FERNANDO AMARAL

APRENDA MINERAÇÃO DE DADOS

teoria e prática_

ALTA BOOKS
EDITORA
Rio de janeiro, 2016

"Sem dados você é apenas outra pessoa com uma opinião"

William Edwards Deming, estatístico, palestrante e autor

Sobre o Autor

Fernando Amaral, nascido em Santa Maria, Rio Grande do Sul, iniciou na área de tecnologia cuidando de CPDs nos anos 90, atuando com Unix e posteriormente Windows NT e Linux. Depois de alguns anos, passou a atuar como desenvolvedor de software na indústria de ERPs, onde utilizou Visual Basic e Delphi. Por volta de 2003, deixou o Rio Grande do Sul e passou a atuar na gestão de projetos de desenvolvimento e implantação de sistemas, quando se tornou especialista na plataforma .NET e em banco de dados SQL Server, desenvolvendo dezenas de soluções nas áreas ECM, produção, contábil, tributos, entre muitos outros. No início de 2010, retornou ao Rio Grande do Sul quando resolveu empreender pelo mundo da análise de dados, o que consequentemente trouxe um grande interesse em tecnologias como R, C#, Hadoop, Visualização, Data Mining, Estatística, BI, OLAP, NoSQL, Weka, entre outros. Desde então, tem atuado como gerente em grandes projetos em análise de dados em todo o Brasil, em projetos de Business Intelligence, análises preditivas, auditoria e monitoramento contínuo e Balanced Scorecard,

nas áreas de recursos humanos, contabilidade, marketing, tributos, fraudes, produção, em áreas de negócios como indústrias, governos, telefonia, energia, varejo, entre outros. Em 2014, passou a lecionar regularmente cursos relacionados à análise de dados, como programação em R, Mineração de Dados e Big Data, além de capacitação em gestão de projetos de análise de dados. Também, ministra palestra sobre o tema em diversas cidades brasileiras, além de manter alguns canais de mídias sociais com temas relacionados à mineração de dados. Atualmente, ocupa o cargo de Diretor de TI na Shelter IT, cuja sede se localiza na cidade de Novo Hamburgo, Rio Grande do Sul.

Agradecimentos

Há alguns meses resolvi escrever um novo livro, já sabendo que seria necessária muita dedicação. E de fato, foram várias noites escrevendo, capturando imagens, testando exemplos e criando atividades. Neste esforço devo agradecer primeiramente a minha família, pela compreensão e apoio. Também a toda a comunidade de "analytics", em especial aos vinculados à Universidade de Waikato que mantém o Weka, que na minha opinião, é desde 1993 o melhor software para ensinar e aprender Mineração de Dados. Por fim, a Editora Alta Books, por acreditar em mais este trabalho.

Sumário

Prefácio xvii

Capítulo 1 - Introdução à Mineração de Dados 1

Conceito 2
Mineração de Dados e Big Data 2
Aplicações 3
Ferramentas de Mineração de Dados 4
Estrutura de Dados 5
Classe, um Atributo Especial 6
Tipos de Dados 7
Tarefas de Aprendizado de Máquina 7
Supervisionado Versus Não Supervisionado 9
Tarefa Não é Algoritmo 9

Capítulo 2 - Introdução ao Weka 13

Projeto Weka 14
Versão 14
Instalação 15
Executando 15
Aplicativos 15
Explorer 16
Experimenter 16
KnowledgeFlow 16
Simple CLI 16
Formato Arff 16
Conjuntos de Dados de Exemplos 18

Weka Explorer: Preprocess 18
Edit 19
Filtros 21
Relação Atual 21
Atributos 22
Para Saber Mais 24

Capítulo 3 - Classificação 27
Introdução à Classificação 28
Aprendizado de Máquina 29
Modelos 30
Avaliando o que Foi Aprendido 32
Matriz de Confusão 34
Generalização Versus Superajuste 35
Problemas de Classe Rara 36
Problemas de Atributos Desconhecidos 36
Maldição da Dimensionalidade 36
Custo 37
Métricas de Modelos 37
Tipos de Algoritmos de Classificação 38
Árvores de Decisão 38
Regras 40
Naïve Bayes 41
Redes Bayesianas 42
Redes Neurais Artificiais 43
Máquina de Vetor de Suporte 44
Métodos de Grupos 46
Aprendizado Baseado em Instância 46

