ESTATÍSTICA APLICADA AO DATA SCIENCE

ANÁLISE EXPLORATÓRIA E ALGORITMOS DE AGRUPAMENTO

Autor: Dr. Antônio Gomes de Mattos Neto

Revisor: Rafael Maltempe

INICIAR

introdução

Introdução

Já discutimos como, a partir de um conjunto de dados, que na estatística chamamos de amostra, podemos desenvolver algoritmos preditivos e aplicá-los a situações da vida.

Especificamente, aplicamos modelos de regressão linear simples e múltipla na predição do valor de imóveis, modelos de regressão logística simples e múltipla na predição da probabilidade de inadimplência com cartões de crédito, e árvores de decisão na predição do volume de vendas de um produto de varejo, uma boneca falante. Dessas aplicações, a primeira é chamada de tarefa de regressão, as outras duas de tarefas de classificação.

Aqui, nesta unidade, deixaremos de lado os modelos preditivos da aprendizagem supervisionada, que são aqueles em que treinamos um algoritmo com base em exemplos da variável resposta, de tal forma que esse algoritmo, alimentado com novos dados de entrada, seja capaz de predizer o resultado dessa variável resposta. Nosso foco, agora, será em algoritmos de aprendizagem não supervisionada. Mais especificamente, discutiremos sobre dois tipos de algoritmos de agrupamento, que também são chamados de algoritmos de clusterização.

Análise Exploratória

Nesta seção abordaremos o que vem a ser mineração de dados e sua relação com a ciência de dados e algoritmos de *machine learning*. Também falaremos sobre análise exploratória de dados.

Mineração de Dados

No início deste *e-book*, apresentamos uma discussão sobre estatística, ciência da computação e ciências dos dados. A estatística é a mais antiga dessas ciências, seguida da ciência da computação e da ciência de dados, que é a mais nova dentre elas. Mas nem na discussão inicial, nem nas unidades que se seguiram, demos atenção à "Mineração de Dados", em inglês, *Data Mining*, a menos de algumas menções esparsas. Aqui, vamos corrigir isso, pois há, entre a mineração de dados e a estatística, a ciência da computação e a ciência dos dados, uma relação muito forte.

Por volta dos anos 70 e 80 do século passado, as grandes corporações já possuíam, como parte de sua infraestrutura e de sua estrutura organizacional, centrais de processamento de dados. Nelas, armazenavam e processavam informações sobre suas transações comerciais, sobre seus estoques, sobre suas

finanças, sobre sua contabilidade, sobre seus clientes etc. Todos esses eram dados que chamamos de dados estruturados. Algumas corporações já começavam também a armazenar dados não estruturados. Essa época pode ser considerada como a antessala do que depois viemos chamar de *big data.*

Por outro lado, processos gerenciais envolvem sempre as etapas de planejamento, execução do que é planejado, monitoramento e controle do que é executado para se averiguar se saiu conforme o plano. Processos gerenciais demandam muitos dados, que se agrupam em coisas que chamamos de informações e geram conhecimento, conhecimento sobre quão bem, ou não, está indo a produção ou os serviços da empresa, quão equilibrados ou desequilibrados estão seus estoques, quão satisfeitos ou insatisfeitos estão seus clientes e tudo mais que seja relevante para a empresa ter sucesso nos seus negócios.

Quando temos poucos dados, um ser humano pode armazená-los em um livro de registros, processá-los em uma calculadora, e exibi-los em uma tabela ou um gráfico em papel milimetrado. Mas, naquela época, essas grandes corporações começaram a ter muitos dados. Tornou-se importante saber processá-los de forma inteligente, extrair deles informações e gerar conhecimento, e fazer isso da forma mais automatizada possível. Para isso, usaram de algoritmos e essa atividade de extrair conhecimento de dados por meio de algoritmos foi designada de *data mining*, ou seja, de mineração de dados. Inicialmente, era realizada por meio de abordagens que alguns autores chamam, em inglês, de "*design approach*", que pode ser traduzido por "abordagem por regras", essas regras definidas pelo ser humano (CHOLLET, 2018).

Ora, explorar dados e procurar descobrir padrões e gerar conhecimento sempre foi a principal tarefa da estatística. Também veio a ser um dos propósitos da ciência da computação, quando deu início às tentativas de emular a inteligência humana com seus algoritmos de aprendizagem supervisionada e não supervisionada. E também é um dos principais propósitos, senão o mais importante, da ciência dos dados. Todas essas ciências, a estatística, a ciência da computação (essa parte da ciência da computação dedicada ao *machine learning*), a mineração de dados e a ciência dos dados, então, com propósitos

similares, interesses similares e abertas a fazerem proveito dos avanços produzidos pelas outras ciências.

A mineração de dados também fez isso e, para além dos mecanismos que criou inicialmente para explorar dados, descobrir padrões e gerar conhecimento, abraçou aqueles da estatística e da ciência da computação. Em especial, abraçou algoritmos de machine learning. Daí que, em muitos aspectos, a mineração de dados ficou muito parecida com a ciência dos dados, e alguns autores não fazem muita distinção entre as duas, mas todos reconhecem que fazem uso de muitas abordagens e ferramentas iguais. Veja o Quadro 4.1. Lá apresentamos uma breve descrição de características de cada uma dessas áreas: a Estatística, a Ciência da Computação, a Mineração de Dados e a Ciência de Dados. Espero que sirva para você entender a relação entre elas.

Talvez a maior diferença esteja no fato de que mineradores de dados dão maior atenção aos aspectos de gerenciamento de bancos de dados, e à ciência de dados um pouco menos. Uma outra diferença pode ser que a mineração de dados trata explicitamente de métodos da “abordagem por regras”, que não são habitualmente discutidos na ciência dos dados, mas poderiam ser. Recomendo a você, ao se deparar com esses termos, mineração de dados e ciência dos dados, imaginar que são áreas irmãs, com muitas similaridades.

Quadro 4.1 - Quadro resumo das áreas relacionadas à ciência dos dados
Fonte: Elaborado pelo autor.

Podemos fechar essa seção observando algo que você já deve ter notado. A terminologia empregada por essas áreas, a estatística, a ciência da computação, a mineração de dados e a ciência de dados possuem similaridade e dissimilaridades, pois, às vezes, usam dos mesmos termos para dizer a mesma coisa, às vezes, usam dos mesmos termos para dizer coisas um pouco diferentes e, muitas vezes, usam termos diferentes para dizer a mesma coisa. Como convivemos com todas elas, o jeito é prestar atenção e não nos confundirmos. Veja o que dizem os autores abaixo citados:

> *A ciência dos dados é uma fusão de múltiplas disciplinas, incluindo estatística, ciência da computação, tecnologia da informação e campos de domínios específicos. Consequentemente, podem-se utilizar de muitos termos diferentes para se referir a um dado conceito (BRUCE; BRUCE, 2019, p. 5-6).*

Passaremos, agora, a um outro assunto, parte importante desse mundo da estatística, da ciência dos dados e das suas ciências irmãs: a análise exploratória de dados.

