and Quadratic Discriminant analysis in the tested applications.

Keywords: Classification methods. Quality coffees. Discriminant Analysis.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Métodos automáticos de classificação têm sido desenvolvidos no Aprendizado de Máquina utilizando a combinação de classificadores. É de conhecimento comum que uma combinação de opiniões induz a uma decisão melhor do que uma decisão tomada individualmente. Em virtude disso, pesquisadores da área da Computação, denominada Aprendizado de Máquina, criaram métodos de classificação baseado na combinação de classificadores, com o intuito de tentar reproduzir a habilidade dos humanos em máquinas para aperfeiçoar o desempenho do computador em algumas tarefas, sendo a classificação uma das mais importantes (OLIVEIRA; NASCIMENTO, 2012).

Dentre os métodos automáticos de classificação disponíveis, destacam-se o Boosting (SCHAPIRE, 1990) e o Bagging (BREIMAN, 1996). Em síntese, o Boosting cria um classificador dado pela combinação dos classificadores gerados em cada iteração a partir de versões reponderadas do conjunto de treinamento, dando maior peso às observações classificadas erroneamente na iteração anterior. Por outro lado, a ideia principal do procedimento Bagging consiste em criar réplicas *bootstrap* do conjunto de treinamento, dando origem a vários classificadores que serão utilizados para formar o classificador final. Esses classificadores são combinados de forma que, as observações do conjunto de treinamento pertencerão à classe de maior frequência.

Convém ressaltar que tanto o Boosting quanto o Bagging podem ser aplicados independentemente da complexidade dos dados em termo do número de observações, características e o número de classes que pertencem a variável resposta. Pois, na prática, as situações de classificação normalmente apresentam o número de categorias superior a dois. Essas características têm permitido que esses métodos de classificação sejam aplicados na solução de um número cada vez maior

de problemas práticos, como análise de DNA, análise de imagens, autenticação de pessoas via dados biométricos entre outros.

Dadas essas considerações, é aceitável que os métodos automáticos relacionados à classificação sejam utilizados no tratamento de dados sensoriais para medir precisamente as características dos produtos a partir de respostas humanas, principalmente em situações que envolvam números expressivos de classes que constituem a variável resposta.

Na Estatística, técnicas multivariadas como Análise Discriminante Linear e Quadrática são utilizadas na Análise Sensorial para fins de previsão, caracterização de produtos, identificação de consumidores, etc. Contudo, existe restrição para a relação entre a variável resposta e as variáveis preditoras que dificultam a aplicação dessas técnicas em algumas situações. Nesse contexto, a utilização dos métodos Boosting e Bagging, para problemas de classificação considerando experimentos sensoriais, permite discriminar pequenas diferenças entre amostras e diferenciar os atributos de forma promissora a produzir resultados mais precisos na atividade de classificação.

Haja vista que, existem poucos relatos na literatura referente à aplicação dos métodos computacionais de classificação em experimentos sensoriais, não foram encontrados estudos que utilizam as técnicas automáticas baseada na combinação de classificadores para situações em que a variável resposta é composta por mais que duas categorias na área de Análise Sensorial.

Portanto, o objetivo deste trabalho é analisar a eficiência de duas técnicas de combinação de classificadores, o Boosting e Bagging, aplicados na Análise Sensorial de Cafés Especiais produzidos na Serra da Mantiqueira. Adicionalmente, serão comparados os métodos tradicionais de classificação com os métodos de Aprendizado de Máquina, para verificar quais das técnicas têm a maior capaci-

dade de discriminação e assim justificar o uso de novas metodologias em dados sensoriais nas situações que envolvam problemas com três ou mais categorias.

2 REFERENCIAL TEÓRICO

As seguintes seções apresentam os temas centrais, os quais abordam a Análise Sensorial e o Aprendizado de Máquina, além de descrever os fundamentos dos métodos de classificação baseados na combinação de classificadores e das técnicas multivariadas. Em seguida será apresentado o critério de verificação do desempenho dos métodos para efetuar as devidas comparações.

2.1 Análise Sensorial

A Análise sensorial é um fator importante para a qualidade de uma bebida ou alimento, pois determina a aceitabilidade do consumidor. De acordo com a Associação Brasileira de Normas e Técnicas (1993), a Análise Sensorial é utilizada para analisar e interpretar reações provocadas pelos alimentos aos sentidos humanos.

Segundo Costel e Duran (1982), a avaliação sensorial dos alimentos é uma função primária do homem, na qual os alimentos são rejeitados ou aceitos conforme as sensações sentidas ao observá-los, envolvendo um conjunto de técnicas elaboradas com o intuito de avaliar um produto através de percepções, sensações e reações do consumidor sobre as características dos produtos (MINIM, 2006).

A avaliação sensorial é feita pelos órgãos dos sentidos, principalmente do paladar, olfato e tato, quando se ingere um alimento. Resultante da interação de nossos sentidos, essa complexa sensação é usada para medir a qualidade dos alimentos e auxiliar no desenvolvimento de novos produtos (MINIM, 2006).

Entre as aplicações na indústria alimentícia e nas instituições de pesquisa, a Análise Sensorial se destaca pela colaboração nas etapas de desenvolvimento de um novo produto, no controle de qualidade e no auxílio da seleção de métodos instrumentais que tenham correlação com os atributos sensoriais de alimentos

(LANCHOTE, 2007).

A indústria faz uso de técnicas modernas de análise sensorial com o objetivo de caracterizar diferenças e similaridades entre produtos que disputam um mesmo mercado consumidor, otimizar atributos sensoriais de aparência, aroma, sabor e textura de alimentos de acordo com as expectativas do mercado, avaliar alterações sensoriais ocorridas em função do tempo e condições de armazenamento, embalagem, processamento, matéria-prima, etc. (PAIVA, 2005). Della et al. (2006) afirmam que um produto pode apresentar excelentes características químicas, físicas e microbiológicas, porém, é imprescindível que as características sensoriais atendam aos anseios e às necessidades do consumidor.

Para avaliar o desenvolvimento de um produto, deve-se basear nas respostas das seguintes questões fundamentais: "O produto é aceito pelos consumidores? Qual a preferência do consumidor? Existe diferença perceptível entre o produto em estudo? Quais qualidades sensoriais estão presentes?" Há bibliografias que definem essas questões em Afetivas, pois testam a aceitação e preferência do consumidor e Analíticas, que incluem os testes descritivos e os discriminativos. Nos testes analíticos, o provador é utilizado como um instrumento que pode ou não ter recebido o mínimo de treinamento.

Os julgadores não treinados, são pessoas selecionadas aleatoriamente que consomem frequentemente o produto. Segundo Chaves (1980), para esse tipo de julgadores, é necessário um grande número de consumidores (no mínimo trinta). Os julgadores treinados são aqueles que possuem uma boa habilidade para perceber algumas propriedades sensoriais. Estes juízes também devem abster-se de hábitos que prejudiquem a habilidade sensorial (MORAES, 1988) e precisam possuir sensibilidade olfativa e gustativa para diferenciar nuances especiais formadas na bebida, identificando com precisão a qualidade do café (ILLY, 2002).

Oliveira (2010) declara que a faixa etária dos provadores deve estar entre 18 e 50 anos. Pois, crianças não têm a capacidade de usar uma terminologia adequada para expressar suas próprias impressões sensoriais e pessoas com mais de 50 anos já não possuem uma boa acuidade sensorial devido à perda da sensibilidade das células da língua.

2.1.1 Cafés Especiais

O segmento de cafés especiais surgiu entre 1970 e 1980, em plena crise de consumo norte-americana. Inicialmente, um grupo de indústrias fundou a Associação Americana de Café Especiais, com o objetivo de estimular a produção. Pode-se dizer que tenha surgido como um meio de driblar preocupações relacionadas à produção ou, até mesmo, apenas para agregar valor ao produto. Os cafés especiais resistem melhor à crise. Na verdade, suas características de sabor e métodos de produção os tornam originais, o que garante um melhor preço, uma vez que são valorizados pelos torreadores e consumidores (AVELINO et al., 2005).

O conceito de café especial está ligado ao prazer que a bebida pode proporcionar, por meio de algum atributo específico, processo de produção ou serviço a ele associado. Portanto, ele diferencia-se dos cafés comuns por características como qualidade superior da bebida, aspecto dos grãos, forma de colheita, processamento pós-colheita, história, origem dos plantios, cultivares e certificações, entre outras. Podem também incluir parâmetros de diferenciação que se relacionam à sustentabilidade econômica, ambiental e social, como sistemas de produção e as condições da mão de obra sob os quais os cafés são produzidos. A rastreabilidade e a incorporação de serviços também são fatores de diferenciação e, portanto, de agregação de valor (SOUZA; SAES; OTANI, 2002; BORÉM et al., 2008).

Embora o café seja a bebida mais popular do mundo, na literatura é ci-

tado como uma bebida estimulante que pode causar irritabilidade e ou insônia, por conter alto teor de cafeína, e pode influenciar no comportamento das pessoas. Entretanto o consumo de café tem sido incentivado por especialistas que afirmam que o consumo do café está associado a efeitos benéficos na prevenção de doenças como: diabetes tipo II, asma, cirrose alcoólica, determinados tipos de cancro, doença de Parkinson e Alzheimer.

Considerando as particularidades da bebida e sua importância na área da saúde, a aceitação dos cafés pelo consumidor se torna uma importante informação para analisar alguns componentes da sua composição, visando melhoria e qualidade.

2.1.2 Análise Sensorial dos Cafés Especiais

Para a avaliação da qualidade do café, a utilização de análise sensorial se tornou uma ferramenta imprescindível (MAMEDE et al., 2010). Sendo utilizada para avaliar o gosto do café e também traçar o perfil sensorial de diferentes tipos de café em relação ao grau de torra. A Speciality Coffe Association Of America (2009) propõe que os atributos da bebida: aroma, doçura, sabor, acidez, corpo, finalização, equilíbrio, defeitos e avaliação global sejam analisados conforme a intensidade com que essas caraterísticas se apresentam. Dessa forma, o provador pode determinar diferentes características sensoriais entre diferentes amostras, além de descrever notas de cada atributo que receberão pontuação conforme a Tabela 1.

A somatória das notas de cada atributo, com exceção do atributo defeito, são correspondentes à classificação final da bebida. A amostra que apresenta pontuação superior a 80 é classificada como café especial (BRAZIL SPECIALTY COFFEE ASSOCIATION, 2008).

Tabela 1 Escala de classificação de cafés especiais.

Escala Atributo	Pontuação	Descrição	Classificação
9 |- 10	90 - 100	Exemplar	Cafés especiais
8 |- 9	85 - 89,99	Excelente	Cafés especiais
7 |- 8	80 - 84,99	Muito Bom	Cafés especiais
6 |- 7	60 - 80	Bom	Cafés não especiais

Fonte: Adaptada de SCAA (2009)

Para detectar pequenas diferenças entre os cafés especiais que sofreram tratamentos incomuns de processamento, secagem e torra, motivos que proporcionam diferentes características no gosto da bebida, são necessárias medidas eficazes que garantam um satisfatório poder discriminatório dos tipos de café.

Uma forma para identificação dos tipos de cafés pode ser realizada pelos métodos desenvolvidos na Área de Aprendizado de Máquina, Boosting e Bagging, e pelos métodos de classificação encontrados na Estatística, como a Análise Discriminante Linear e Quadrática.

2.2 Aprendizado de Máquina

O Aprendizado de Máquina, ou Aprendizagem Automática, é uma área da Inteligência Artificial (AI) dedicada a buscar métodos ou dispositivos computacionais que possuam a capacidade racional do ser humano de resolver problemas (LUGER, 2004). Pode também ser definido como o ramo da ciência da computação que se ocupa do comportamento inteligente ou ainda, o estudo de como fazer os computadores realizarem atividades que, atualmente, os humanos fazem melhor (RICH; KNIGHT, 1994).

Weiss e Kulikowski (1991) definem um sistema de aprendizado como um algoritmo que toma decisões baseado em experiências acumuladas por meio da solução bem sucedida de problemas anteriores. Segundo Lorena e Carvalho (2007),

as técnicas de aprendizado de máquinas empregam um princípio de inferência denominado indução, no qual é possível obter conclusões genéricas a partir de um conjunto particular de exemplos.

No contexto de classificação, o método de aprendizado de máquina tenta prever a classificação de novas observações a partir de exemplos que foram previamente rotulados com suas classificações corretas (SCHAPIRE; FREUND, 2012), tornando possível classificar melhor com o tempo à medida que observamos novos exemplos.

Para alguns sistemas de aprendizagem é necessário predizer se uma certa ação irá fornecer uma certa saída. Nesse sentido, é possível classificar os sistemas de aprendizado de máquina em supervisionado e não supervisionado (RUSSSEL; NORVIG, 2003).

2.2.1 Aprendizado não supervisionado

No aprendizado não supervisionado, tem-se a incerteza sobre a classificação esperada, isto é, não tem o conhecimento prévio sobre os atributos das classes. Dessa forma, para um conjunto de observações sem a presença de suas respectivas classes, é necessário utilizar métodos probabilísticos para simular uma experiência não vivida (CONDUTA; MAGRIN, 2010).

Lorena e Carvalho (2007) destacam que o algoritmo de aprendizado de máquina não supervisionado aprende a representar (ou agrupar) as observações submetidas segundo medidas de similaridade.

2.2.2 Aprendizado supervisionado

O aprendizado supervisionado tem o intuito de induzir conceitos a partir de um conjunto de treinamento para que o conhecimento adquirido permita classi-

ficar novas observações com categorias desconhecidas. Ou seja, dado um conjunto de observações em que a classe de cada exemplo é conhecida, o objetivo é encontrar uma hipótese capaz de classificar novas observações entre as classes já existentes (PRATI, 2006).

2.2.3 Conjunto de teste e treinamento

Um conjunto de treinamento é composto por exemplos contendo valores de atributos bem como a classe associada. Na Tabela 2, é mostrado o formato padrão de um conjunto de treinamento com n exemplos e m atributos. Nessa tabela, a linha i refere-se ao i-ésimo exemplo $(i = 1, 2, ..., n)$ e a entrada x_{ij} refere-se ao valor do j-ésimo $(j = 1, 2, ..., m)$ atributo X_j do exemplo i.

Tabela 2 Conjunto de treinamento no formato atributo - valor.

	X_1	X_2	$\cdots$	X_m	Y
z_1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_1
z_2	x_{21}	x_{22}	$\cdots$	x_{2m}	y_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
z_n	x_{n1}	x_{n2}	$\cdots$	x_{nm}	y_n

A amostra (ou conjunto) de treinamento é dado por (x_i, y_i), em que y_i denota a classe e fica subentendido o fato que x_i é um vetor. Este conjunto é formado para produzir um modelo que seja capaz de predizer a classificação de novas observações.