Capítulo 4 - Regressão 51
Correlação 52
Regressão Linear Simples e Múltipla 52
Regressão Logística 54
Árvores de Regressão e Outras Técnicas 55

Capítulo 5 - Classificação e Regressão no Weka 57
Compreendendo a Interface 58
Opções de Teste 60

Classe 62
Executando 62
Testando seu primeiro Classificador com Naïve Bayes 63
Árvore de Decisão com BFTree 70
Máquina de Vetor de Suporte com SMO 75
Visualizando a Árvore Graficamente com J48 76
Aprendizado Baseado em Instância com IB1 e Cálculo de Custo 77
Criando seu Próprio Classificador de Árvore de Decisão 81
Redes Neurais Artificiais: Multilayer Perceptron 88
Construindo uma Camada Oculta 89
Regressão Linear 92
Analisando Graficamente a Taxa de Erros 94

Capítulo 6 - Agrupamentos 97
Compreendendo Agrupadores 98
Funcionamento Básico 98
DBSCAN 100
Hierárquico 101

Capítulo 7 - Agrupamentos no Weka 105
Agrupamento com DBSCAN 106
Avaliando a Performance 106
Avaliando a Performance Graficamente 107
Agrupamento com Kmeans 108
Avaliando a Performance 109
Avaliando a Performance Graficamente 111

Capítulo 8 - Associadores 115
Entendendo Suporte e Confiança 116
Arquivo de Transações 117
Apriori 118
FP-Grow 118

Capítulo 9 - Associadores no Weka 121
Minerando uma Relação Tradicional com Apriori 122
Minerando Regras de Associação 123
Minerando Regras de Classificação 125
Minerando Cesta de Compras FPGrow 126

Capítulo 10 - Seleção de Atributos 129
Seleção de Atributos no Weka 130
Selecionando Atributos 130

Capítulo 11 - Filtros 133
Filtros no Weka 134
Criando um Atributo com o Grupo 135
Removendo Atributos Menos Importantes 136
Discretizando um Atributo 137
Adicionando Atributos Calculados 139

Capítulo 12 - Visualização 143
Visualizando Dados 144
Gerando Gráficos Individuais 145
Boundary Visualizer 146

Capítulo 13 - Mais Classificação 151
Aprendizado com Métodos de Grupos com Random Forest 152
Salvando o Modelo para Posterior Utilização 153
Fundamentos de Mineração de Texto no Weka 154
Preparando o Arquivo 156
Minerando Texto 158
Redes Bayesianas 159
Classificação com Redes Bayesianas 159
Potencializando Redes Bayesianas 160
Bayes Network Editor 162
Aprendendo uma Árvore Bayesiana 162
Construindo uma Árvore Bayesiana 165

Capítulo 14 - Experimenter 169
Preparando o Experimento 170
Executando o Experimento 171
Avaliando os Resultados 172

Capítulo 15 - KnowledgeFlow 175
Construindo um Fluxo 176
Avaliando o que Foi Produzido 178

Capítulo 16 - Linha de Comando com Simple CLI 183
Estrutura de Pacotes do Weka 184
Simple CLI 184
Principais Parâmetros para Classificadores 186
Obtendo Parâmetros do Weka Explorer 187

Capítulo 17 - Weka e Mineração de Dados, Alguns Insights Sobre o Futuro 189
Weka Versão 3.7 190
Pacotes 191
Instalando e Testando um Pacote 192

Referências 195

Solução de Exercícios 197
1. Introdução à Mineração de Dados 197
2. Introdução ao Weka 198
3. Classificação 199
4. Regressão 200
5. Classificação e Regressão no Weka 201
6. Agrupamentos 203
7. Agrupamentos no Weka 203
8. Associadores 205
9. Associadores no Weka 206
10. Seleção de Atributos 208
11. Filtros 208
12. Visualização 209
13. Mais Classificação 210
14. Experimenter 210
15. KnowledgeFlow 211
16. Linha de Comando com Simple CLI 212
Referências Bibliográficas 213