Análise Exploratória

Não há um sentido estrito, único e universal, para o que se convencionou chamar de análise exploratória de dados. Alguns autores definem análise exploratória de dados como as atividades iniciais do processo de análise dos dados, e chamam atividades posteriores de atividades de modelagem de dados. Nesse sentido, explorar dados é o exame inicial dos dados, as primeiras descobertas, as primeiras ideias que você tem sobre o que eles parecem dizer. Outros admitem que análise exploratória de dados pode englobar o esforço completo de examinar, depurar, tratar, visualizar e modelar dados. Vamos examinar, como exemplo, as palavras de um muito famoso cientista de dados,

Hadley Wickham que, em um livro que escreveu em coautoria, diz (tradução livre do autor deste *e-book*):

> *O objetivo da primeira parte deste livro é fazer com que você ganhe velocidade com as ferramentas básicas de exploração de dados o mais rápido possível. Exploração de dados é a arte de olhar os seus dados, rapidamente gerar hipóteses sobre eles, e rapidamente testar essas hipóteses. E repetir isso outra vez, outra vez, outra vez. O objetivo da exploração de dados é a geração de pistas sobre o que os dados nos revelam, pistas que você poderá explorar, mais tarde, em maior profundidade (WICKHAM; GROLEMUN, 2017, p.1).*

E ele avança nesse tema escrevendo (tradução livre do autor deste *e-book*):

> *A análise exploratória dos dados não é um processo formal, com um conjunto estrito de regras. Análise exploratória de dados é um estado de espírito. Durante as fases iniciais você deve se sentir livre para investigar todas as ideias que lhe ocorrerem. Algumas dessas ideias vingarão, outras não trarão resultados (WICKHAM; GROLEMUN, 2017, p. 81).*

Veja que ponto interessante! Aqui, um dos maiores cientistas de dados da atualidade dá maior ênfase a um estado de espírito do que às técnicas propriamente ditas: a curiosidade, a vontade de explorar o desconhecido, a capacidade de gerar ideias (hipóteses) sobre o que os dados nos contam. As técnicas são, e continuarão a ser, cada vez mais importantes. Mas elas não serão nada se o seu espírito não for investigativo e criativo, o verdadeiro espírito de um cientista de dados.

Neste e-book, entenderemos análise exploratória de dados como as fases iniciais do processo de investigação dos dados, incluindo a etapa de análise descritiva dos dados. Essa etapa inicial de análise nos leva a descobertas de padrões e nos ajuda a formular hipóteses, que podem ser investigadas em mais profundidade depois, por exemplo, com modelos preditivos.

Passaremos agora a uma atividade para, em seguida, prosseguirmos com o estudo de caso que, nesta unidade, servirá de base para a discussão sobre

algoritmos de agrupamento.

praticar

Vamos Praticar

> *"Este livro é sobre descoberta. Nele, o conceito de descoberta assume dois sentidos. Primeiro, o romântico, no qual a descoberta é vista como um fenômeno emocionante e prazeroso. E é esse fenômeno que se espera provocar durante a sua leitura. Segundo, o sentido técnico, no qual a descoberta continua sendo igualmente emocionante e prazerosa, mas passa também a ser o resultado de um criterioso estudo sobre dados. A partir do estudo e da mineração de dados, a descoberta acontece, e então novo conhecimento é produzido, contribuindo para a melhoria de produtos, sistemas, processos, negócios etc.[...] Contudo, minerar dados para descobrir conhecimento não é uma tarefa trivial. É preciso conhecer os dados, o processo de análise e descoberta, as tarefas e técnicas de mineração e as ferramentas matemáticas e computacionais que se aplicam nesse contexto. Portanto, a descoberta é um processo. Ainda, é preciso conhecer o ambiente em que os dados são produzidos e que tipo de conhecimento esse ambiente necessita e espera receber."*

SILVA, L. A.; PERES, S. M.; BOSCARIOLI, C. **Introdução à mineração de dados:** com aplicações em R. Rio de Janeiro: Elsevier, p.3, 2016.

A respeito da mineração de dados e sua relação com áreas, tais como a estatística, a ciência da computação e a ciência dos dados, assinale a alternativa correta:

- ○ **a)** Vimos que a estatística, a ciência da computação e a ciência de dados são áreas relacionadas, mas a mineração de dados e ciência de dados são áreas independentes, sem nenhuma relação.
- ○ **b)** *Machine learning* são algoritmos desenvolvidos, principalmente, pela ciência da computação. São usados na ciência de dados e não são usados na mineração de dados.
- ○ **c)** Minerar dados para descobrir conhecimento é uma tarefa trivial. Basta conhecer dos dados, o processo de análise e descoberta, as tarefas de mineração e as ferramentas matemáticas e computacionais que se aplicam nesse contexto.

- **d)** Não é importante conhecer o ambiente em que os dados são produzidos e que tipo de conhecimento esse ambiente necessita e espera receber. Importante, de fato, é que o minerador de dados ou cientista de dados seja curioso.
- **e)** O processo de descoberta de padrões e geração de conhecimento por meio de dados tem um sentido romântico, por ser emocionante e prazeroso, e um sentido técnico, pois demanda estudos técnicos criteriosos.

Estudo de Caso - Violência Urbana

Nesta seção, apresentaremos a você o caso que nos servirá de principal exemplo na discussão sobre algoritmos de agrupamento. Algoritmos de agrupamento, também chamados de algoritmos de clusterização, são, habitualmente, considerados parte do instrumental do estatístico (e do cientista de dados, do minerador de dados etc.) para a análise exploratória dos dados. Dois personagens nos guiarão por esse caminho, um estatístico sênior americano e um jovem cientista de dados brasileiro.

Algoritmos de Agrupamento

Agrupamento é uma das tarefas da mineração de dados. Também é da estatística e da ciência dos dados, como já explicamos. É uma tarefa que faz parte dos métodos de aprendizagem não supervisionada, ou seja, nesse tipo de aprendizagem, não há uma variável, dentre as variáveis observadas, que sirva de exemplo (que funcione como um supervisor), para o treinamento de algoritmos preditivos.