O conjunto de teste é formado pelas observações restantes ao conjunto de treinamento e utilizado para avaliar a capacidade do modelo proposto em fazer previsões. Sendo assim, um conjunto de teste é utilizado como fonte extra de informação para avaliar a taxa de acerto do modelo (SILVA, 2008).

A divisão do conjunto total de dados em dois subconjuntos mutuamente exclusivos, um para treinamento e outro para teste, geralmente é realizada pelo Método Holdout, em que os conjuntos de dados podem ser separados em quantidades iguais ou não. Uma proporção muito comum é considerar $\frac{2}{3}$ para treinamento e o $\frac{1}{3}$ restante para o teste (KOHAVI, 1995). Essa abordagem é indicada quando está disponível uma grande quantidade de dados. Caso o conjunto de dados seja pequeno, o erro calculado na predição pode sofrer muita variação.

2.3 Métodos Automáticos de Classificação

Em Aprendizado de Máquina e Estatística, a classificação é definida como o problema de identificar a qual conjunto de categorias uma nova observação pertence, com base em um conjunto de treinamento de dados contendo observações cuja categoria é conhecida. Podendo ser determinada por dois problemas distintos: classificação binária e classificação multiclasse.

A classificação binária é o ato de classificar observações de um conjunto de dados em dois grupos. Ou seja, consiste em atribuir a uma observação, com base em covariáveis, uma variável dependente composta por apenas duas classes. Um cenário simples poderia ser um pesquisador classificar transações de cartões de crédito como legítimas ou fraudulentas com base em informações como: o intervalo de tempo entre uso do cartão, faixa de valores usados em comparação com o mês anterior, o cartão ser usado com valores expressivos, entre outras. Para este exemplo, a categorização é dada por sim ou não de acordo com a presença da característica de interesse no conjunto das variáveis explicativas.

Considerando uma situação em dados sensoriais, um exemplo seria a classificação de cafés como especiais ou não, em que as variáveis explicativas poderiam ser dadas pelas condições de temperatura, altitude, colheita, secagem e etc.

A classificação multiclasse é quando o problema em questão é constituído de resposta com três ou mais categorias. Para exemplificação, considere uma situação em que se deseja categorizar a temperatura de uma determinada cidade onde os seguintes valores são: Quente, Ameno e Frio. Neste tipo de situação, as variáveis explicativas podem ser pressão, temperatura, umidade, velocidade do vento, entre outras. Na análise sensorial, pode-se considerar a classificação de cafés especiais como exemplar, excelente, muito bom e bom. A variável resposta é composta por quatro categorias explicadas pelas notas atribuídas às caraterísticas sensoriais como: aroma, sabor, acidez e doçura.

Recentemente, em uma área de pesquisa formada por pesquisadores na área da computação e denominada aprendizado de máquina (machine learning), muitos pesquisadores têm se esforçado na busca por métodos de classificação automáticos que combinem as potencialidades de vários classificadores, tais como o método Boosting e o método Bagging.

Para se executar a tarefa de classificação baseada nos métodos automáticos, são usados dados que consistem em um conjunto de atributos denominados previsores, e um atributo denominado preditor (classe). Os atributos previsores são utilizados para definir uma classificação efetiva dos registros pertencentes à base de dados em estudo. O atributo preditor por sua vez é utilizado como uma hipótese de classificação que será validada ou não pela análise resultante da classificação através dos atributos previsores (CARVALHO, 2001).

Neste contexto, um algoritmo de classificação, dito algoritmo indutor, consiste em dividir a base de dados em dois conjuntos de instâncias mutuamente exclusivos. Um dos subconjuntos é o conjunto de treinamento e o outro um conjunto de teste. Inicialmente, o conjunto de treinamento é percorrido, analisando as relações existentes entre os atributos previsores e o atributo preditor. Estas relações

são então usadas para prever a classe dos registros presentes no conjunto de teste, que será a próxima ação do classificador (MITCHELL, 1997).

2.3.1 Método Boosting

Freund e Schapire (1996) desenvolveram o método Boosting para melhorar o desempenho da classificação em um determinado conjunto de dados. Este procedimento é considerado como uma das mais importantes e bem sucedidas metodologias desenvolvidas nos últimos anos na literatura de classificação e tem recebido bastante atenção dos estatísticos desde seu surgimento na literatura de aprendizado automático (SCHAPIRE, 1990).

A utilização do método Boosting como ferramenta de classificação tem despertado interesse em diversas áreas como Reconhecimento de padrões, Mineração de dados, Medicina e outras, devido ao baixo custo computacional e vantagem da simplicidade e facilidade de implementação, além do seu alto potencial de classificação.

De maneira geral, o Boosting consiste em aplicar, sequencialmente, uma regra de classificação final a versões iterativamente reponderadas da amostra de treinamento. A cada iteração, o algoritmo atribui pesos maiores às observações classificadas erroneamente na iteração anterior. A regra final de classificação é obtida através de uma combinação linear dos classificadores parciais construídos.

A Figura 1 ilustra de forma geral como o Boosting funciona, em que w_i são os pesos atribuídos às observações em cada iteração t; os C_i são os classificadores parciais encontrados em cada amostra ponderada e H é a hipótese final de classificação.

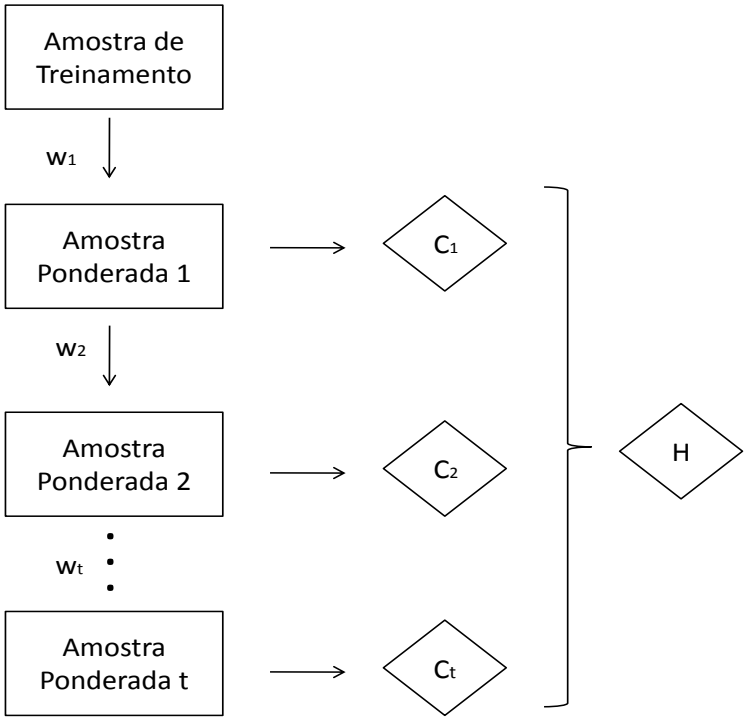


Figura 1 Esquema de funcionamento do Algoritmo Boosting

Várias versões do procedimento boosting foram desenvolvidas, tais como: PBoost, TotalBoost, BrownBoost, MadaBoost, LogitBoost. Porém, a mais utilizada e difundida na literatura e na prática é o Adaboost (FREUND; SCHAPIRE, 1996), uma vez que foi o primeiro algoritmo originalmente desenvolvido para a classificação de problemas com respostas binárias com o objetivo de aumentar a precisão de qualquer outro algoritmo de aprendizagem (SCHAPIRE; FREUND, 2012).

Uma breve descrição do Algoritmo Boositng (AdaBoost) para o problema de classificação binária é: considere os dados de treinamento $T = \{x_i, y_i\}_{i=1}^{N}$,

em que, x_i é um vetor de características; y_i a variável resposta que recebe os valores $+1$ e -1 e N é o tamanho total do conjunto de treinamento. Definimos o classificador final

$$F(x) = sign\left(\sum_{i=1}^{N} c_m f_m(x)\right)$$

em que, c_m são constantes que medem a importância do classificador dada por

$$C_m = \frac{1}{2} ln\left(\frac{1-\varepsilon_m}{\varepsilon}\right)$$

Em cada iteração m, um classificador $f_m(x)$ que retorna os valores das classes é construído em versões ponderadas da amostra de treinamento. A ponderação das amostras é realizada através dos pesos ${w_i}^{(m)}$. Inicialmente, todos os pesos são iguais, mas em cada execução, os pesos de observações classificadas incorretamente são incrementados para que o algoritmo seja forçado a focar nas obervações mais difíceis de serem classificadas, sujeito ao chamado fator de normalização Z_m para garantir que a soma dos pesos não ultrapasse o valor um, dado que os pesos são probabilidades de classificação correta da observação. O classificador final é produzido com uma combinação linear dos classificadores construídos em cada uma das amostras ponderadas. No Algoritmo 1, é apresentado o procedimento de execução do AdaBoost para os problemas de classificação binária.

Algoritmo 1. AdaBoost para classificação binária

1. Conjunto de dados: $(x_1, y_1), ..., (x_N, y_N)$ em que $x_i \in X$ e $y_i \in \{-1, +1\}$

Inicie os pesos ${w_i}^{(m)} = \frac{1}{N}, i = 1, 2, ..., N$

2. Repita $m = 1, ..., M$

Ajuste o classificador $f_m(x) \in -1, +1$ usando os pesos w_i e os dados de treinamento;

Calcule:

$$\varepsilon_m = \frac{\sum_{i=1}^{N} {w_i}^{(m)} I[Yi \neq f_m(x_i)]}{\sum_{i=1}^{N} {w_i}^{(m)}} \quad \text{e} \quad c_m = \frac{1}{2} ln \left(\frac{1 - \varepsilon_m}{\varepsilon_m} \right)$$

3. Faça a atualização dos pesos:

$${w_i}^{m+1} = \frac{{w_i}^m}{Z_m} \begin{cases} e^{-c_m}, & \text{se} \quad y_i = f_m(x_i), \\ e^{c_m}, & \text{se} \quad y_i \neq f_m(x_i). \end{cases}$$

em que $Z_m = \sum_{i=1}^{N} w_i^m e^{-c_m y_i f_m(x_i)}$ é um fator de normalização

4. Classificador final, $F(x)$:

$$F(x) = sign \left(\sum_{i=1}^{N} c_m f_m(x) \right)$$

Para entender melhor o funcionamento do Boosting, observe o problema de aprendizagem mostrado na Figura 2, em que os pontos, ou observações, serão classificados como + ou -. Neste exemplo, tem-se que o conjunto de treinamento, ou seja, os pontos pertencentes a este plano, é dado por 10 pontos, sendo cinco

positivos e cinco negativos.

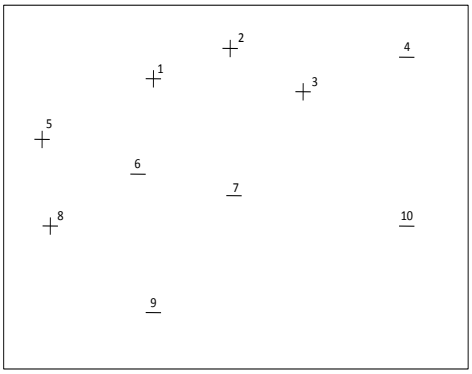


Figura 2 Conjunto de treinamento

Considere que o classificador base encontrado seja definido por uma linha vertical no plano como mostra a Figura 3. Nesse caso, apenas linhas verticais e horizontais são definidas para formar polígonos que facilitam a separação das duas classes em várias iterações, dado que a utilização de uma reta diagonal realiza a classificação dos dados em uma única iteração.

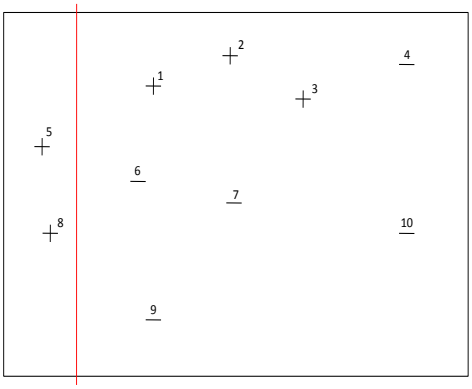


Figura 3 Classificador base para a primeira iteração

Inicialmente, o classificador define todos os pontos à direita da linha ver-

tical sendo negativos e todos os outros pontos à esquerda como positivos. Dessa forma, note que as observações 1, 2 e 3 foram classificadas incorretamente com a regra definida pelo classificador. Dessa forma, o AdaBoost para este exemplo é definido como a seguir:

1. O primeiro ponto de interesse do algoritmo AdaBoost é a distribuição igual dos pesos atribuídos às observações com o valor $\frac{1}{10}$.

2. O cálculo da taxa de erro do classificador é feito, numericamente, da seguinte forma: a função indicadora $I[Yi \neq f_m(x_i)]$ recebe o valor um para as três observações classificadas incorretamente e zero para as demais observações corretamente classificadas:

$$\varepsilon_1 = \frac{3\ (0,10 \times 1) + 7\ (0,10 \times 0)}{1} = 0,3$$

Portanto, a importância do classificador é:

$$c_1 = \frac{1}{2} \ln \left(\frac{1 - 0,3}{0,3} \right) = 0,42$$

3. A constante de normalização para esta iteração é:

$$Z_1 = 3\ (0,10 \times e^{0,42}) + 7\ (0,10 \times e^{-0,42}) = 0,92$$

Os novos pesos para as observações que foram classificadas corretamente e incorretamente serão, respectivamente:

$$w_i^{(2)} = \frac{0,10}{0,92} \times e^{-0,42} = 0,07 \quad \text{para} \quad y_i = f_m(x_i)$$

$$w_i^{(2)} = \frac{0,10}{0,92} \times e^{0,42} = 0,17 \quad \text{para} \quad y_i \neq f_m(x_i)$$

Para a segunda iteração, suponha que o segundo classificador define uma linha vertical em outra posição, como apresentado na Figura 4. A nova regra define os pontos à esquerda da linha como positivos e à direita como negativos.