Índice 215

Prefácio

A tecnologia da informação está mudando o mundo de forma muito veloz: Internet, popularização da computação pessoal, smartphones e uma infinidade de dispositivos conectados, a chamada Internet das Coisas, estão transformando nossa vida em algo que nem os mais famosos filmes de ficção científica conseguiram prever. O fenômeno conhecido como Big Data, em que dados são produzidos em grande volume, velocidade e variedade, traz outra característica dessa nova era: dados produzidos de todas as formas, por dispositivos espalhados por toda a parte. Mas dados, mesmo que em grande volume, são apenas dados: é preciso produzir informação e conhecimento para explorar os benefícios que essa matéria-prima bruta pode trazer. Para isso, o dado precisa de alguma forma ser analisado. Dados analisados podem tornar as empresas mais lucrativas, os carros mais econômicos, podem reduzir fraudes, ajudar a criar campanhas publicitárias mais eficientes, remédios com menos efeitos colaterais, reduzir o efeito estufa e até salvar a vida de milhões de pessoas. É disto que este livro trata: analisar dados. Podemos analisar dados de diversas formas: ordenar uma coluna para ver qual produto custou mais caro é uma forma simples e eficiente de analisar dados. Mas não é disso que este livro trata, e sim da forma mais sofisticada e complexa de análise de dados: a mineração de dados, uma ciência irmã do aprendizado de máquina e da inteligência artificial. Com a mineração de dados, somos capazes de extrair informação e conhecimento desta fantástica matéria-prima que é o dado.

Falar em ciência sofisticada e complexa é de certa forma intimidador. O principal objetivo desta obra é mostrar que é possível aprender a aplicar a mineração de dados abstraindo a complexidade: é preciso entender os conceitos e as suas aplicações: não precisamos aprender a desenvolver algoritmos

complexos para minerar dados: usamos o que já existe pronto. Nesta obra, escolhemos a ferramenta Weka, um produto open source que você conseguirá utilizar em poucos minutos, que está repleta de dados de exemplo para praticarmos, e que tem os principais algoritmos e boas práticas do mercado. Os capítulos estão estruturados normalmente em teoria, seguidos por partes práticas e ordenados de forma lógica. Ao final de cada capítulo, exercícios ajudam a fixar os conceitos e a praticar o que foi aprendido. Ao longo do nosso curso iremos abordar desde os fundamentos, passando pela teoria fundamental da mineração até as principais tarefas, como classificação, regressão, agrupamentos, mineração de regressas de associação e seleção de atributos. Ao final da obra, o leitor deverá saber o que é mineração de dados, diferenciar as principais tarefas e saber quando cada uma delas deve ser aplicada, entender os conceitos por traz das principais famílias de algoritmos, compreender problemas que o minerador de dados pode encontrar e como resolvê-los e, acima de tudo, será capaz de resolver problemas de análise de dados com dados encontrados no seu dia a dia, de forma prática e rápida.

Esta obra é voltada a todos os interessados no assunto: analistas de sistemas, administradores de banco de dados, estatísticos, especialistas em marketing, desenvolvedores, cientistas de dados, analistas de negócios ou simplesmente curiosos.

1

Introdução à Mineração de Dados

Neste capítulo inicial, antes de falarmos de técnicas ou ferramentas, é preciso entender o que é a mineração de dados, qual seu papel na análise de dados, onde ela pode ser aplicada e que ferramentas existem no mercado. Depois, vamos entrar em alguns conceitos mais técnicos, como a forma em que os dados devem estar estruturados na mineração, quais os tipos principais de dados que são minerados e vamos entender o que é uma classe, um conceito fundamental e importantíssimo para todo o restante do estudo. Finalmente, vamos estudar as tarefas e algoritmos em mineração de dados e entender qual a diferença entre tarefa supervisionada e não supervisionada.