Quadro 4.2 - Dados biométricos sobre um animal desconhecido
Fonte: Elaborado pelo autor.

Ora, se é assim, o que é que a gente faz ao tomar a decisão de fazer uma análise de agrupamento? Vamos voltar aos dados estruturados na forma de uma tabela, com as variáveis dispostas nas colunas e as observações dispostas nas linhas. Imagine que alguém nos passe dados biométricos sobre uma espécie animal. Os dados estão em uma tabela com 10 observações do comprimento do animal e da sua massa corporal (coloquialmente, o seu peso). Esses dados são exibidos no Quadro 4.2.

Em uma análise preditiva, poderíamos pensar em escolher uma dessas variáveis com a variável resposta e a outra seria a variável de entrada. Por exemplo, o comprimento como a variável de entrada e o peso como a variável resposta e,

assim, poderíamos investigar se há uma relação entre as duas, tal que o valor de uma servisse para informar o valor da outra.

Em uma análise de agrupamento não fazemos assim. Nós procuramos observar cada linha da tabela e compará-la a outras linhas. Formamos um grupo com todas linhas que forem parecidas e outro grupo - ou outros grupos - com outras linhas que forem parecidas entre si. Ao final, teremos dois, três ou mais grupos, cada um contendo observações (linhas) similares. Depois de formarmos esses grupos, podemos dar nomes a eles.

No nosso exemplo, temos apenas dez observações. Cada linha é uma observação e refere-se a um determinado indivíduo da amostra de animais. Vamos imaginar que dividimos essas observações (linhas) em dois grupos similares entre si. Não sabemos ainda, porque ainda não separamos os grupos, quantas observações (linhas) terão cada um desses grupos. Feita essa separação das observações (linhas) em dois grupos, podemos dar nome a eles. Por exemplo, Grupo A e Grupo B. E cada indivíduo pode, então, ser classificado com indivíduo A ou indivíduo B. E um novo indivíduo, se for mais parecido com os indivíduos do Grupo B, será denominado indivíduo B. O mesmo para um novo indivíduo que seja mais parecido com os indivíduos do Grupo A.

Veja que, nesse processo, primeiramente, criamos grupos por similaridades entre si, depois damos nomes aos grupos e só, depois, classificamos novos indivíduos como sendo de uma classe, A ou B, conforme mais parecidos com um grupo ou com o outro. O ser humano faz isso com enorme facilidade, agrupar e depois classificar. Nós classificamos tudo com tanta naturalidade, que nem percebemos que primeiramente criamos os grupos para só depois classificarmos.

Em princípio, somos nós mesmos que devemos decidir em quantos grupos agrupar as observações, com base em algum critério de similaridade. Veja, então, que esse tipo de análise de agrupamento, que é parte da análise exploratória de dados, é um pouco “mais solta” que a análise para o desenvolvimento de modelos preditivos. As palavras de especialistas nos ajudarão a reforçar esse ponto (tradução livre do autor deste *e-book*):

> *Em contraste à aprendizagem supervisionada, aprendizagem não supervisionada é, frequentemente, mais desafiadora. Tende a ser mais subjetiva. Não há um objetivo claro para a análise, tal como a predição de uma variável resposta. Aprendizagem supervisionada é, frequentemente, parte da análise exploratória dos dados. Além disso, pode ser difícil avaliar os resultados obtidos da aprendizagem supervisionada. Não há um mecanismo universal, por todos aceitos, de validação cruzada do resultado, ou validação do resultado com um conjunto de dados independentes dos dados de treino. A razão para essa diferença é simples. Quando ajustamos um modelo preditivo usando uma técnica de aprendizado supervisionado, é possível checar o resultado verificando quão bem o nosso modelo prediz a resposta Y para observações não usadas no ajuste do modelo (treino do modelo). Entretanto, na aprendizagem não supervisionada, não é possível fazer essa checagem, pois não conhecemos uma resposta verdadeira. Afinal, o problema não é supervisionado (JAMES et al., 2013, p. 374).*

Voltando ao nosso exemplo, como temos apenas duas variáveis, o comprimento e a massa corporal do animal, podemos tentar plotar os dados usando a função plot() do sistema básico de gráficos do R ou outra ferramenta qualquer. Fazendo isso, obtemos o gráfico exibido na Figura 4.1. A visualização dessa pequena amostra é muito ilustrativa. Claramente, parece haver dois grupos de animais: um grupo com comprimento menor que 1,8 metros e outro com comprimento maior que 1,8 metros. No primeiro grupo, as massas corporais são menores que as massas corporais do segundo grupo. Vamos chamar o primeiro grupo de Grupo A e o segundo grupo de Grupo B.

Bem, dessa forma, com uma análise visual, acabamos de realizar nossa primeira tarefa de agrupamento, para uma pequena amostra de dados biométricos de alguns animais, que não sabemos quais são. A partir desse ponto, podemos usar um critério de similaridade: Qualquer outro animal que cair perto do Grupo A classificaremos como A e qualquer outro animal que cair perto do Grupo B classificaremos como B.

Mas o que aconteceria se, ao invés de uma amostra com apenas duas variáveis e dez observações, tivéssemos de lidar com uma amostra com 30 variáveis e 2000 observações, ou mesmo maiores, como é comum hoje em dia? Não teríamos como fazer isso visualmente. É justamente aqui que aparecem os algoritmos de agrupamento. Eles usam algum critério de similaridade, por nós definido, e automatizam a tarefa de agrupamento. Caberá a nós decidirmos se os grupos fazem algum sentido ou não. É por esse motivo que dizemos que é uma análise mais difícil ou "mais solta". E a presença de um especialista (especialista na área que está sendo investigada) é sempre importante, mais importante ainda fica nessa situação.

Mas, afinal, quem são aqueles animais? Na verdade, são leoas e leões. Os dados foram obtidos em consulta ao site do Zoológico de San Diego, nos Estados Unidos da América (SAN DIEGO ZOO). Lá, podemos ver que leoas medem de 1,4 a 1,7 metros (sem contar o rabo) e pesam de 122 a 180 quilos, e leões medem de 1,7 a 2,5 metros e pesam de 150 a 260 quilos. Criamos uma amostra aleatória com base nesses dados, com propósito didático. É como se realmente estivéssemos medindo o comprimento e o peso de 10 leões. Aqui, nessa simulação, acabamos com 6 fêmeas e 4 machos. Como são apenas 2 variáveis e 10 observações, ficou fácil agrupá-los. Na verdade, o Grupo A é o grupo das leoas (as fêmeas, menores e menos pesadas) e o Grupo B é o grupo dos leões (os machos, maiores e mais pesados).