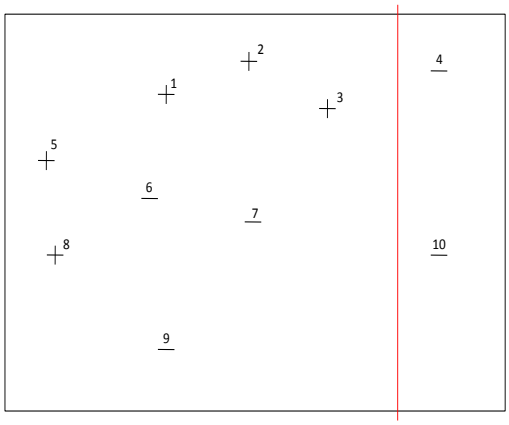


Figura 4 Classificador base para a segunda iteração

Nesta iteração, as observações 6, 7 e 9 foram erroneamente classificadas. A execução do algoritmo para a segunda iteração é dada a seguir:

1. O erro de treinamento:

$$\varepsilon_2 = \frac{3\,(0,17 \times 0) + 4\,(0,07 \times 0) + 3\,(0,07 \times 1)}{(0,17 \times 3) + (0,07 \times 4) + (0,07 \times 3)} = 0,21$$

Importância do classificador:

$$c_2 = \frac{1}{2} \ln\left(\frac{1 - 0,21}{0,21}\right) = 0,66$$

3. A constante de normalização para a segunda iteração é:

$$z_2 = 3\,(0,17 \times e^{-0,66}) + 4\,(0,07 \times e^{-0,66}) + 3\,(0,07 \times e^{0,66}) = 0,82$$

Atualização dos pesos para a terceira iteração:

$$w_i^{(3)} = \frac{0,17}{0,82} \times e^{-0,66} = 0,11 \quad \text{para} \quad y_i = f_m(x_i)$$

$$w_i^{(3)} = \frac{0,07}{0,82} \times e^{-0,66} = 0,05 \quad \text{para} \quad y_i = f_m(x_i)$$

$$w_i^{(3)} = \frac{0,07}{0,82} \times e^{0,66} = 0,16 \quad \text{para} \quad y_i \neq f_m(x_i)$$

Considerando o próximo classificador uma linha horizontal como apresentado na Figura 5, os pontos abaixo da linha serão classificados como negativos e os pontos acima da linha como positivos, para a terceira iteração tem-se que as observações 4, 5 e 8 foram classificadas incorretamente.

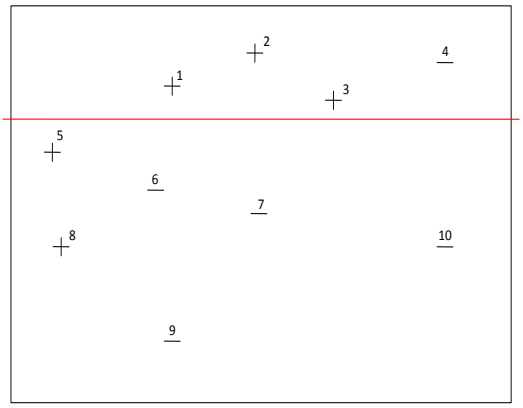


Figura 5 Classificador base para a terceira iteração

O erro de treinamento e a importância do classificador para a atual iteração são, respectivamente:

$$\varepsilon_3 = \frac{3\,(0,11 \times 0) + 3\,(0,05 \times 1) + 3\,(0,16 \times 0) + (0,05 \times 0)}{(0,11 \times 3) + (0,05 \times 3) + (0,16 \times 3) + (0,05 \times 1)} = 0,15$$

$$c_3 = \frac{1}{2} \ln\left(\frac{1 - 0,15}{0,15}\right) = 0,87$$

Após as iterações, o algoritmo Boosting combina linearmente todos os coeficientes que medem a importância dos classificadores parciais encontrados, denominados classificadores fraco, num único classificador, chamado de forte, como apresenta a Figura 6.

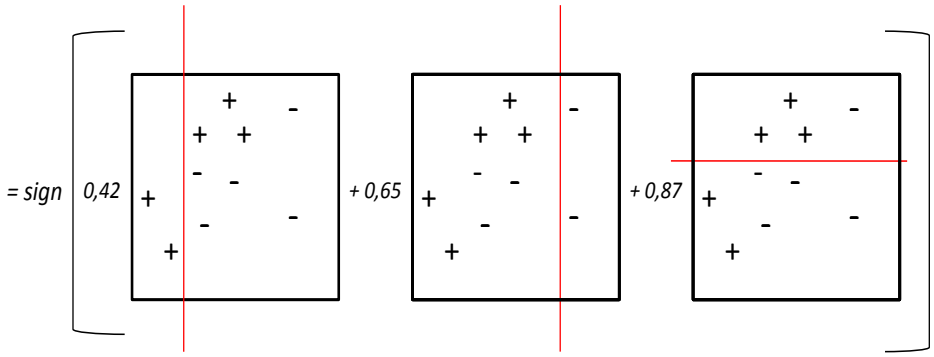


Figura 6 Combinação linear dos classificadores fracos

Dessa forma, as classes estimadas para cada observação são:

$$\hat{y}_1 = sign[(-1 \times 0,42) + (1 \times 0,66) + (1 \times 0,87)] > 0 \longrightarrow +$$

$$\hat{y}_2 = sign[(-1 \times 0,42) + (1 \times 0,66) + (1 \times 0,87)] > 0 \longrightarrow +$$

$$\hat{y}_3 = sign[(-1 \times 0,42) + (1 \times 0,66) + (1 \times 0,87)] > 0 \longrightarrow +$$

$$\hat{y}_4 = sign[(-1 \times 0,42) + (-1 \times 0,66) + (-1 \times 0,87)] < 0 \longrightarrow -$$

$$\hat{y}_5 = sign[(1 \times 0,42) + (1 \times 0,66) + (1 \times 0,87)] > 0 \longrightarrow +$$

$$\hat{y}_6 = sign[(-1 \times 0,42) + (1 \times 0,66) + (-1 \times 0,87)] < 0 \longrightarrow -$$

$$\hat{y}_7 = sign[(-1 \times 0,42) + (1 \times 0,66) + (-1 \times 0,87)] < 0 \longrightarrow -$$

$$\hat{y}_8 = sign[(1 \times 0,42) + (1 \times 0,66) + (1 \times 0,87)] > 0 \longrightarrow +$$

$$\hat{y}_9 = sign[(-1 \times 0,42) + (1 \times 0,66) + (-1 \times 0,87)] < 0 \longrightarrow -$$

$$\hat{y}_{10} = sign[(-1 \times 0,42) + (-1 \times 0,66) + (-1 \times 0,87)] < 0 \longrightarrow -$$

Observe que o AdaBoost foi capaz de construir um classificador combinado que classifica corretamente todas as observações em apenas três iterações.

Uma importante propriedade do AdaBoost, que também pode ser observada neste exemplo, é que o algoritmo tem a capacidade de reduzir rapidamente o erro de treinamento. Após a primeira iteração, o objetivo do classificador base é corrigir os pesos dos objetos erroneamente classificados para obter o erro mínimo. Se o erro de treino de cada hipótese do classificador fraco for ligeiramente melhor, temos que o erro de treinamento decresce exponencialmente (SCHAPIRE; FREUND, 2012).

Uma forma de determinar o número de iterações para que o algoritmo não seja executado indefinidamente é analisar o erro de treinamento, a partir do momento em que não existe variação no erro, temos o número de iterações suficientes para a construção do classificador final.

Apesar de todas as teorias do AdaBoost serem voltadas para gerar classificadores para problemas binários, existe um modo simples de se transformar o AdaBoost em um algoritmo de classificação mullticlasse.

Muitas vezes substituída por questões com duas categorias, por exemplo, a classificação de uma carta em um baralho, pode-se substituir a pergunta multiclasse "Qual é essa carta?"que tem 52 repostas possíveis, por questões binárias como:

"É uma letra ou não?"

"É um número ou não?"

"É um dez ou não?"

Mas isso pode se tornar uma tarefa muito exaustiva em certas situações, principalmente quando se tem um número considerável de classe. Portanto, essa abordagem de reduzir um problema de aprendizagem mais complicada para um simples problema de classificação deve ser realizada de forma automática. Diante disso, Freund e Schapire (1997) estenderam o AdaBoost a um caso multiclasse, denominado AdaBoost.M1, que possui a vantagem da simplicidade e a facilidade de implementação. A configuração do algoritmo é basicamente a mesma que no caso binário, exceto para o número de classes que podem assumir 3 ou mais categorias. O objetivo do classificador fraco é gerar, em cada iteração m, uma hipótese f_m que atribui a categorização a cada observação com baixo erro de classificação.

A atualização para a distribuição ${w_i}^m$ também é idêntica ao AdaBoost, como dada na primeira forma de atualização do Algoritmo 1. A principal diferença é a hipótese final $F(x)$ que tem como saída o valor Y que maximiza a soma dos valores da importância de um classificador c_m que predizem uma classe.

Na classificação multiclasse o erro de treinamento não tem a mesma propriedade como no caso binário. Durante a execução do AdaBoost.M1, se o erro $\epsilon_m > \frac{1}{2}$, a iteração corrente é abortada e o classificador fraco retorna a calcular outra hipótese de separação. Na eventualidade do classificador fraco não conseguir determinar nenhuma hipótese de separação com um erro $\epsilon_m \leq \frac{1}{2}$, o Adaboost.M1

termina.

2.3.2 Bagging

Proposto por Breiman (1996), o Bagging (Bootstrap Aggregating) baseia-se em criar amostras *bootstraps* dos dados com novos conjuntos de treinamento, dando origem a vários preditores, como mostra a Figura 7.

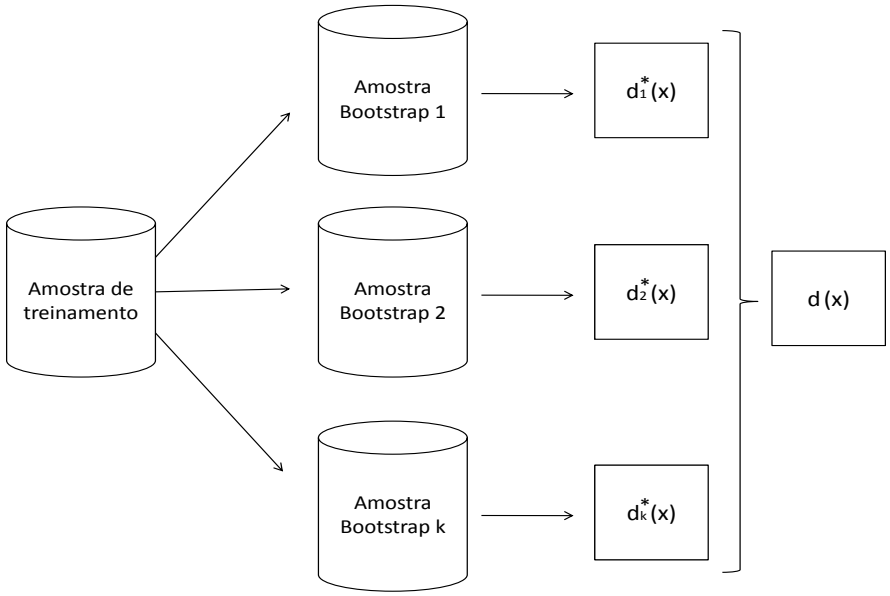


Figura 7 Esquema de funcionamento do Bagging

A fim de gerar o preditor final, que é utilizado para classificar o conjunto de teste, estes preditores são combinados de forma que a classe mais prevista pelos classificadores é a classe escolhida, e para o caso de regressão utiliza-se a média dos preditores.

Em outras palavras, o Bagging funciona da seguinte forma: A partir do conjunto de dados, são retiradas k amostras de tamanho N com reposição. Isto sig-

nifica que em cada amostra *bootstrap* k, uma dada observação do conjunto original pode não aparecer, ou aparecer mais de uma vez. Em cada uma destas amostras, computa-se um preditor. O preditor final agrega os k preditores e a classe mais frequente é a escolhida (RUBESAM, 2004).

Para melhor compreensão do Bagging, na Tabela 3 é apresentada uma base de dados fictícios na qual mostra o tipo de café de acordo com os atributos previsores, genótipo, altitude e processamento natural.

Tabela 3 Conjunto fictício de treinamento da base de dados.

	Variáveis explicativas			Variável resposta
ID	Genótipo	Altitude	Processamento Natural	Tipo de café
1	Bourbon	Alta	Não	Especial
2	Bourbon	Alta	Não	Não Especial
3	Bourbon	Alta	Sim	Especial
4	Bourbon	Baixa	Não	Não Especial
5	Bourbon	Baixa	Sim	Não Especial
6	Catuaí	Alta	Não	Especial
7	Catuaí	Alta	Sim	Não Especial
8	Catuaí	Baixa	Sim	Especial
9	Catuaí	Baixa	Não	Especial
10	Icatu	Alta	Não	Não Especial
11	Icatu	Alta	Não	Especial
12	Icatu	Alta	Sim	Não Especial
13	Icatu	Baixa	Sim	Não Especial
14	Icatu	Baixa	Não	Não Especial

A primeira etapa consiste em construir amostras *bootstrap*, a partir da amostra original de dados, com reposição e de mesmo tamanho do conjunto de treinamento. A Tabela 4 apresenta as amostras geradas.

Tabela 4 Amostras *bootstrap* para a base de dados.

Amostra 1					Amostra 2					Amostra 3				
ID	G	A	PN	TC	ID	G	A	PN	TC	ID	G	A	PN	TC
1	B	A	N	E	1	B	A	N	E	2	B	A	N	NE
2	B	A	N	NE	3	B	A	S	E	2	B	A	N	NE
3	B	A	S	E	3	B	A	S	E	2	B	A	N	NE
4	B	B	N	NE	6	C	A	N	E	3	B	A	S	E
5	B	B	S	NE	6	C	A	N	E	3	B	A	S	E
6	C	A	N	E	7	C	A	S	NE	3	B	A	S	E
6	C	A	N	E	8	C	B	S	E	4	B	B	N	NE
7	C	A	S	NE	9	C	B	N	E	5	B	B	S	NE
7	C	A	S	NE	10	I	A	N	NE	6	C	A	N	E
9	C	B	N	E	10	I	A	N	NE	8	C	B	S	E
10	I	A	N	NE	11	I	A	N	E	12	I	A	S	NE
10	I	A	N	NE	13	I	B	S	NE	14	I	B	N	NE
10	I	A	N	NE	13	I	B	S	NE	14	I	B	N	NE
14	I	B	N	NE	13	I	B	S	NE	14	I	B	N	NE

Posteriormente, deve-se encontrar os preditores de cada amostra para resultar o preditor final. Na literatura sobre Bagging, o preditor favorito são as árvores de classificação e regressão, que têm variabilidade bastante grande quando se perturba o conjunto de dados. Isto significa que, numa árvore de classificação, são tomadas decisões do tipo "se certo critério é satisfeito, o caso pertence a certa classe. Caso contrário, pertence a outra classe". É isso o que cria a instabilidade, pois ao perturbar o conjunto de dados, esse critério pode mudar bastante (RUBESAM, 2004).

Para o exemplo em questão, considere uma observação com as características apresentadas na Tabela 5 cujas predições são desconhecidas.

Tabela 5 Observação a ser classificada pelo método Bagging.

Observação	Variáveis			
	Previsores			Classificação
	Genótipo	Altitude	Processamento Natural	Tipo de café
a_1	Bourbon	Alta	Não	?