Isso é o que os biólogos chamam de dimorfismo sexual. Para algumas espécies, pode se manifestar com machos maiores, como é o caso dos leões, para outras

espécies, com fêmeas maiores. Para essa amostra didática, a fronteira entre eles é relativamente clara e foi suficiente observar apenas duas variáveis quantitativas para fazer o agrupamento. Mas poderíamos incluir outras variáveis, quantitativas ou qualitativas. Novas variáveis podem, eventualmente, ajudar a desvendar características de grupos de indivíduos que, de fato, não conhecemos. Aqui, as leoas e leões, que são uma das espécies de mamíferos mais estudada do mundo, serviram como um exemplo simples, para lhe mostrar o princípio de funcionamento de algoritmos de agrupamento.

Estudo de Caso - Violência Urbana

O jovem cientista de dados brasileiro, já mencionado no início dessa Seção 2, foi fazer um estágio em uma empresa de pesquisa e análise de dados sociais. Lá soube que seu supervisor seria um estatístico sênior americano, muito respeitado em análise de dados sociais. E uma das primeiras tarefas que lhe passou foi a de análise exploratória com algoritmos de agrupamento, com bases em dados reais de violência urbana, coletados de uma pesquisa feita em 50 estados americanos (McNEIL). Esses dados, que são facilmente acessíveis em uma base de dados do R, chamada de USArrest (R DOCUMENTATION), são de uso livre (de domínio público) e adequados a propósitos didáticos. O estatístico sênior americano explicou ao jovem cientista de dados que os mesmos métodos que ele usaria, no seu treinamento, poderiam ser aplicados a estudos similares, relativos a estados ou municípios brasileiros.

Quadro 4.3 - Dados USArrest sobre violência urbana
Fonte: R Documentation (2020).

O conjunto de dados USArrest consiste em um data-frame com 50 observações e 4 variáveis (ver alguns exemplos no Quadro 4.3). As 50 observações são os cinquenta Estados dos EUA e as 4 variáveis são:

Murder Númerode homicídios por 100.000 habitantes

Assault Número de assaltos por 100.000 habitantes

UrbanPop Porcentagem de habitantes urbanos no estado

Rape Números de estupros por 100.000 habitantes

Na próxima seção, o nosso jovem cientista de dados se dedicará à boa prática de começar fazendo uma análise descritiva desses dados. Como já explicado,

tanto a análise descritiva como a análise de agrupamento podem ser consideradas parte do que chamamos de análise exploratória dos dados. Há outras técnicas que também são consideradas parte integrante da análise exploratória dos dados, mas não serão discutidas aqui.

praticar

Vamos Praticar

O aprendizado não supervisionado pode ter diferentes objetivos possíveis. Em alguns casos, pode ser usado para criar uma regra preditiva na ausência de uma resposta rotulada. Os métodos de agrupamento podem ser usados para identificar grupos de dados significativos. Por exemplo, usando os cliques da web e dados demográficos de usuários de um site, podemos ser capazes de agrupar diferentes tipos de usuários. O site poderia, então, ser personalizado para esses diferentes grupos. (BRUCE; BRUCE, 2019, p. 5-6)

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019.

A respeito da análise de agrupamento, assinale a alternativa correta:

- ○ **a)** Análise de agrupamento faz parte do método de aprendizado supervisionado.
- ○ **b)** Não é possível realizar análise de agrupamento se não temos definida qual é a variável resposta na amostra analisada.
- ○ **c)** Agrupar diferentes espécies de animais e diferentes espécies de plantas não são exemplos de análise de agrupamento.
- ○ **d)** Agrupar diferentes usuários de um site ou diferentes grupos de clientes de uma empresa são exemplos de análise de agrupamento.
- ○ **e)** Análise de agrupamento é um dos métodos preditivos que faz parte da chamada aprendizagem supervisionada.

Análise Descritiva dos Dados

Deu início à sua análise descritiva examinando a variável Murder, que é o número de homicídios por cada 100.000 habitantes, para cada um dos 50 Estados americanos. Como recomendado, gerou alguns sumários estatísticos, para o que fez uso da função summary() do R, obtendo:

Min.	1stQu.	Median	Mean	3rd Qu.	Max.
0.800	4.075	7.250	7.788	11.250	17.400

Percebeu haver uma grande variação do número de homicídios entre os Estados americanos, já que o mínimo é de 0,8 por cada 100.000 mil habitantes, contra um máximo de 17,4 por cada 100.000 habitantes. Quase 22 vezes mais.

Em seguida, o jovem cientista de dados, para a visualização desses dados, lançou mão da função gráfica hist() do R, apropriada para a visualização de dados quantitativos, e obteve como resultado o gráfico exibido na Figura 4.2. Essa figura demonstra visualmente aquilo que já observamos com os sumários estatísticos. Da figura, vemos que a maior frequência se dá com Estados que apresentam entre dois a quatro homicídios por cada 100.000 habitantes.

O cientista de dados aprendiz voltou sua atenção, agora, para os dados relativos à variável *Assault*. Novamente, gerou sumários estatísticos com a função summary() do R e obteve os seguintes resultados:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
45.0	109.0	159.0	170.8	249.0	337.0

A variação entre o valor mínimo e máximo observado não é tão grande quanto no caso anterior da variável *Murder*, porém, em termos absolutos, o número de assaltos é maior do que o de homicídios por cada 100.000 habitantes, como se espera que seja: menos homicídios que assaltos. Passou, então, à visualização desses dados, por meio de um histograma, exibido na Figura 4.3, onde podemos observar dois picos (modas) na distribuição da frequência de assaltos para os 50 Estados.

Em seguida, o cientista de dados aprendiz voltou sua atenção para os dados relativos à variável *Rape* e, novamente, gerou resumos estatísticos com a função summary() do R, obtendo os seguintes resultados:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
7.30	15.07	20.10	21.23	26.18	46.00

A variação entre o valor mínimo e máximo observado não é tão grande quanto no caso da variável *Murder*, porém, em termos absolutos, o número de estupros é maior do que o de homicídios por cada 100.000 habitantes. O jovem cientista de dados também gerou o histograma exibido na Figura 4.4, onde podemos observar uma certa assimetria, com a mediana se situando entre 20 e 21 casos por cada 100.000 habitantes para os 50 Estados.

Finalmente, o nosso cientista de dados aprendiz voltou sua atenção para os dados relativos à variável *UrbanPop*. Gerou sumários estatísticos com a função summary() do R, obtendo os seguintes resultados:

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
32.00	54.50	66.00	65.54	77.75	91.00

Percebeu que a variação entre o valor mínimo e máximo observado para a porcentagem de população urbana dos Estados era o menor dentre todas as variáveis observadas. Também viu que havia ao menos um Estado com uma baixa porcentagem de população urbana, de apenas 32% e, ao menos, um Estado com população quase que inteiramente urbana, de 91%. Na média, os Estados apresentaram cerca de 65,5% da sua população vivendo nas cidades. Gerou o histograma dos dados relativos à porcentagem da população urbana, exibido na Figura 4.5, onde podemos observar uma certa assimetria, com a mediana se situando entre 60% e 70% de população urbana para os 50 Estados, o que o sumário estatístico confirma.

Com isso, o nosso jovem cientista de dados concluiu a etapa da análise descritiva de cada variável individualmente e passou à etapa de análise descritiva de possíveis relações entre essas variáveis.

Análise Descritiva da Relação entre as Variáveis

O nosso jovem cientista de dados passou ao exame da possível relação entre as variáveis. Todas elas são variáveis quantitativas. Para casos como esse, ele se lembrou de que existe, no software estatístico R, a função cor(), a qual calcula a correlação entre múltiplas variáveis quantitativas simultaneamente. Ele obteve o seguinte resultado:

	Murder	Assault	UrbanPop	Rape
Murder	1.00	0.80	0.07	0.56
Assault	0.80	1.00	0.26	0.67
UrbanPop	0.07	0.26	1.00	0.41
Rape	0.56	0.67	0.41	1.00

Imediatamente, notou que todas correlações são positivas. Se há uma relação entre duas variáveis quantitativas, e elas se comportam uma em relação à outra de forma aproximadamente linear, correlações positivas indicam que, quando uma aumenta, a outra também aumenta. Veja que correlações entre dados da mesma variável - a correlação da variável com ela mesma - são sempre iguais a 1 (um). O jovem cientista de dados também percebeu que as menores correlações são aquelas entre quaisquer das outras variáveis com a variável *UrbanPop*, a porcentagem de população vivendo nas cidades nos 50 Estados americanos.

Ao consultar seu supervisor, este o fez notar que, para dados sociais, não se devem desprezar correlações nessa faixa. Por exemplo, a correlação de 0,41 entre a porcentagem de população urbana (*UrbanPop*) e o número de estupros por cada 100.000 habitantes (*Rape*) é significativa, indicando haver um efeito da concentração da população em cidades com a frequência de estupros, isto é, um aumento da população urbana do Estado associado a um aumento dos casos de estupro, por 100.000 habitantes, para os dados da amostra analisada. O supervisor americano também disse que, até mesmo, a correlação de 0,21 entre assaltos e a porcentagem da população urbana não deve ser desprezada.

Melhor investigada sim, desprezada não, mesmo se não tão forte quanto os 0,41 do caso anterior.

O jovem cientista de dados percebeu o valor do apoio de alguém mais experiente, ainda mais em situações desse tipo, de análise de dados associados aos fenômenos sociais, como a violência urbana. Fenômenos sociais estão entre os mais complexos que conhecemos. Refletiu um pouco sobre isso e decidiu avançar com a visualização dessas relações. Lembrou-se do R e da sua função gráfica pairs(). Aplicou-a aos dados de USArrests e obteve o gráfico exibido na Figura 4.6.

Esse é o *output* gráfico da função pairs() do R aplicada aos dados de USArrest. São múltiplos gráficos de dispersão, de cada uma das quatro variáveis contra cada uma das outras três. Como já dissemos, só pode ser aplicada a variáveis quantitativas, pois gráficos de dispersão só podem ser aplicados para a visualização da possível relação entre variáveis quantitativas. No entanto, o nosso jovem cientista de dados não sabia muito bem como ler este gráfico. Seu supervisor o ajudou, explicando assim:

- O nome de cada variável aparece uma vez ocupando uma das células do gráfico. Essa é a célula que seria do gráfico de dispersão da variável contra ela mesma, com cor = 1. Esse seria um gráfico de dispersão sem utilidade, por não trazer informação relevante, a relação de uma variável com ela mesma;

- Ao caminharmos na horizontal, partindo de qualquer uma dessas variáveis, ela deve ser vista como se estivesse no eixo vertical y, e cada uma das outras variáveis como se estivessem no eixo horizontal x;
- Dessa forma, como temos 4 variáveis, 4 x 3 resultam nos 12 gráficos de dispersão exibidos na figura.

O supervisor americano continuou e explicou que esses 12 gráficos são:

- y = Murder versus x = Assalt ou UrbanPop ou Rape (3 gráficos);
- y = Assault versus x = Murder ou UrbanPop ou Rape (3 gráficos);
- y = UrbanPop versus x = Murder ou Assault ou Rape (3 gráficos);
- y = Rape versus x = Murder ou Assault ou UrbanPop (3 gráficos).

Com essa explicação, o jovem cientista de dados finalmente entendeu que se são 4 variáveis e cada uma com 1 gráfico de dispersão com cada uma das outras 3 variáveis, ao final, chega-se aos 12 gráficos. Tudo estava se encaixando até aqui.

As correlações e os gráficos da Figura 4.6 conversam entre si. Veja, por exemplo, a correlação entre *Murder* (homicídios por 100.000 habitantes) e *Assault* (assaltos por 100.000 mil habitantes), de 0,80. É uma correlação positiva forte, indicando que um aumento no número de assaltos leva a um aumento no número de homicídios, o que pode ser visto também no gráfico de dispersão entre *Murder versus Assault*. Os pontos dos pares ordenados (x, y), onde x é *Assault* e y é *Murder*, apresentam uma dispersão não muito alta e uma tendência de subida, pois com um aumento de x = *Assault*, y = *Murder* aumenta.

A análise descritiva, feita pelo jovem cientista de dados, ainda rendeu muitas discussões entre ele e seu supervisor, mas ficaremos por aqui com esse assunto, pois devemos progredir e ver como ele aplicou algoritmos de agrupamento a este e outros casos.

praticar

Vamos Praticar

O agrupamento é uma técnica para dividir dados em diferentes grupos, na qual os registros de cada grupo são semelhantes uns aos outros. Um objetivo do agrupamento é identificar grupos de dados significantes e significativos. Os grupos podem ser usados diretamente, analisados mais a fundo ou passados como características ou resultado para um modelo de regressão ou classificação. (BRUCE; BRUCE, 2019, p. 265).

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019.