Para conhecer a predição da observação, o processo consiste em submetê-la a três classificadores. Cada classificador informa como resposta um valor de predição. O valor de predição mais frequente é a classificação. A Figura 8 mostra uma regra de classificação definida pelo primeiro classificador obtido a partir da amostra *bootstrap* 1 da Tabela 4.

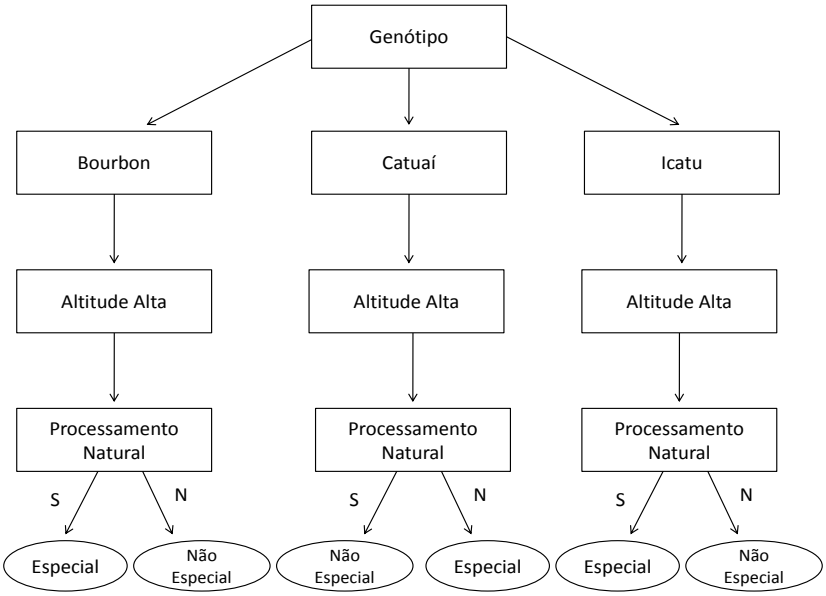


Figura 8 Regra de classificação definida pelo classificador 1

A Figura 9 se refere à regra de classificação do classificador 2, C_2, gerado a partir da amostra *bootstrap* 2.

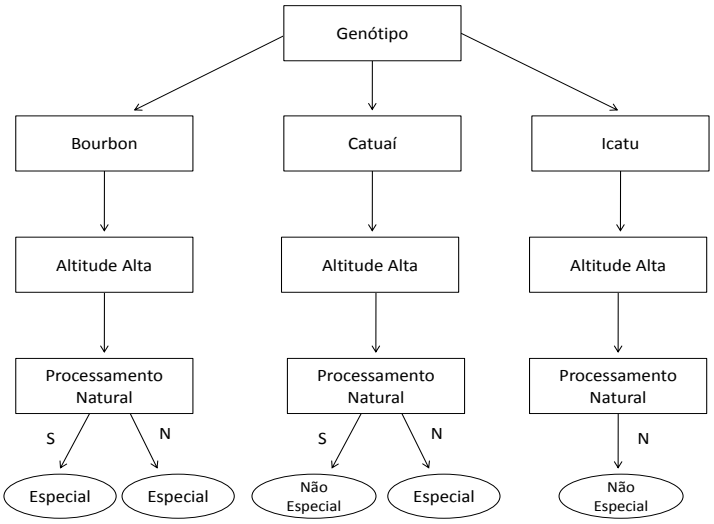


Figura 9 Regra de classificação definida pelo classificador 2

Por fim, a Figura 10 se refere à regra de classificação do classificador 3, C_3, gerado a partir da amostra *bootstrap* 3.

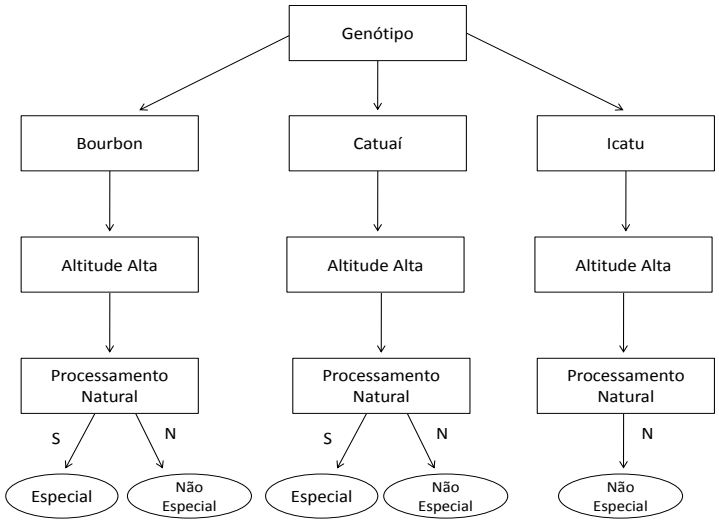


Figura 10 Regra de classificação definida pelo classificador 3

A classificação da observação a_1 resulta na resposta não especial, para o Classificador C_1; especial, para o Classificador C_2 e não especial, para o Classificador C_3. Consequentemente, a resposta não especial será considerada, visto que a mesma obteve dois votos contra um.

Nos casos em que o número de classificadores é ímpar, tem-se uma maioria absoluta dos classificadores. No entanto, quando a quantidade de classificadores é par, pode ocorrer o empate. Portanto, pode-se considerar, que para criação de sistemas de classificação, usando votação majoritária com número restrito de classificadores, é adequado a utilização de número ímpar de classificadores individuais, evitando assim situações não conclusivas de empate (DAISTER, 2007).

2.4 Métodos clássicos de classificação

A análise discriminante é uma técnica da estatística multivariada utilizada para discriminar e classificar objetos. Segundo Khattree e Naik (2000), é uma técnica que estuda a separação de objetos de uma população em duas ou mais classes. A discriminação é a primeira etapa, sendo a parte exploratória da análise e consiste em procurar características capazes de serem utilizadas para alocar objetos em diferentes grupos previamente definidos. Segundo Jonhson & Wichern (1999), a classificação ou alocação pode ser definida como um conjunto de regras que serão usadas para alocar novos objetos.

O problema da discriminação entre dois grupos ou mais, visando posterior classificação, foi proposto por Fisher (1936), e consiste em obter funções matemáticas capazes de classificar um indivíduo $\tilde{x}$ em uma de várias populações π_i, com base em medidas de um número de p características, buscando minimizar o erro de má classificação, isto é, minimizar a probabilidade de classificar incorretamente uma observação em uma população π_i. Um exemplo da função linear e quadrática

é dado na ilustração a seguir.

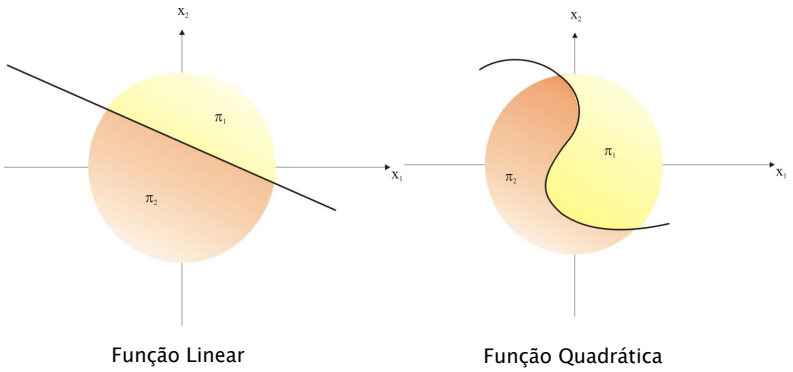


Figura 11 Regiões de alocação para o caso de duas populações

2.4.1 Análise de Discriminação

Segundo Johnson e Wichern (1999), uma boa classificação deve resultar em pequenos erros, ou seja, pouca probabilidade de má classificação, e para isso, a regra de classificação deve considerar as probabilidades a priori e os custos de má classificação. Outro fator a considerar é se as variâncias das populações são ou não iguais. Quando a regra de classificação assume variâncias populacionais iguais, as funções discriminantes são ditas lineares, caso contrário, as funções discriminantes são quadráticas.

A análise discriminante é aplicada às observações com suas classes definidas a priori, descritas por diversas variáveis explicativas. Dessa forma, é construída

uma regra de decisão que permita alocar novos indivíduos, minimizando os erros de alocação.

2.4.2 Análise Discriminante Linear de Fisher

A função discriminante linear de Fisher é uma combinação linear de características originais a qual se caracteriza por produzir separação máxima entre duas populações. Considerando que temos n_1 e n_2 observações, descritas por um conjunto de p variáveis $X_{11}, X_{12}, ..., X_{1n_1}$ amostradas da população π_1 e $X_{21}, X_{22}, ..., X_{2n_2}$ da população π_2, temos que a essência do método de Fisher é transformar as observações multivariadas X em observações univariadas Y tal que os $Y's$ nas populações π_1 e π_2 sejam separadas tanto quanto possível.

Uma das principais vantagens ao se adotar o critério de Fisher (1936) na discriminação, é não ser necessário o conhecimento das densidades populacionais, nem assumir que estas sejam gaussianas. Porém, consideramos o pressuposto de que as matrizes de covariância (Σ) das diferentes populações são iguais. Dessa forma, o melhor estimador não-viesado de Σ é dado por:

$$S_p = \frac{(n_1 - 1)S_1 + (n_2 - 1)S_2}{n_1 + n_2 - 2}$$

em que,

$$S_1 = \frac{1}{n_1 - 1}\left[\sum_{j=1}^{n_1}(X_{1j} - \bar{X}_1)(X_{1j} - \bar{X}_1)^T\right] \quad \text{e} \quad \bar{X}_1 = \frac{\sum_{j=1}^{n_1} X_{1j}}{n_1}$$

$$S_2 = \frac{1}{n_2 - 1}\left[\sum_{j=1}^{n_2}(X_{2j} - \bar{X}_2)(X_{2j} - \bar{X}_2)^T\right] \quad \text{e} \quad \bar{X}_2 = \frac{\sum_{j=1}^{n_2} X_{2j}}{n_2}$$

Apresentado o pressuposto de covariâncias iguais para ambas as populações, temos que a abordagem mais simples para elaborar regras de classificação é a função discriminante linear de Fisher, dada por:

$$Y = (\bar{X}_{\pi_1} - \bar{X}_{\pi_2})\, {S_p}^{-1}\, X$$

A combinação linear é realizada de forma a maximizar a distância ao quadrado entre as médias relativas à variabilidade de Y. O valor médio entre as duas amostras univariadas são apresentadas abaixo:

$$\hat{m} = \frac{1}{2}\,(\bar{Y}_{\pi_1} - \bar{Y}_{\pi_2}) = \frac{1}{2}\,(\bar{X}_{\pi_1} - \bar{X}_{\pi_2})^T\, {S_P}^{-1}\,(\bar{X}_{\pi_1} - \bar{X}_{\pi_2})$$

Utilizando as funções discriminantes de Fisher como base para alocação, uma regra de classificação razoável é aquela que atribui um indivíduo x_o à população π_1, se:

$$Y_o = (\bar{X}_{\pi_1} - \bar{X}_{\pi_2})\, {S_p}^{-1}\, x_o \geq \hat{m}$$

consequentemente, um indivíduo x_o pertence à população π_2 se:

$$Y_o = (\bar{X}_{\pi_1} - \bar{X}_{\pi_2})\, {S_p}^{-1}\, x_o < \hat{m}$$

Segundo Bishop (2006), a representação mais simples de uma função discriminante linear é obtida tendo uma função linear do vetor de entrada, de forma que

$$y(x) = w^T x + w_0$$

no qual w é chamado um vetor de ponderação e w_0 é um viés. Um vetor de entrada x é atribuído à classe C_1 se $y(x) \geq 0$ e para a classe C_2, caso contrário.

A fronteira de decisão, cujas características são cruciais para o desempenho da máquina de classificação a partir da existência de diversas regiões de decisões, é definida pela relação de $y(x) = 0$, que corresponde a um hiperplano $(D-1)$ dimensional dentro da entrada D-dimensional no espaço. Na Figura 12, apresentamos um exemplo em que estão definidas as regiões e fronteiras de decisão para um contexto em que os dados são vetores bidimensionais e há três classes. As fronteiras do exemplo são lineares apenas para facilitar a visualização, mas poderiam perfeitamente ser curvas não-lineares.

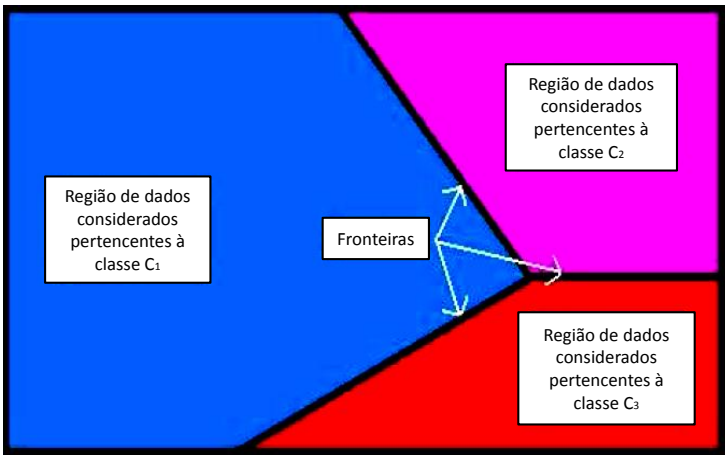


Figura 12 Exemplo com Regiões e Fronteiras de Decisão

Considere dois pontos, x_A e x_B, os quais se encontram na superfície de decisão. Dado $y(x_A) = y(x_B) = 0$, tem-se que $w^T = (x_A - x_B) = 0$, e portanto, o vetor w é ortogonal a cada vetor localizado dentro da superfície e determina a orientação da superfície de decisão. Da mesma forma, se x é um ponto na superfície de decisão, então $y(x) = 0$. Logo a distância normal desde a origem até a superfície de decisão é dado por

$$\frac{w^T x}{|w|} = -\frac{w_o}{||w||}$$

Assim, o parâmetro de polarização w_0 determina a localização de decisão da superfície. Estas propriedades são ilustradas para o caso de $D = 2$ na Figura 13.

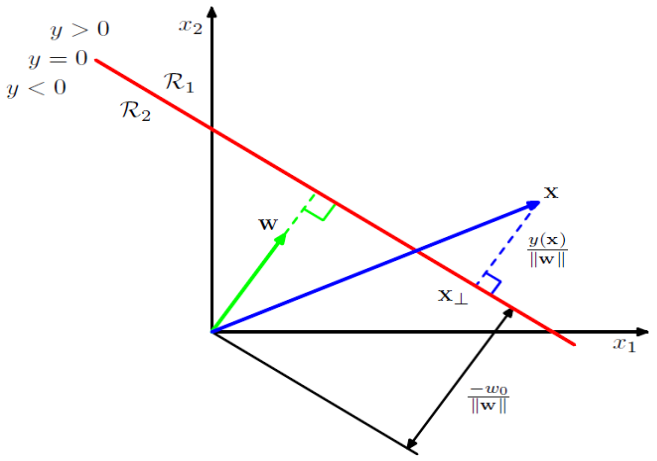


Figura 13 Ilustração geométrica de uma função discriminante linear em duas dimensões.

Note que, a superfície de decisão, mostrada em vermelho, é perpendicular a w, e o seu deslocamento desde a origem é controlado pelo parâmetro de polarização w_0. Além disso, a distância ortogonal assinado de um ponto x geral a partir

da superfície de decisão é dada por $\frac{y(x)}{||w||}$. Além disso, nota-se que o valor de $y(x)$ é a medida da distância r perpendicular do ponto x da superfície de decisão. Para melhor entendimento, considere um ponto arbitrário x e deixe $x_\perp$ ser sua projeção ortogonal sobre a superfície de decisão, de modo que

$$x = x_\perp + r\frac{w}{||w||}$$

Multiplicando ambos os lados deste resultado por w^T, adicionando w_0 e fazendo uso de $y(x) = w^T x + w_0$ e $y(x_\perp) = w^T x_\perp + w_0 = 0$, temos

$$r = \frac{y(x))}{||w||}.$$

Tal como acontece em modelos de regressão linear, é conveniente usar uma notação mais compacta em que se introduz um valor fictício $x_o = 1$ e em seguida definir $\widetilde{w} = (w_0, w)$ e $\widetilde{x} = (x_0, x)$ então temos

$$y(x) = \widetilde{w}^T \widetilde{x}$$

Neste caso, as superfícies de decisão são hiperplanas D-dimensional que passam através da origem do espaço de entrada expandida a $D + 1$-dimensional.

Para o caso em que se tem múltiplas categorias, pode-se tentar construir uma discriminante para k classes através da combinação de uma série de funções discriminantes de duas classes. Para isso, considere o uso de $k - 1$ classificadores para separar uma determinada classe C_K das demais, abordagem essa conhecida como um-contra-todos. O exemplo do lado esquerdo na Figura 14 mostra um exemplo que envolve três classes em que esta abordagem leva às regiões do espaço que são classificadas.

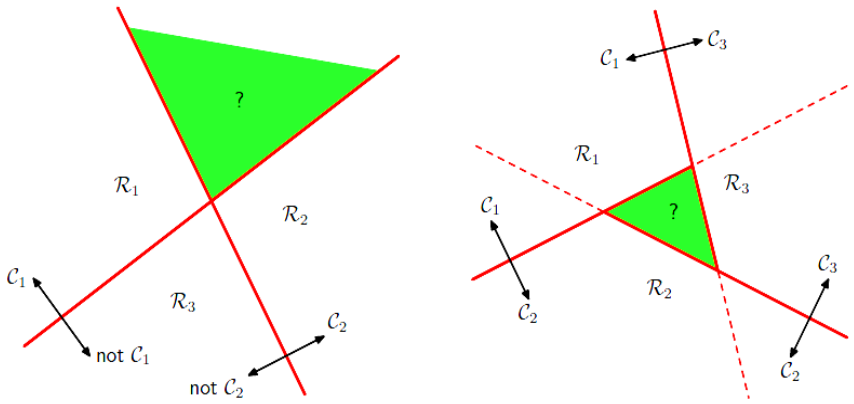


Figura 14 Regiões dos espaço a serem classificadas.

A tentativa de construir uma discriminante para a classe K a partir de um conjunto de duas classes leva a ambas regiões, mostrado em verde na Figura 14. À esquerda, um exemplo que envolve a utilização de dois discriminantes destinados a distinguir pontos na classe C_k a partir de pontos não presentes na classe C_k. À direita, está um exemplo que envolve três funções discriminantes de cada um que são usadas para separar um par de classes de C_k e C_j.

Uma alternativa consiste em introduzir $\frac{K(K-1)}{2}$ funções discriminantes binárias, uma para cada par possível de classes. Isto é conhecido como um classificador de um contra-um. Cada ponto é classificado de acordo com o voto da maioria entre as funções discriminantes. No entanto, isso também corre para o problema das regiões ambíguas, tal como ilustrado no diagrama da direita da Figura 14. Podemos evitar essas dificuldades, considerando uma única função discriminante linear às k classes da seguinte forma:

$$y(x) = w_k^T x + w_{k_0}$$

e, em seguida, atribuir um ponto x de classe C_k se $y_k(x) > y_j(x)$ para todo $j \neq k$. A fronteira de decisão entre a classe C_k e classe C_j é dada por $y_k(x) = y_j(x)$ e, portanto, corresponde a um hiperplano $D-1$-dimensional definido por

$$(w_k - w_j)^T x + (w_{k_0} - w_{j_o}) = 0.$$

As regiões de decisão de tal discriminante estão sempre ligadas individualmente e convexas. Para ver isto, considere dois pontos x_A e x_B ambos que se encontram dentro da região de decisão R_k, como ilustrado na Figura 15.

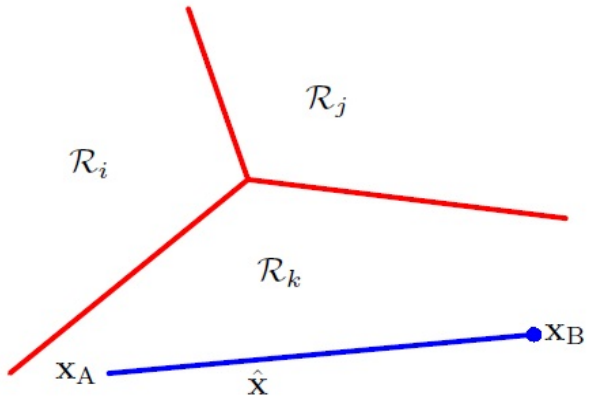


Figura 15 Ilustração das regiões de decisão para um discriminante linear multiclasse.

Na Figura 15, os limites de decisão são mostrados em vermelho. Se dois pontos, x_A e x_B, estão dentro da mesma região decisão R_k, então qualquer ponto $\hat{x}$ que fica na linha e liga esses dois pontos também devem estar em R_k, e, portanto, a região de decisão deve ser individualmente ligada e convexa.

Qualquer ponto $\hat{x}$ que se situa na linha de conexão x_A e x_B pode ser

expresso sob a forma

$$\hat{x} = \lambda x_A + (1 - \lambda) x_B$$

em que $0 \leq \lambda \leq 1$. A partir da linearidade das funções discriminantes, segue-se que

$$y_k(\hat{x}) = \lambda y_k(x_A) + (1 - \lambda) y_k(x_B).$$

Ambos os pontos, x_A e x_B, estão dentro de R_k. Logo, segue que $y_k(x_A) > y_j(x_A)$, e $y_k(x_B) > y_j(x_B)$, para todo $j \neq k$. Portanto, $y_k(\hat{x}) > y_j(\hat{x})$, e então $\hat{x}$ também se encontra em R_k. Assim R_k é individualmente ligada e convexa.

2.4.3 Mínimos quadrados para a classificação

Considere um problema geral de classificação com as classes K, com um esquema de codificação binária para o vetor alvo t. Uma justificação para a utilização de mínimos quadrados em tal contexto é que ele aproxima-se da esperança condicional $E[t|x]$ dos valores-alvo dado o vetor de entrada. Para o esquema de codificação binária, essa expectativa condicional é dada pelo vetor de probabilidades de classe posterior. Infelizmente, estas probabilidades são tipicamente mal aproximadas. Na verdade, as aproximações podem ter valores fora do intervalo $(0, 1)$, devido à flexibilidade limitada de um modelo linear.

Cada classe C_k é descrito pelo seu próprio modelo linear de modo a que

$$y_k(x) = {w_k}^T x + w_{k_0}$$

em que $k = 1, ..., K$. Convenientemente, pode-se agrupá-los usando a notação de vetor para que

$$y(x) = \widetilde{W}^T \widetilde{x}$$

em que $\widetilde{W}$ é uma matriz cuja coluna k^{th} compreende o vetor $D+1$-dimensional $\widetilde{w}_k = (w_{k_0}, {w_k}^T)^T$ e $\widetilde{x}$ é o correspondente vetor de entrada aumentado $(1, x^T)^T$ com uma *dummy* de entrada $x_o = 1$. Uma nova entrada x é então atribuída à classe para a qual a saída $y_k = \widetilde{w_k}^T \widetilde{x}$ é maior. Em seguida, determina-se a matriz parâmetro $\widetilde{W}$ minimizando a função de erro da soma de quadrados.

Considere um conjunto de dados de treinamento $\{x_n, t_n\}$ em que $n = 1, ..., N$, e defina uma matriz T cuja n^{th} linha é o vetor ${t_n}^T$, juntamente com uma matriz $\widetilde{X}$ cuja n^{th} linha $\tilde{x}_n^T$ é $\widetilde{x}_n^T$. A função de erro de soma de quadrados pode então ser escrito como

$$E_D(\widetilde{W}) = \frac{1}{2} Tr(\widetilde{X}\widetilde{W} - T)^T(\widetilde{X}\widetilde{W} - T)$$

Definindo a derivada em relação a $\widetilde{W}$ para zero e reorganizando. Em seguida, obtém-se a solução para, $\widetilde{W}$, sob a forma

$$\tilde{W} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}^T T = \widetilde{X}^\dagger T$$

em que $\widetilde{X}^\dagger$ é a pseudo-inversa da matriz $\widetilde{X}$. Dessa forma, obtém-se a função discriminante na forma

$$y(x) = \widetilde{W}^T\widetilde{x} = T^T(\widetilde{X}^\dagger)^T\widetilde{x}.$$

Uma propriedade interessante de soluções de mínimos quadrados com múltiplas variáveis-alvo é que, se cada vetor de destino no conjunto de treinamento satisfaz a restrição linear

$$a^T t_n + b = 0$$

para algumas constantes de a e b, o modelo de previsão para qualquer valor de x

satisfaz a mesma restrição, de modo que

$$a^T y(x) + b = 0.$$

Assim, se usarmos um esquema de codificação para as K classes, então as previsões feitas pelo modelo terão a propriedade de que os elementos de $y(x)$ irão resumir a 1 para qualquer valor de x. No entanto, esta soma de restrição por si só não é suficiente para permitir que o os resultados do modelo sejam interpretados como probabilidades, porque eles não são limitados a ficarem dentro do intervalo $(0, 1)$.

A abordagem dos mínimos quadrados dá uma solução de forma fechada e exata para os parâmetros da função discriminante. No entanto, mesmo com uma função discriminante (na qual podemos usá-la para tomar decisões diretamente e dispensar qualquer interpretação probabilística), pode surgir alguns problemas graves, como por exemplo, não ter robustez a *outliers* conforme ilustrado na Figura 16.

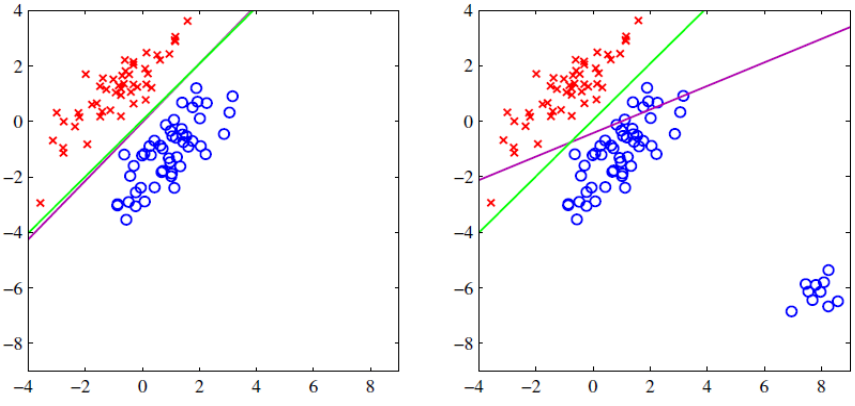


Figura 16 Exemplo de um conjunto de dados sintéticos com presença de outliers.

O gráfico da esquerda, na Figura 16, mostra os dados de duas classes, denotado por cruzes vermelhas e círculos azuis, juntamente com a fronteira de decisão encontrado por mínimos quadrados (curva roxa) e também pelo modelo de regressão logística (curva verde). A trama do lado direito mostra os resultados obtidos quando os pontos de dados são adicionados no canto inferior esquerdo do diagrama, mostrando que mínimos quadrados é altamente sensível a *outliers*, ao contrário de regressão logística. Os pontos de dados adicionais na figura da direita produzem uma mudança significativa na localização da fronteira de decisão, mesmo que estes pontos sejam corretamente classificados pelo limite de decisão original na figura do lado esquerdo. A função de erro de soma de quadrados penaliza previsões que são "muito correta"na medida em que eles se encontram num longo caminho no lado correto da fronteira de decisão.

No entanto, problemas com mínimos quadrados podem ser mais graves do que simplesmente a falta de robustez, tal como ilustrado na Figura 17.

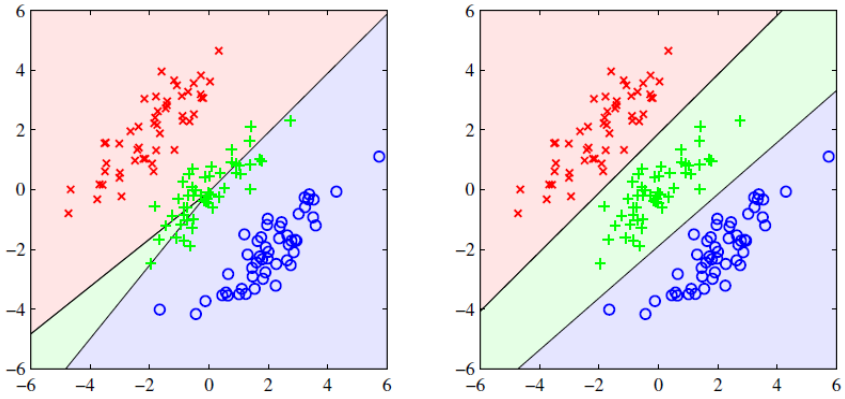


Figura 17 Exemplo de um conjunto de dados sintéticos que compreende três classes.

Na Figura 17, os pontos de dados de treinamento estão indicados em vermelho (x), verde $(+)$ e azul $(\circ)$. As linhas denotam os limites de decisão, e as cores de fundo denotam as respectivas classes das regiões de decisão. À esquerda, está o resultado do uso de um discriminante de mínimos quadrados. A região do espaço de entrada atribuído à classe verde é muito pequena e assim os pontos desta classe são classificados erroneamente. À direita, está o resultado do uso de regressão logística, mostrando a classificação correta dos dados de treinamento. Isso mostra um conjunto de dados sintéticos de três classes em um espaço de entrada bidimensional (x_1, x_2), que têm a propriedade de que limites lineares de decisão podem dar uma excelente separação entre as classes. Na verdade, a técnica de regressão logística dá uma solução satisfatória como pode ser visto no gráfico do lado direito. No entanto, a solução de mínimos quadrados dá resultados pobres, com apenas uma pequena região do espaço de entrada atribuída à classe verde.