A respeito da técnica de agrupamento, assinale a alternativa correta:

- ○ **a)** O agrupamento não é uma técnica para dividir dados em diferentes grupos, na qual os registros de cada grupo são semelhantes uns aos outros
- ○ **b)** Como agrupamento é um método de aprendizagem supervisionada, não é possível realizar análise de agrupamento se não há uma variável resposta nos dados analisados.
- ○ **c)** Agrupar registros de dados semelhantes é o mesmo que agrupar observações semelhantes, pois os termos, observações e registros são empregados como sinônimos em mineração de dados e na ciência de dados.
- ○ **d)** Depois de realizado o agrupamento, os grupos identificados não podem ser nomeados e seus exemplares usados como variáveis de resposta em modelos preditivos.
- ○ **e)** Análise de agrupamento é um método de análise relativamente fácil, pois faz parte dos métodos de aprendizagem supervisionada. A variável supervisora (que é a variável resposta) nos ajudaria a avaliar o quão bom é o resultado obtido.

Agrupamento ou Análise de Cluster

Nesta seção, veremos como o nosso jovem cientista de dados faz uso de alguns algoritmos para realizar análise de agrupamento. Especificamente, instruído pelo seu supervisor, ele utiliza dois dos mais famosos desses algoritmos, agrupamento por k-médias e agrupamento hierárquico.

Agrupamento por k-Médias

O supervisor americano do nosso jovem cientista de dados decidiu que já estava na hora de treiná-lo em alguns métodos de aprendizagem não supervisionada, muito aplicados na análise exploratória de dados. Achou melhor começar com algoritmos de agrupamento, e escolheu o de k-médias por ser o de mais simples entendimento. Chamou o jovem cientista de dados e explicou como esse algoritmo funciona:

1. É aplicado em agrupamento de dados oriundos de variáveis quantitativas, como no caso dos leões e das leoas, onde as variáveis são comprimento (m) e massa corporal (kg) dos felinos;
2. O analista decide em quantos grupos dividir as observações. Como nunca se sabe se a divisão fará sentido, convém experimentar alguns

números até que, em consenso com o especialista da área que está sendo investigada, se chegue a um resultado que tenha algum sentido;
3. Depois de decidido o número de grupos que serão formados, o próprio algoritmo faz uma escolha randômica das observações formando os grupos aleatoriamente;
4. A partir desse ponto, ele calcula a centroide dos grupos e verifica, para uma observação, de qual centroide ela está mais próxima. Troca essa observação de grupo se ela estiver no grupo errado (ela está no grupo errado se estiver mais longe da centroide do seu grupo do que do outro grupo);
5. O passo 4 é repetido até não mais haver troca de observações entre os grupos.

O jovem cientista de dados quase não acreditou que esse algoritmo funcionasse, e perguntou se ele - o algoritmo - não entraria em um loop infinito. Na verdade, não, explicou o estatístico sênior americano, pois esse algoritmo sempre converge para uma solução final, cada observação alocada a um grupo, cuja centroide está mais perto do que a centroide de qualquer outro grupo. E pediu ao jovem cientista de dados que praticasse o algoritmo de k-médias com os dados relativos aos leões e leoas. Vamos, agora, colocar os sexos na tabela dos dados, conforme o Quadro 4.4.

Quadro 4.4 - Comprimento, massa corporal e sexo dos leões
Fonte: Elaborado pelo autor.

A ideia era esconder do algoritmo o sexo dos animais e ver se ele agruparia de forma correta por sexo, mas logo se percebeu um problema, pois o comprimento (m) e massa corporal (kg) estão em escalas diferentes. Para evitar distorções no resultado, ele decidiu padronizar essas variáveis subtraindo de cada uma sua média, e dividindo o resultado pelo desvio padrão. Essa padronização, às vezes, chamada de reescalonamento, é muito comum na estatística. Feita essa operação, o jovem cientista de dados usou da função kmeans() do R, especificando querer 2 grupos, plotou o resultado e obteve o gráfico exibido na Figura 4.7.

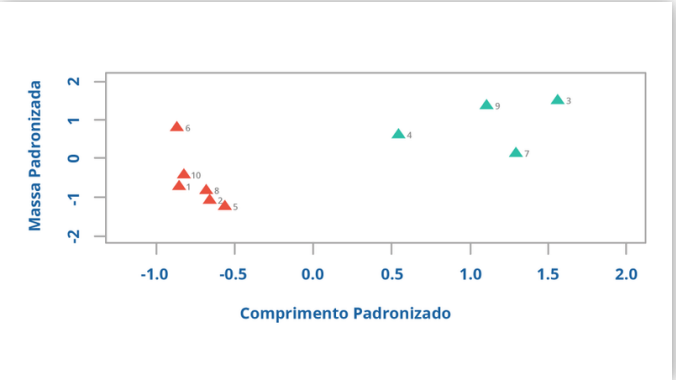


Figura 4.7 - Agrupamento dos dados relativos a leões e leoas
Fonte: Elaborado pelo autor.

Veja que interessante: o algoritmo de agrupamento por k-médias, tendo sido informado pelo cientista de dados que deveria formar dois grupos das 10 observações das leoas e leões, usando apenas como conhecimento o comprimento e a massa corporal padronizados dos felinos, agrupou-os exatamente em 6 fêmeas (em vermelho) e 4 machos (em verde). Ao lado de cada uma das observações individuais, há um número, que é o número da observação (linha) da tabela de dados.

Por exemplo, vamos ver quem é a leoa da linha 6 na tabela. Ela está destoando um pouco das demais leoas, com comprimento de 1,50 metros e massa de 204,0 quilos. Seu comprimento está na faixa normal dos comprimentos de leoas, mas sua massa corporal está acima do normal. A Figura 4.7 mostra isso visualmente, de forma clara, mas note que, nesse exemplo simples, atribuído ao jovem cientista de dados com o propósito de treinamento, nós já sabíamos quem eram as fêmeas e quem eram os machos e suas características. Para um conjunto de dados oriundos de um fenômeno ou processo desconhecido, esses agrupamentos permitem descobertas, se os grupos formados fizerem algum sentido, por isso a importância de um especialista da área ao lado do estatístico ou do cientista de dados para ajudar a interpretar os resultados.

Também é importante, na análise exploratória com algoritmos de agrupamento, fazer diversas tentativas, pois, como já dissemos, é um processo de descoberta e não há uma resposta certa, como em problemas supervisionados. Nada é garantido e não se sabe exatamente onde se vai chegar, pois é uma exploração!

Mas havia outra tarefa atribuída ao jovem cientista de dados que o aguardava desde o início dessa jornada, a análise dos dados de USArrest. Vamos ver, em seguida, como ele enfrentou esse desafio.