O desempenho ruim dos mínimos quadrados não surpreende quando se lembra que corresponde à máxima verossimilhança sob a hipótese de uma condicional distribuição normal, enquanto que os vetores binários alvo têm uma distribuição que está longe de ser normal. Através da adoção de modelos probabilísticos mais adequados, obtém-se técnicas de classificação com melhores propriedades do que os mínimos quadrados. Para o momento, no entanto, é melhor explorar métodos alternativos não probabilísticos para definir os parâmetros nos modelos de classificação lineares.

2.4.4 Análise Discriminante quadrática de Fisher

A função discriminante quadrática é semelhante à função discriminante linear, exceto que, para classificar uma observação p-variada x na população π_i, considera-se o caso em que as matrizes de covariância das populações são dife-

rentes entre si, mas pressupõe-se que as densidades populacionais sejam conhecidas, em particular, que estas tenham distribuição gaussiana p-variada (FERREIRA, 2008).

Supondo que para a população π_1 a variável X tem distribuição normal com média μ_1 e que a população π_2 tem distribuição normal com média μ_2 e variâncias nos dois grupos iguais a σ^2, tem-se que, para cada valor de x, é possível calcular a razão entre as duas distribuições de probabilidades, chamada de razão de verosimilhança entre as duas populações, e definida por:

$$\lambda(x) = \frac{\text{função densidade de x na população 1}}{\text{função densidade de x na população 2}} = \frac{f_1(x)}{f_2(x)}$$

que no caso da distribuição normal, torna-se:

$$\lambda(x) = \frac{\frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2}\left(\frac{x-\mu_1}{\sigma}\right)^2\right\}}{\frac{1}{\sqrt{2\pi}\sigma} exp\left\{\frac{-1}{2}\left(\frac{x-\mu_2}{\sigma}\right)^2\right\}} = exp\left\{\frac{-1}{2}\left[\left(\frac{x-\mu_1}{\sigma}\right)^2 - \left(\frac{x-\mu_2}{\sigma}\right)^2\right]\right\}$$

Para uma observação x, quando $\lambda(x) > 1$, o valor da função densidade da população π_1 calculada para o respectivo valor da observação x é maior do que aquele obtido usando-se a distribuição da população π_2. Assim, pelo princípio da maior probabilidade, se $\lambda(x) > 1$, será razoável classificar a observação x como sendo da população π_1. Por outro lado, se $\lambda(x) < 1$, seria razoável classificá-lo como sendo pertencente da população π_2. Quando $\lambda(x) = 1$, as duas funções densidades têm numericamente o mesmo valor, podendo, portanto, classificar tanto como pertencente à população π_1 quanto à população π_2 (MINGOTI, 2007).

No entanto, uma regra de classificação razoável, seria aquela que minimizasse o ECM. Johnson e Wichern (2002) mostram que o ECM é minimizado

quando a seguinte regra de classificação é adotada:

$$\frac{f_1(x)}{f_2(x)} \geq \left[\frac{c(1|2)}{c(2|1)}\right]\left(\frac{p_2}{p_1}\right)$$

Nesse contexto, classifica-se x em π_1 se a desigualdade for verdadeira e, caso contrário, em π_2.

2.5 Avaliação dos métodos de classificação

Segundo Congalton (1991), uma das técnicas mais utilizadas para avaliar a acurácia na classificação de dados é a utilização da Tabela Verdade, que pode ser usada como ponto de partida para uma série de técnicas de Aprendizado de Máquinas e possibilita a verificação do desempenho de um algoritmo. Essa tabela é uma forma de apresentar a classe real e a classe prevista pelo modelo, como mostra a Tabela 6.

Tabela 6 Tabela Verdade

Predito	Real		
	Positivo	Negativo	
Positivo	Verdadeiro Positivo (VP)	Falso Positivo (FP) Erro tipo I (α)	Preditos Positivos
Negativo	Falso Negativo (FN) Erro tipo II (β)	Verdadeiro Negativo (VN)	Preditos Negativos

A taxa de Falso Positivo (FP) é equivalente à taxa de erro tipo I, que consiste em rejeitar a hipótese apresentada sobre determinado fato, quando a mesma é verdadeira. A probabilidade de cometer um erro do tipo I num teste de hipóteses é denominada significância do teste e é representada por α. Da mesma forma, a taxa de Falsos negativos (FN) é equivalente à taxa de erro tipo II, representada por β e ocorre quando a análise estatística não rejeita uma hipótese, quando esta hipótese

é falsa. A taxa de Verdadeiro Positivo (VP) é definida como o poder do teste, dado por $1 - \beta$, e equivale à sensibilidade (SENS) do modelo. E a taxa de Verdadeiro negativo (VN) que é equivalente à especificidade (ESPEC) é dada por $1 - \alpha$.

A sensibilidade e especificidade são conhecidas em Aprendizado de Máquina como função de classificação e são calculadas da seguinte forma:

$$\text{SENS} = \frac{VP}{VP + FN}$$

ou seja, a sensibilidade mede a proporção de positivos reais, devidamente identificados.

A especificidade mede a proporção de verdadeiros negativos, isto é, avalia a capacidade de um indivíduo não pertencer a uma classe dado que este realmente não pertence.

$$\text{ESPEC} = \frac{VN}{VN + FP}$$

Deste modo, a proporção de observações que foram classificadas corretamente pelo classificador em um conjunto de dados, isto é, a acurácia (ACC), é dada por:

$$\text{ACC} = \frac{VP + VN}{VP + FP + VN + FN}$$

Por fim, a taxa de erro total de classificação é dada pelo complementar da proporção de observações que foram devidamente classificadas, ou seja,

$$Erro = \frac{FP + FP}{VP + FP + VN + FN}$$

Em estatística, outra forma de avaliar o desempenho de um algoritmo de classificação é dada pela Característica de Operação do Receptor (COR), ou *Receiver Operating Characteristic* (ROC), ou simplesmente curva ROC, que é uma representação gráfica que ilustra o desempenho (ou performance) de um sistema classificador binário e como o seu limiar de discriminação é variado. A Curva ROC é uma representação da sensibilidade da predição do modelo versus o complemento da especifidade. Porém, uma das principais desvantagens de se utilizar gráficos ROC é a sua limitação para apenas duas classes. A extensão das curvas ROC para problemas de classificação multiclasse sempre foi complicado, pois os graus de liberdade aumenta quadraticamente com o número de classes, e o espaço ROC tem $c(c-1)$ dimensões, onde c é o número de classes. Diante disso, neste trabalho, o desempenho dos métodos de classificação será analisado através das taxas de erros.

3 MATERIAIS E MÉTODOS

3.1 Dados Simulados

O conjunto de dados obtido por simulação, como apresentado por Culp et al. (2006), será utilizado para analisar a aplicação dos métodos de classificação em problemas multiclasse. A base de dados é composta por um total de 1.200 observações geradas a partir de uma distribuição normal padrão. Tais observações eram descritas por $X_1, ..., X_{10}$ e uma variável resposta, na qual foi construída da seguinte forma:

$$Y = \begin{cases} 1, & \text{se} \quad \sum_{r=1}^{p} {X_r}^2 \leq {\chi_{10}}^2(0,33) \\ 2, & \text{se} \quad {\chi_{10}}^2(0,33) < \sum_{r=1}^{p} {X_r}^2 \leq {\chi_{10}}^2(0,66) \\ 3, & \text{se} \quad \sum_{r=1}^{p} {X_r}^2 > {\chi_{10}}^2(0,66) \end{cases}$$

Os valores $\chi^2_{10}(0,33)$ e $\chi^2_{10}(0,66)$ são equivalentes aos valores tabelados da distribuição Qui-quadrado, dados respectivamente por 7,58 e 11,23. Os quantis $0,33$ e $0,66$ são pontos estabelecidos em intervalos a partir da função distribuição acumulada, e foram utilizados para dividir a amostra e assim determinar os limites entre os subconjuntos das classes.

3.2 Dados reais

O conjunto de dados reais é referente a um experimento sensorial da qualidade de quatro tipos distintos de cafés especiais produzidos na Serra da Mantiqueira. O ponto de torra foi determinado visualmente, utilizando o sistema de classificação de cor por meio de discos padronizados (SCAA/AgtronRoast Color Classification System). Em relação ao preparo da bebida, utilizou-se água filtrada pronta para consumo, livre de qualquer contaminante e sem adição de açúcar.

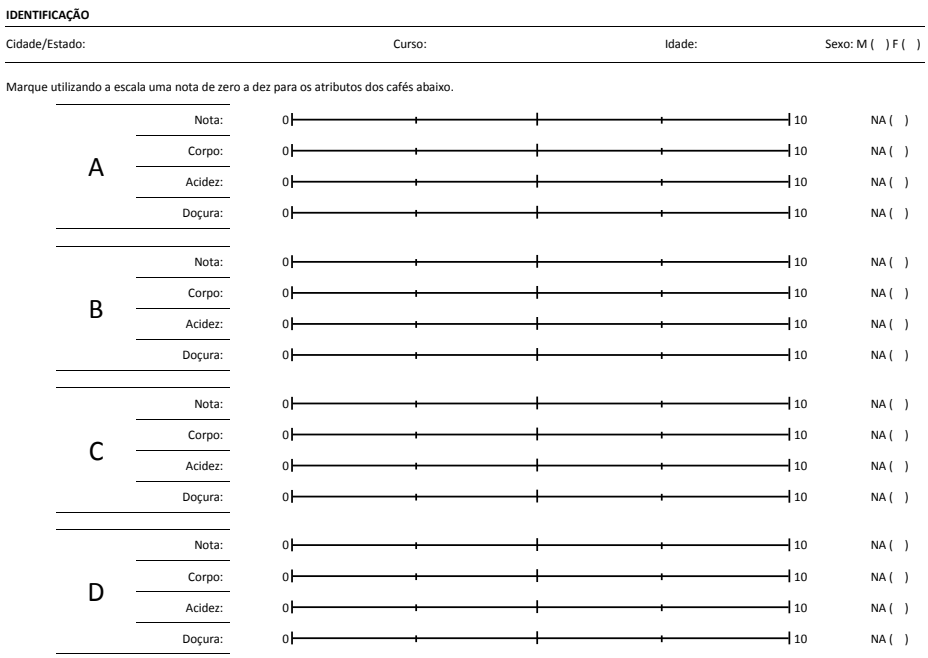
Com estas especificações, os quatro tipos de cafés especiais foram codificados nas amostras por A, B, C e D dada a descrição na Tabela 7.

Tabela 7 Descrição dos cafés especiais avaliados na análise sensorial.

Tipo	Genótipo	Altitude	Processamento
A	Bourbon Amarelo	acima de 1.200 m	Natural
B	Acaiá	abaixo de 1.100 m	Cereja descascado
C	Acaiá	abaixo de 1.100 m	Natural
D	Bourbon Amarelo	acima de 1.200 m	Cereja descascado

Fonte: Liska (2015)

Para cada tipo de café, foram avaliadas as seguintes características sensoriais no teste de aceitação: aroma, corpo, dureza e nota final. A cada provador foi apresentado um questionário, como mostra a Figura 18, no qual eram atribuídas notas para formar uma base de dados.



Fonte: Liska (2015)

Figura 18 Questionário da pesquisa de análise sensorial de café

O experimento foi realizado na Universidade Federal de Lavras contando com a participação de um grupo de provadores voluntários que receberam um treinamento básico em relação às avaliações sensoriais de café e um outro grupo que não recebeu qualquer tipo de treinamento. Porém, como os indivíduos eram voluntários, não foi possível impedir a participação de pessoas com idade acima de 50 anos. Dessa forma, a fim de assegurar que a condição da idade dos provadores fosse satisfeita, os indivíduos foram classificados em treinados e não treinados respeitando a faixa etária. A categorização da variável resposta foi dada da seguinte forma:

T_1: provadores treinados com idade entre 19 e 50 anos;

T_2: provadores não treinados com idade entre 19 e 50 anos;

T_3: provadores não treinados com idade acima de 50 anos.

Após classificar os indivíduos em treinados e não treinados com faixa etária dentro das características dos perfis de provadores, foi verificado se os consumidores apresentam alguma habilidade sensorial advinda de um treinamento para fazer a discriminação de quatro tipos de cafés especiais.

Vale ressaltar que o banco de dados para a discriminação dos provadores e a discriminação dos cafés especiais é o mesmo. Porém, a forma como os dados foram dispostos na tabela diferenciam em dois cenários. A Tabela 8 relaciona os provadores bem como as características sensoriais avaliadas por cada provador, em que a_{ij} representa a nota dada pelos provadores treinados com idade entre 19 e 50 anos, b_{ij} nota dada pelos provadores não treinados com idade entre 19 e 50 anos e c_{ij} nota dos provadores não treinados acima de 50 anos, tal que n_1, n_2 e n_3 são os números de indivíduos que compõem cada classe. Já a Tabela 9 é referente à entrada dos dados para a discriminação dos cafés especiais.

Tabela 8 Representação tabular para a classificação dos provadores.

		Atributo Sensorial 1				$\cdots$	Atributo Sensorial 4			
Condição	Provador	A	B	C	D	$\cdots$	A	B	C	D
T_1	1	a_{11}	a_{12}	a_{13}	a_{14}	$\cdots$	a_{113}	a_{114}	a_{115}	a_{116}
	2	a_{21}	a_{22}	a_{23}	a_{24}	$\cdots$	a_{213}	a_{214}	a_{215}	a_{216}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_1	$a_{n_1 1}$	$a_{n_1 2}$	$a_{n_1 3}$	$a_{n_1 4}$	$\cdots$	$a_{n_1 13}$	$a_{n_1 14}$	$a_{n_1 15}$	$a_{n_1 16}$
T_2	1	b_{11}	b_{12}	b_{13}	b_{14}	$\cdots$	b_{113}	b_{114}	b_{115}	b_{116}
	2	b_{21}	b_{22}	b_{23}	b_{24}	$\cdots$	b_{213}	b_{214}	b_{215}	b_{216}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_2	$b_{n_2 1}$	$b_{n_2 2}$	$b_{n_2 3}$	$b_{n_2 4}$	$\cdots$	$b_{n_2 13}$	$b_{n_2 14}$	$b_{n_2 15}$	$b_{n_2 16}$
T_3	1	c_{11}	c_{12}	c_{13}	c_{14}	$\cdots$	c_{113}	c_{114}	c_{115}	c_{116}
	2	c_{21}	c_{22}	c_{23}	c_{24}	$\cdots$	c_{213}	c_{214}	c_{215}	c_{216}
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	n_3	$c_{n_3 1}$	$c_{n_3 2}$	$c_{n_3 3}$	$c_{n_3 4}$	$\cdots$	$c_{n_3 13}$	$c_{n_3 14}$	$c_{n_3 15}$	$c_{n_3 16}$

Tabela 9 Representação tabular para a classificação dos cafés especiais.

Condição	Provador	Nota final	Corpo	Aroma	Doçura	Tipo de café
	1	a_{11}	a_{12}	a_{13}	a_{14}	A
	2	a_{21}	a_{22}	a_{23}	a_{24}	B
	3	a_{31}	a_{32}	a_{33}	a_{34}	C
	4	a_{41}	a_{42}	a_{43}	a_{44}	D
T_1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$n_1 - 3$	$a_{(n_1-3)1}$	$a_{(n_1-3)2}$	$a_{(n_1-3)3}$	$a_{(n_1-3)4}$	A
	$n_1 - 2$	$a_{(n_1-2)1}$	$a_{(n_1-2)2}$	$a_{(n_1-2)3}$	$a_{(n_1-2)4}$	B
	$n_1 - 1$	$a_{(n_1-1)1}$	$a_{(n_1-1)2}$	$a_{(n_1-1)3}$	$a_{(n_1-1)4}$	C
	n_1	$a_{n_1 1}$	$a_{n_1 2}$	$a_{n_1 3}$	$a_{n_1 4}$	D
	1	b_{11}	b_{12}	b_{13}	b_{14}	A
	2	b_{21}	b_{22}	b_{23}	b_{24}	B
	3	b_{31}	b_{32}	b_{33}	b_{34}	C
	4	b_{41}	b_{42}	b_{43}	b_{44}	D
T_2	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	$n_2 - 3$	$b_{(n_2-3)1}$	$b_{(n_2-3)2}$	$b_{(n_2-3)3}$	$b_{(n_2-3)4}$	A
	$n_2 - 2$	$b_{(n_2-2)1}$	$b_{(n_2-2)2}$	$b_{(n_2-2)3}$	$b_{(n_2-2)4}$	B
	$n_2 - 1$	$b_{(n_2-1)1}$	$b_{(n_2-1)2}$	$b_{(n_2-1)3}$	$b_{(n_2-1)4}$	C
	n_2	$b_{n_2 1}$	$b_{n_2 2}$	$b_{n_2 3}$	$b_{n_2 4}$	D

Um ponto importante a ser destacado é que apenas os indivíduos treinados e com idade entre 19 e 50 anos foram utilizados para a discriminação dos tipos de cafés especiais. Dado que, por falta de conhecimento básico da qualidade do café, as notas atribuídas às características sensoriais pelos provadores não treinados não serviram como base para verificar diferentes nuances especiais da bebida.

3.3 Treinamento e Teste

Neste trabalho, o foco concentra-se no aprendizado supervisionado com o propósito de estimular seu uso na realização de pesquisas em áreas do conhecimento relacionadas ao problema da classificação de dados. Portanto, os conjuntos de dados foram divididos em conjunto de treinamento, destinada aos ajustes dos classificadores, e conjunto de teste, destinado à validação dos classificadores.

A divisão do conjunto total de dados em dois subconjuntos mutuamente exclusivos foi realizada através do Método Houdout. A Tabela 10 apresenta o número de observações que compõem cada subconjunto.

Tabela 10 Partição do conjunto total de dados em treinamento e teste.

Conjunto	Simulado	Reais	
		Provadores	Cafés
Treinamento (2/3)	800	92	466
Teste (1/3)	400	46	234
Total de observações	1200	138	700

Para fins comparativos, o desempenho dos classificadores será avaliado utilizando os mesmos conjuntos de treinamento e teste.

3.4 Implementação dos algoritmos

3.4.1 AdaBoost para classificação multi-classe

A implementação do Boosting foi realizada através da função *boosting* do pacote *adabag* do *software* R (CORE TEAM, 2014). A configuração do AdaBoost.M1 tem a mesma fundamentação teórica do Algoritmo Boosting descrita na seção 2.3.1.

Algoritmo 2. AdaBoost.M1

---

1. Conjunto de dados: $(x_1, y_1), ..., (x_N, y_N)$ em que $x_i \in X$ e Y

   Inicie os pesos ${w_i}^{(m)} = \frac{1}{N}$, para $i = 1, 2, ..., N$

2. Para $m = 1, ..., M$

   i) Selecione $f_m(x)$ para minimizar o erro ponderado:

   Se $\epsilon_m \geq \frac{1}{2}$, então $M = m - 1$;

   ii) Escolha $c_m = \frac{1}{2} ln \left(\frac{1-\epsilon_m}{\epsilon_m} \right)$

   iii) Atualize, para $i = 1, .., m$:

$$ {w_i}^{(m+1)} = \frac{{w_i}^{(m)}}{Z_m} \begin{cases} e^{-c_m}, & \text{se} \quad y_i = f_m(x_i), \\ e^{c_m}, & \text{se} \quad y_i \neq f_m(x_i). \end{cases} $$

   e Z_m é o fator de normalização.

3. A Classificação final, $F(x)$:

$$ F(x) = argmax \sum_{m=1}^{N} c_m \{ f_m(x) = y \} $$

3.4.2 Bagging

Para a aplicação do Bagging, foi utilizado o procedimento descrito no Algoritmo 3. É importante destacar que os passos gerados foram baseados na teoria descrita na seção 2.3.2. Dessa forma, foram geradas 25 amostras *bootstrap* e o classificador base utilizado foi a Árvore de classificação. O classificador final via Bagging foi dado pela classe mais frequente determinada pelos classificadores parciais. A implementação do Bagging foi realizada utilizando diretamente a função *bagging* do pacote *ipred* do *software* R (CORE TEAM, 2014).

Algoritmo 3. Bagging

---

1. Construa uma amostra *bootstrap* $L^* = \{x_i, y_i\}_{i=1}^N$.

2. Compute o preditor via *bootstrap* $\hat{d}_n(x)$, utilizando o mesmo procedimento para construir $\hat{d}_n(x)$, mas com a amostra L^*.

3. O preditor via Bagging é $\hat{d}_{n;B} = E^*[\hat{d}_n^*(x)]$ no caso de regressão. Em classificação, a classe predita pelo classificador via Bagging é a mais votada pelos classificadores $\hat{d}_n^*(x)$. Ou seja,

$$\hat{d}_{n;B} = argmax N_j$$

em que, $N_j \# \{k; \hat{d}_{n;k}^*(x) = j\}$

3.5 Avaliação dos Métodos de classificação

A avaliação da classificação das observações, para cada problema exposto, será dada pelas Tabelas Verdades disponíveis no anexo. A partir destas Tabelas, os métodos de classificação serão avaliados por meio das taxas de erro estimadas.

4 RESULTADOS E DISCUSSÃO

Neste capítulo, serão apresentados os resultados obtidos nas aplicações dos métodos estatísticos tradicionais e dos métodos de Aprendizado de Máquina. Inicialmente, serão apresentados as aplicações de cada método nos dados simulados e em seguida, as aplicações na análise sensorial.

4.1 Aplicação dos métodos de classificação aos dados simulados

A aplicação dos métodos de classificação nos dados simulados foi feita utilizando a divisão em conjunto de treinamento e teste discutida anteriormente. As taxas de erro estimadas estão apresentados na Tabela 11.

Tabela 11 Taxa de erro global obtida pelos métodos de classificação do conjunto de dados simulados.

Métodos de classificação	Taxa de erro (%)
Boosting	26,25
Bagging	35,25
Análise Discriminante Quadrática	29,0
Análise Discriminante Linear	68,5

Nota-se, a partir dos resultados apresentados na Tabela 11, que a Análise Discriminante Linear resultou em um desempenho inferior aos resultados obtidos pela Análise Discriminante Quadrática, supostamente, tal resultado pode ser justificado pela falta de linearidade dos dados analisados. Quando observado os métodos de Aprendizado de Máquina, os menores índices de erros de classificação foram obtidos pelo Boosting. Para este exemplo, os resultados sugerem que o Bagging não produz um resultado especialmente bom, sendo inferior à Análise Discriminante Quadrática. Porém, a taxa de erro indica ser viável a utilização do Bagging como forma de classificação de dados em problemas multiclasse.

Dado que, neste estudo de simulação considerou-se variáveis geradas por

uma distribuição normal, portanto, em uma escala contínua, em algumas aplicações realizadas por diferentes autores, os resultados foram concordantes e serão mencionados a seguir.

Rubesam (2004) aplicou os métodos Boosting e Bagging a um conjunto de dados relativo a clientes de uma loja de varejo. O objetivo era obter uma regra que permitisse classificar futuros clientes em um de 10 grupos de maneira precisa e rápida. Através das taxas de erro, concluiu-se que o algoritmo AdaBoost e o Bagging foram superiores ao método clássico de Análise Discriminante Linear. Outro ponto importante por ele observado é que os métodos automáticos são úteis mesmo para problemas extremamente complexos de classificação (no caso testado havia 160 variáveis preditoras e 10 classes).

Um estudo realizado para verificar a eficiência do Bagging foi feito por Granatyr (2011), o autor aplicou a técnica de combinação de classificadores a métodos baseados em análise formal de conceitos, que tem como principal objetivo a identificação de estruturas conceituais de relacionamento entre dados. O Bagging foi utilizado a fim de se obter melhores taxas de acerto e concluiu-se que este método produziu resultados superiores que os demais classificadores combinados.

Para discriminar pacientes com presença/ausência de doença cardíaca coronariana, Liska (2012) ajustou o algoritmo Boosting e obteve as melhores taxas de acurácia, sensibilidade, especificidade, taxas de falsos positivos e taxas de falsos negativos quando comparadas com as taxas obtidas pelo modelo de regressão logística com seus parâmetros estimados via máxima verossimilhança.

Pereira (2014) também realizou uma comparação do Boosting com os métodos tradicionais de classificação através da alocação de pacientes em grupos com presença e ausência de uma doença da coluna vertebral. A partir da aplicação foi possível concluir que o Algoritmo Boosting apresentou melhores resultados em

relação aos resultados obtidos nas Análises Discriminante Linear e Quadrática.

4.2 Aplicação dos métodos de classificação à Análise Sensorial

4.2.1 Classificação dos grupos de provadores

Uma vez obtida a Tabela Verdade de cada método de classificação, foram obtidas as taxas de erro para verificar o desempenho dos classificadores. Portanto, a Tabela 12 apresenta os métodos de classificação e suas respectivas taxas de erros.

Tabela 12 Taxa de erro global obtida pelos métodos de classificação para a discriminação dos grupos de provadores.

Métodos de classificação	Taxa de erro (%)
Boosting	19,56
Bagging	28,26
Análise Discriminante Quadrática	36,95
Análise Discriminante Linear	45,65

Em concordância com os resultados obtidos na Tabela 12, pode-se concluir que é perceptível que os métodos de Aprendizado de Máquina apresentam resultados superiores aos métodos estatísticos tradicionais, pois as taxas de erro de classificação obtidas pelo Boosting e Bagging foram menores. Note também, que o segundo melhor método de Aprendizado de Máquina teve taxa de erro igual a 28,26% contra 36,95% do primeiro melhor método tradicional. Ou seja, os métodos Boosting e Bagging de fato apresentam um alto poder discriminatório dos provadores envolvendo análise sensorial de cafés especiais respeitando as faixas etárias.

Um estudo na análise sensorial foi realizado por Liska et al. (2015) considerando o mesmo conjunto de dados usado nesta dissertação. Porém, o objetivo principal era discriminar grupos de provadores treinados e não treinados. A aná-

lise considerava o problema de classificação binária, em que a variável resposta era dada pelas categorias de consumidores com a presença ou ausência de treinamento básico desconsiderando a idade dos provadores. O método Boosting foi aplicado com sucesso e aumentou em 23,96% a capacidade do classificador LDA em discriminar corretamente os provadores treinados.

4.2.2 Classificação dos cafés especiais

A Tabela 13 resume o poder de discriminação dos cafés especiais de cada método de classificação.

Tabela 13 Taxa de erro global obtida pelos métodos de classificação para a discriminação dos cafés especiais.

Métodos de classificação	Taxa de erro (%)
Boosting	24,79
Bagging	25,64
Análise Discriminante Linear	37,72
Análise Discriminante Quadrática	38,04

Observe que a maior taxa de erro obtida foi pela Análise Discriminante Linear. Barbosa et. al (2014) encontrou taxa de erro próxima na discriminação de tipos de processamento de cafés especiais considerando diferentes isótopos estáveis em sementes dos cafés especiais. A taxa de erro de classificação obtida pela Análise Discriminante Linear ao analisar a cor da semente foi de 31,3%.

Ahmad et. al (1999) realizaram a separação automática de grãos de café por visão computacional. O problema consistia em diferenciar quatro tipos de grãos, sendo três deles caracterizados pelo tamanho e o quarto, podendo ser definido por tamanho ou cor. A classificação foi realiza com algoritmos de processamento de imagens para detecção dos objetos e extração de características, classifi-

cação por padrões de cor, definição de descritores de forma. O resultado foi uma taxa de erro de 21,68%.

Pereira et. al (2011) utilizou a técnica multivariada discriminante com o objetivo de comparar a qualidade sensorial do café do Banco de Germoplasma de Café de Minas Gerais pertencentes aos grupos Bourbon e Híbrido de Timor. Seis acessos de Bourbon Amarelo, sete de Bourbon Vermelho e nove de Híbrido de Timor foram avaliados de acordo com os critérios da "Brazil Speciality Coffee Association - BSCA". Com base na Análise Discriminante, observou-se alto erro aparente, evidenciando que os grupos são bem similares e a classificação final dos acessos podia ser confundida.

Para realizar a discriminação da qualidade do Bourbon Amarelo produzido em diferentes altitudes (abaixo de 1000 m, entre 1000 e 1200 m e acima de 1200 m) de acordo com diferentes pontuações obtidas através de avaliações sensoriais realizadas por provadores treinados qualificados, como os juízes de cafés especiais, Ramos et. al (2015) utilizaram árvores de decisão a fim de agregar resultados que permitam ao pesquisador tomar decisões em um conjunto de dados multivariados de grande dimensão. Analisando a precisão dos modelos com base nas faixas de altitude avaliadas, os autores observaram que os resultados encontrados para as altitudes acima de 1100 m exibiram valores de precisão perto de 80% e que esses valores são considerados aceitáveis no que diz respeito a corrigir as taxas de classificação para os modelos obtidos com o método CHAID.