Agrupamento Hierárquico

Por recomendação do estatístico sênior americano, o jovem cientista de dados deveria usar na análise dos dados de USArrest um outro algoritmo de agrupamento, denominado de agrupamento hierárquico. A ideia do seu supervisor era apenas de treiná-lo nesses dois importantes algoritmos de agrupamento, o de k-médias e o de agrupamento hierárquico. Não se sabe, a priori, para um dado conjunto de dados, qual deles trará melhores descobertas. É um processo de tentativa e erro. Além disso, há outros algoritmos de agrupamento, mas não os discutiremos aqui.

A forma de funcionamento do algoritmo de agrupamento hierárquico é diferente do de k-médias. O estatístico sênior americano chamou o jovem cientista de dados e, como fez anteriormente para o algoritmo de k-médias, explicou como o algoritmo de agrupamento hierárquico funciona:

1. É aplicado em agrupamento de dados oriundos de variáveis quantitativas, como no caso das variáveis *Murder*, *Assault*, *UrbanPop* e *Rape* de *USArrest*;
2. O analista não decide em quantos grupos dividir as observações. O algoritmo hierárquico sempre vai começar com grupos com um só indivíduo, ou seja, tanto os grupos quanto o número de observações na amostra analisada;
3. O analista escolhe um critério de medida de distância de qualquer grupo do conjunto de dados até qualquer outro grupo já formado. Por exemplo, pode ser a distância mais próxima entre indivíduos dos dois grupos, a mais afastada entre indivíduos dos dois grupos ou a distância entre as centroides dos dois grupos;
4. A partir desse ponto, o algoritmo calcula, para cada grupo, de qual outro grupo ele está mais próximo e funde esses grupos mais próximos em um só;
5. O passo 4 é repetido até se chegar a apenas um grupo, com todas as observações do conjunto de dados.

O estatístico sênior americano continuou e disse que, por exemplo, no caso do conjunto de dados de USArrest, com suas 4 variáveis quantitativas e 50 observações (ou seja, 50 é o tamanho da amostra), o algoritmo de agrupamento hierárquico começaria com 50 grupos de um só indivíduo cada e, aos poucos, fundiria os mais próximos entre si, dois a dois, até acabar com um só grupo com 50 indivíduos.

Nesse ponto, o jovem cientista de dados fez uma pergunta óbvia ao seu supervisor: bem, afinal, nesse exemplo, se o algoritmo começa com 50 grupos e acaba com 1 grupo, que agrupamento devo escolher? O seu supervisor já esperava por essa pergunta e deu uma resposta que é típica para análises de agrupamento: é. você que escolhe qual agrupamento faz mais sentido. Por isso, se você não é um expert no assunto analisado, tenha ao seu lado um especialista da área para lhe ajudar com esse processo de descoberta. Em seguida, seu supervisor pediu que fizesse a análise de agrupamento dos dados USArrest, mas, para simplificar, limitou o conjunto de observações a apenas 5 Estados, como mostra o Quadro 4.5.

	Murder	*Assault*	*UrbanPop*	*Rape*
Arkansas	8,8	190	50	19,5
Louisiana	15,4	249	66	22,5
New Mexico	11,4	285	70	32,1
Oklahoma	6,6	151	68	20,2
Texas	12,7	201	80	25,5

Quadro 4.5 - Dados de USArrests para apenas cinco estados americanos

Fonte: Elaborado pelo autor.

Só restava, agora, ao jovem cientista de dados, arregaçar as mangas da sua camisa e colocar suas mãos à obra. Aqui, também decidiu fazer o reescalonamento das variáveis (para cada variável subtrair sua média e dividir pelo seu desvio padrão). Escolheu um critério de medida de distância entre grupos de observações e, por meio do R, obteve o resultado exibido na Figura 4.8.

Esse gráfico é denominado de dendrograma, que significa diagrama na forma de uma árvore. Esse é o *output* típico da função hclust() do R que realiza agrupamento hierárquico. Veja que ele exibe, de baixo para cima, primeiro todas observações, que são os grupos de um só indivíduo, e vai agrupando-as até o topo, onde só há um grupo com todos indivíduos.

Ao analisar esse gráfico de cima para baixo, e comparar com o Quadro 4.5, o jovem cientista de dados viu que o algoritmo forma dois grupos, um com dois Estados, Arkansas e Oklahoma, que têm taxas de homicídios por 100.000 mil habitantes entre 6,6 a 8,8, e o outro com três Estados, Louisiana, Texas e New Mexico, que têm taxas de homicídios entre 11,4 a 15,4, ou seja, são dois grupos onde os com menores taxas de homicídio estão em um primeiro grupo e os com maiores taxas de homicídio estão em um segundo grupo.

Ele continuou analisando o segundo grupo e viu que dentro dele há um subgrupo com um só indivíduo, New Mexico, e outro subgrupo com dois Estados, Louisiana e Texas. Ao examinar o Quadro 4.5, percebeu que a taxa de estupros por 100.000 habitantes do New Mexico é de 32,2, enquanto as taxas da Louisiana e do Texas ficam entre 22,5 e 25,5. Esse novo exemplo de agrupamento o fez ter uma ideia melhor de como os algoritmos de agrupamento podem nos ajudar, sugerindo “padrões similares de comportamento” entre observação (registros, linhas) de um conjunto de dados, principalmente, em casos em que há muitas observações (milhares ou mais) e muitas variáveis (dezenas ou centenas). Também ficou feliz que o seu supervisor não pediu, de imediato, que analisasse todos os 50 Estados de USArrest. Ele iria precisar de vários dias!

reflita

Reflita

"As regras de negócios mudaram. Em todos os setores de atividades, a difusão de tecnologias novas digitais e o surgimento de novas ameaças disruptivas estão transformando modelos e processos de negócios. A revolução digital está virando de cabeça para baixo o velho guia de negócios[...] Empresas constituídas antes do surgimento da internet enfrentam um grande desafio: muitas das regras e pressupostos fundamentais que governavam e orientavam a atuação e o progresso dos negócios na era pré-digital não mais se aplicam. A boa notícia é que a mudança é possível. As empresas pré-digitais não são dinossauros condenados à extinção. A ruptura não é inevitável. As empresas podem transformar-se e florescer na era digital."

Fonte: Rogers (2017, p. 11).

praticar

Vamos Praticar

Uma companhia internacional de vendas on-line deseja agrupar seus clientes com base em suas características comuns. Os gestores da companhia não têm rótulos predefinidos para esses grupos. Com base no resultado do agrupamento, eles definirão campanhas de marketing e de divulgação específicas para cada um dos diferentes grupos que vierem a ser definidos. As informações que dispõe sobre seus clientes incluem renda, idade, número de filhos, estado civil e grau de educação.