5 CONCLUSÕES

Concluiu-se que os métodos automáticos de classificação baseados na combinação de classificadores, Boosting e Bagging, resultam em erro de classificação inferior à Análise Discriminante Linear (LDA) e Quadrática (QDA), podendo ser considerados como uma eficiente forma de classificação de dados em problemas multiclasse.

Em se tratando da aplicação, as classificações em dados sensoriais apresentadas neste trabalho, permitiram ilustrar a capacidade do Boosting e do Bagging em captar pequenas diferenças entre as amostras de cafés especiais e, dessa forma, mostrar habilidade em realizar a tarefa de classificação de forma simples e eficaz. Portanto, foi possível concluir que o objetivo principal proposto nesta dissertação foi alcançado, ou seja, os métodos automáticos de classificação podem ser aplicados a problemas multiclasses em dados de natureza Sensorial.

REFERÊNCIAS

AHMAD, I. S. et al. Color classifier for symptomatic soybean seeds using image-processing. **Plant Disease**, Saint Paul, v. 83, n. 4, p. 320-327, Apr. 1999.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. Análise sensorial dos alimentos e bebidas: terminologia. Rio de janeiro: ABNT, 1993

AVELINO, J. et al. Effects of spole exposure, altitude and yield on coffee qualityin two altitude terroirs of Costa Rica, Orosi and Santa María de Dota. **Journal of the Science of Food and Agriculture**, London, v.85, n.11, p.1869-1876, 2005.

BARBOSA, J. et al. Discrimination of production environments of specialty coffees by means of stable isotopes and discriminant model. **Journal of Agricultural Science**, Cambridge, v.6, n.5, p. 55-64, 2014.

BISHOP, M. C. **Pattern recognition and machine learning**. New York: Springer-Verlag, 2006.

BORÉM, F. M. et al. Qualidade do café natural e despolpado após secagem em terreiro e com altas temperaturas.**Ciência e Agrotecnologia**, Lavras, v. 32, n. 5, p. 1609-1615, set./out. 2008.

BRAZIL SPECIALTY COFFEE ASSOCIATION. **Colômbia fará concurso de cafés especiais com metodologia brasileira**. 2008. Disponível em: <http://www.bsca.com.br>. Acesso em: 12 jan. 2015.

CARVALHO, L. A. V. **Data mining**, 2.ed, São Paulo: Erica, 2001.

CHAVES, J. B. P. **Avaliação sensorial de alimentos: métodos de análise**. Viçosa: Ed. UFV, 1980,

CONGALTON, R. G. A review of assessing the accuracy of classification of remotely sensed data. **Remote Sensing of the Environment**, v. 37, n. 1, p. 35-46, Jul.1991.

CONDUTA, B.C; MAGRIN, D.H. **Aprendizagem de Máquina**, Dissertação (Mestrado em Tecnologia) - Universidade Estadual de Campinas, Limeira, 2010.

CULP, M; JOHNSON, K; MICHAILIDIS, G. Ada: an R package for stochastic boosting. **Journal of Statistical Software**, [S.l.], v. 17, n. 2, p. 1-27, Oct. 2006.

BREIMAN, L. Bagging predictors. **Machine Learning**, v. 24, n. 2, p. 123-140, 1996.

COSTEL, E.; DURAN, L. El análisis sensorial en el control de calidad de los alimentos, IV. Realización y análisis de los datos, **Revista de Agroquímica y Tecnología de Alimentos**, Valencia, v. 22, n. 1, p. 1-21, mar. 1982.

DAISTER, L. P. **Estratégias para desenvolvimento de sistemas de múltiplos classificadores em aprendizado supervisionado**. 2007. 98 p. Dissertação (Mestrado em Ciências em Engenharia Civil)-Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007.

DELLA LUCIA, S. M.; MININ, V. P. R. M.; CARNEIRO, J. D. S. Análise sensorial de alimentos, In: MININ, V. P. R. M. **Análise sensorial**: estudos com consumidores. Viçosa, MG: Ed. UFV, 2006.

FERREIRA, D. F. **Estatística Multivariada**, 2. ed. Lavras: ED. UFLA, 2008.

FISHER, R. A. The use of multiple measurements in taxonomic problems. **Annals of Eugenics**, London, v.7, p. 179-188, 1936.

FREUND, Y.; SCHAPIRE, R. E. A decision-theoretic generalization of on-linelearning

and an application to boosting. **Journal of Computer and System Sciences**, New York, v. 55, n. 1, p. 119-139, Aug. 1997.

FREUND, Y.; SCHAPIRE, R. E. A short introduction to boosting. **Japonese Society for Artificial Intelligence**, Tokyo, v.14, n.5, p. 771-780, Sept. 1999.

FREUND, Y.; SCHAPIRE, R. E. Experiments with a new Boosting algorithm. In:INTERNATIONAL CONFERENCE ON MACHINE LEARNING, 13.,1996, Bari. **Proceedings**...Bari: IMLS. 1996.

GRANATYR, J. **Descoberta de regras de classificação utilizando análise formal de conceitos**. 2011. 61 p. Dissertação (Mestrado em Agentes de Software)-PontifíciaUniversidade Católica do Paraná, Curitiba, 2011.

ILLY, E. A saborosa complexidade do café. **Scientific American**, New York, v. 286, n. 6, p. 48-53, 2002.

JOHNSON, R. A; WICHERN, D. W. **Applied Multivariate Statistical Analysis**, New Jersey, Prentice-Hall, 1999.

KHATTREE, R; NAIK, D. N. **Multivariate data reduction and discrimination with SAS software**, Cary: SAS Institute, 2000.

KOHAVI, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELIGENCE, 14. San Francisco. **Proceedings**... San Francisco: Morgan Kaufmann. p. 1137-1145, 1995.

LANCHOTE, L. N. **Estudos com mapas de preferência: associação com Procrustes e construção com valores faltantes**. 2007. 75 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária)-Universidade Federal de Lavras,

Lavras, 2007.

LISKA, G. R. **Classificação de dados em modelos com resposta binária via algoritmo Boosting e regressão logística**. 2012. 105 p. Dissertação (Mestrado em Estatística e Experimentação Agropecuária) - Universidade Federal de Lavras, Lavras, 2012.

LISKA, G. R. et al. Evaluation of sensory panels of consumers of specialty coffee beverages using the boosting method in discriminant analysis. **Semina: Ciências Agrárias**, Londrina, v.36, n.6, p. 3671-3680, nov./dez. 2015.

LORENA, A. C.; CARVALHO, A. C. P. L. F. Uma introdução às support vector-machines. **Revista de Informática Teórica e Aplicada**, Santo André, v. 14, n. 2,p. 43-67, 2007.

LUGER, G. F. **Inteligência artificial**: estruturas e estratégias para a solução de problemas complexos. 4.ed. Porto Alegre: Bookman, 2004.

MAMEDE, M. E. O. et al. Sensory and chemical evaluation of decaffeinated soluble coffee. **Alimentos e Nutrição**, Araraquara, v. 21, n. 2, p. 311-324, 2010.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**. Belo Horizonte: Ed. UFMG, 2007.

MINIM, V. P. R. **Análise sensorial**: estudos com consumidores. Viçosa, MG: Ed. UFV,2006.

MITCHELL, T. **Machine Learning**. New York: McGraw-Hill, 1997.

MORAES, M. A. C.**Métodos para avaliação sensorial dos alimentos**. 6.ed.Campinas: Ed. Unicamp, 1988.

PAIVA, E. F. F. **Análise Sensorial dos cafés especiais do Estado de Minas Gerais**. 2005. 55 p. Dissertação (Mestrado em Ciências dos Alimentos) - Universidade Federal de Lavras, Lavras, 2005.

PASSOS, U. R. C. **Computação evolutiva e aprendizado de máquina aplicadosao apoio do diagnóstico da cardiopatia isquêmica**. 2014. 83 p. Dissertação (Mestrado em Pesquisa Operacional e Inteligência Computacional) - Universidade Federal Candido Mendes, Rio de Janeiro, 2014.

PEREIRA, A. P. et. al. Caracterização da qualidade de bebida e outras características de acessos do banco de germoplasma de café de Minas Gerais. In: SIMPÓSIO DE PESQUISA DOS CAFÉS DO BRASIL, 7., 2011, Araxá. **Anais**... Araxá: Consórcio Pesquisa Café, 2011.

PEREIRA, E. A.; PEREIRA, T. M. Alocação de pacientes em grupos via Boosting: uma comparação com métodos tradicionais de classificação. **Revista da Estatística UFOP**, Ouro Preto, v.3, n. 3, p. 814-819, 2014.

PRATI, R. C. **Novas abordagens em aprendizado de máquina para a geraçãode regras, classes desbalanceadas e ordenação de casos**. 2006. 191 p. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Universidade de São Paulo, São Carlos, 2006.

OLIVEIRA, A. F. **Análise sensorial dos alimentos**. Londrina: Universidade Tecnológica Federal do Paraná, 2010. Apostila do Curso de Tecnologia em Alimentos.

OLIVEIRA, F.A; NASCIMENTO, J. C. **Algoritmos de aprendizado de máquina para tarefas de classificação morfossintática**. 2012. Monografia (Graduação em Engenharia de Computação) - Instituto Militar de Engenharia, Rio de Janeiro, 2012.

RAMOS, M. F. et al. Discrimination of the sensory quality of the Coffea arabica L. (cv. Yellow Bourbon) producedin different altitudes using decision trees obtained by the CHAID metod. **Journal of the Science of Food and Agriculture**, London, Nov. 2015.Eprint.

R CORE TEAM. **R**: a languageandenvironment for statisticalcomputing. Vienna: R Foundation for StatisticalComputing, 2014. Disponível em: <http://www.R-project.org/>. Acesso em: 15 jan. 2015.

RICH, E; KNIGHT, K. **Inteligência artificial**. 2.ed. Rio de Janeiro: Makron Books, 1994.

RODRIGUES, M. A. S. **Árvores de classificação**. 2005. 34 p. Monografia (Graduação em Matemática)-Universidade dos Açores, Ponta Delgada, 2005.

RUSSEL, S; NORVING, P. **Artificial intelligence**: a modern approach. 2nd ed. Upper Saddle River, N.J: Prentice Hall, 2003.

RUBESAM, A. **Estimação não paramétrica aplicada a problemas de classificação via bagging e boosting**. 2004. 127 p. Dissertação (Mestrado em Estatística)-Universidade Estadual de Campinas, Campinas, 2004.

SCHAPIRE, R. E. The strength of weak learnability. **Machine learning**, Boston, v.5, n. 2, p 197-227, June 1990.

SCHAPIRE, R. E.; FREUND, Y. **Boosting**: foundations and algorithms. Cambridge: MIT, 2012.

SILVA, M. M. **Uma abordagem evolucionária para o aprendizado semi-supervisionado em máquinas de vetores de suporte**. 2008. 106 p. Dissertação (Mestrado em Engenharia Elétrica) - Universidade Federal de Minas Gerais, Belo Horizonte, 2008.

SOUZA, M. C. M.; SAES, M. S. M.; OTANI, M. N. Pequenos agricultores familiares e sua inserção no mercado de cafés especiais: uma abordagem preliminar. **Informações Econômicas**, São Paulo, v. 32, n. 11, p. 16-26, nov. 2002.

SPECIALITY COFFE ASSOCIATION OF AMERICA. **SCAA protocols**: cupping specialty coffee. Long Beach: SCAA, 2009.

WEISS. S; KULIKOSKI, C. **Computer Systems that learn: classification and-prediction methods from statisticas, neural nets, machine learning, and expert system**. San Francisco: Morgan Kaufmann, 1991.

ANEXOS

Tabela 14 Classificação obtida método Boosting dos dados simulados.

	Real			
Predito	Classe 1	Classe 2	Classe 3	Total
Classe 1	120	25	0	145
Classe 2	26	72	39	137
Classe 3	3	12	103	118
Total	149	109	142	400

Tabela 15 Classificação obtida pelo método Bagging dos dados simulados.

	Real			
Predito	Classe 1	Classe 2	Classe 3	Total
Classe 1	107	29	7	143
Classe 2	29	79	36	144
Classe 3	3	42	68	113
Total	139	150	111	400

Tabela 16 Classificação obtida pela Análise Discriminante Quadrática dos dados simulados.

	Real			
Predito	Classe 1	Classe 2	Classe 3	Total
Classe 1	99	41	1	141
Classe 2	16	96	25	137
Classe 3	0	33	89	122
Total	115	170	115	400

Tabela 17 Classificação obtida pela Análise Discriminante Linear dos dados simulados.

	Real			
Predito	Classe 1	Classe 2	Classe 3	Total
Classe 1	55	41	49	145
Classe 2	54	32	51	137
Classe 3	37	42	39	118
Total	146	115	139	400

Tabela 18 Classificação dos grupos de provadores obtida pelo método Boosting.

	Real			
Predito	T_1	T_2	T_3	Total
T_1	27	1	4	32
T_2	0	1	0	1
T_3	3	1	9	13
Total	30	3	13	46

Tabela 19 Classificação dos grupos de provadores obtida pelo método Bagging.

	Real			
Predito	T_1	T_2	T_3	Total
T_1	24	0	7	31
T_2	0	3	0	3
T_3	6	0	6	12
Total	30	3	13	46

Tabela 20 Classificação dos grupos de provadores obtida pela Análise Discriminante Quadrática.

	Real			
Predito	T_1	T_2	T_3	Total
T_1	23	2	9	34
T_2	0	3	0	3
T_3	4	2	3	9
Total	27	7	12	46

Tabela 21 Classificação dos grupos de provadores obtida pela Análise Discriminante Linear.

	Real			
Predito	T_1	T_2	T_3	Total
T_1	18	1	11	30
T_2	0	2	1	3
T_3	7	1	5	13
Total	25	4	17	46

Tabela 22 Classificação dos cafés especiais pelo método Boosting.

	Real				
Predito	A	B	C	D	Total
A	6	0	0	0	6
B	2	76	3	17	98
C	0	0	29	0	29
D	2	32	2	65	101
Total	10	108	34	82	234

Tabela 23 Classificação dos cafés especiais obtida pelo método Bagging.

	Real				
Predito	A	B	C	D	Total
A	6	1	0	3	10
B	0	76	0	32	108
C	1	3	29	1	34
D	0	18	1	63	82
Total	7	98	30	99	234

Tabela 24 Classificação dos cafés especiais obtida pela Análise Discriminante Linear.

	Real				
Predito	A	B	C	D	Total
A	0	8	3	2	13
B	0	72	1	28	101
C	0	8	29	1	38
D	0	36	0	46	82
Total	0	124	33	77	234

Tabela 25 Classificação dos cafés especiais obtida pela Análise Discriminante Quadrática.

	Real				
Predito	A	B	C	D	Total
A	0	8	1	1	10
B	1	66	6	30	103
C	2	13	20	2	37
D	0	27	0	57	84
Total	3	114	27	90	234