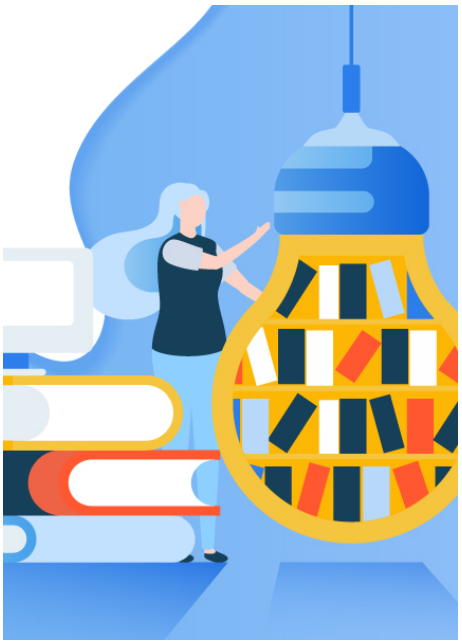
DUHAM, M. H. **Data mining:** introductory and advanced topics. Upper Saddle River, NJ: Pearson Education, 2003, p.125.

A respeito de algoritmos de agrupamento, assinale a alternativa correta:

- ○ **a)** Algoritmos de agrupamento só conseguem lidar com variáveis quantitativas. Sendo assim, parte das variáveis disponíveis para esse caso são irrelevantes.
- ○ **b)** Algoritmos de agrupamento podem ter as suas soluções verificadas por um supervisor e, dessa forma, saberemos se o resultado é bom ou ruim.
- ○ **c)** Seres humanos não possuem habilidade natural para agrupar e depois classificar, já que isso só pode ser realizado por meio de algoritmos.
- ○ **d)** O resultado de um problema de agrupamento pode ser usado, depois, como entrada para um problema de classificação.
- ○ **e)** Algoritmos de agrupamento são especializados no tratamento de conjuntos de dados exclusivamente qualitativos.

indicações

Material Complementar



LIVRO

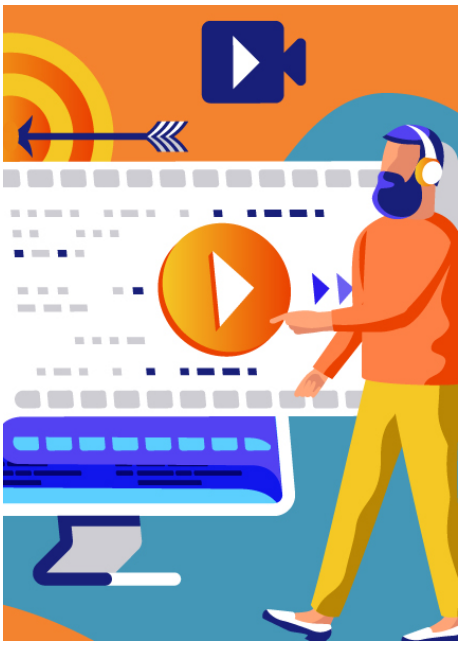
Segmentação de mercado em mídias sociais.

Marina Lauritzen Jácome.

Editora: Nova Edições Acadêmicas.

ISBN: 9786202042734

Comentário: Nos últimos anos, o mercado on-line varejista se expandiu no mundo de forma surpreendente e continua a crescer em ritmo forte. Todas grandes empresas que atuam nesse mercado têm interesse em estudar e segmentar o seu mercado e usam rotineiramente algoritmos de aprendizado de máquina para esse tipo de tarefa, incluindo algoritmos de agrupamento. Este livro trata desse tema, de segmentação de mercado, em mídias sociais. Sem dúvida, um tema atualíssimo.



FILME

A Rede Social (The Social Network)

Ano: 2010

Comentário: Este é um filme biográfico sobre o criador do Facebook. Nada mais atual no que se refere ao uso de computadores e da internet a serviço das pessoas. Além disso, sabemos que das redes sociais se extraem dados das pessoas, que são usados depois por empresas nas suas campanhas de marketing e divulgação. Essas empresas fazem uso, no seu dia a dia, de algoritmos de aprendizado de máquina, tanto internamente como na sua interação conosco. Acesso em: 30 jan. 2020.

TRAILER

conclusão

Conclusão

Muito bem, chegamos ao fim, mas todo fim é um recomeço, de outro ponto de partida, para uma nova jornada. Nessa que acabou, estudamos como a estatística é aplicada à ciência dos dados e, com a ajuda de alguns personagens, vimos exemplos de aplicações de: (1) Como fazer a predição do valor de imóveis; (2) Como fazer a predição da inadimplência com cartões de crédito; (3) Como fazer a predição do volume de vendas de um produto de varejo; e (4) Como estudar a violência urbana com algoritmos de agrupamento. Tudo isso de forma introdutória, apenas para lhe propiciar uma visão geral do grande potencial de aplicação dessas técnicas ao mundo dos negócios e à ciência em geral. Caberá a você decidir se tudo isso lhe interessa e, se assim for o seu desejo, prosseguir estudando esse tema e temas correlatos. Sem dúvida, há muitas oportunidades que, hoje, o mercado de trabalho oferece, voltadas às aplicações desse tipo, mas há também tantas outras mais que, seja qual for a sua decisão, estude, cuide do seu desenvolvimento, e tenha muito sucesso!

referências

Referências Bibliográficas

BRUCE, P.; BRUCE, A. **Estatística prática para cientistas de dados:** 50 conceitos iniciais. Rio de Janeiro: Alta Books, 2019.

CHOLLET, F. **Deep learning with Python.** Shelter Island, NY: Manning, 2018.

JAMES, G.; WITTEN, D.; HASTIE, T.; TIBSHIRANI, R. **An introduction to statistical learning:** with applications in R. New York: Springer, 2013.

McNEIL, D. R. **Interactive data analysis:** a practical primer. New York: Wiley, 1977.

R DOCUMENTATION. **USArrests**: violent crime rates by US states. Disponível em: https://www.rdocumentation.org/packages/datasets/versions/3.6.1/topics/USArrests. Acesso em: 15 jan. 2020.

ROGERS, D. L. **Transformação digital:** repensando o seu negócio para a era digital. São Paulo: Autêntica Business, 2017.

SAN DIEGO ZOO. **Animals and plants.** Disponível em: https://animals.sandiegozoo.org/animals/lion. Acesso em: 15 jan. 2020.

WICKHAM, H.; GROLEMUN, G. **R for data science:** import, tidy, transform, visualize, and model dada. Sebastopol (CA): O'Reilly Media, 2017.