Conceito

Mineração de dados são processos para explorar e analisar grandes volumes de dados em busca de padrões, previsões, erros, associações entre outros. Normalmente a mineração de dados está associada ao aprendizado de máquina: uma área da inteligência artificial que desenvolve algoritmos capazes de fazer com que o computador aprenda a partir do passado: usando dados de eventos que já ocorreram.

O aprendizado de máquina é capaz de identificar padrões que dificilmente seriam identificados a "olho nu" ou mesmo usando técnicas triviais de análise de dados, como filtros, junções, pivôs ou agrupamentos. Para exemplificar, observe os dados na Figura 1.1: temos um registro de uma solicitação de crédito. Observe que seu histórico de crédito anterior na coluna *credit_history* indica que houve atrasos anteriores. A tendência numa primeira análise é de não conceder um novo crédito. Porém, um algoritmo de classificação nos recomenda o contrário: ele será um bom pagador de novos empréstimos, com uma chance de acerto de 75% na sua previsão.

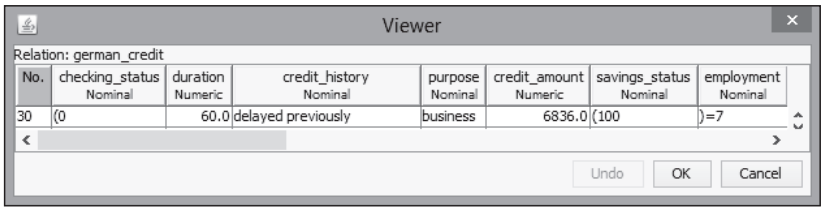


Figura 1.1: Solicitação de Crédito

Mineração de Dados e Big Data

Se você está interessado em aprender Mineração de Dados, provavelmente deve ter lido muita coisa sobre Big Data ultimamente. Nessa seção vamos entender como Big Data e Mineração de Dados estão intimamente relacionados.

Big Data é o fenômeno de produção de informação com velocidade, volume e variedade. Estes três "Vs" devem ser vistos neste contexto de forma extrema: muito embora Big Data esteja primariamente associado a produção de grande volume de dados, tão grandes que as tecnologias e modelos existentes até então

não eram capazes de processá-los, os dados também são produzidos em grande variedade, e aqui entra principalmente a questão dos dados não estruturados, aquela informação produzida sem uma estrutura fixa, diferente da tradicional linha e coluna que encontramos em planilhas e em bancos de dados tradicionais, e tudo isso produzido em grande velocidade.

Como todo fenômeno, o Big Data tem uma causa. Entre as principais, está a popularização da internet, o baixo custo de equipamentos e dispositivos tecnológicos, em especial os sensores, custo de armazenamento de dados cada vez mais baixos e dispositivos conectados. Essa última causa, dispositivos conectados, é outro fenômeno de proporções épicas e que está mudando o mundo como conhecemos: a Internet das Coisas, ou *Internet of Things* em inglês, é o nome dado à bilhões de dispositivos que são conectados à internet, e é claro, produzindo dados.

Entendido o que é Big Data, vamos ver então qual sua relação com Mineração de Dados. Já falamos brevemente no prefácio da obra que o dado é uma matéria-prima bruta. Seu valor surge quando ele é analisado. Analisar dados produz informação e conhecimento. Aqui então a coisa começa a ficar interessante: não adianta ter terabytes de dados sobre seus clientes se você não sabe se ele está satisfeito, que tipo de produto ele procura, o quanto ele está disposto a mudar para a concorrência etc. Mineração de Dados não é a única forma de transformar dados em informação e conhecimento: uma simples ordenação de clientes pelo valor da compra já produz informação. Porém, Minerar Dados é a forma mais sofisticada, complexa e difícil de analisar dados. Em consequência, o resultado pode trazer insights sobre o negócio que nenhuma outra técnica seria capaz de produzir. A Mineração de Dados já existia antes de começarmos a falar em Big Data, mas com o fenômeno, esta ciência ganhou uma posição de destaque: tenha volumes gigantescos de dados, mas seja capaz de analisá-los.

Aplicações

A mineração de dados tem sua aplicação cada vez mais difundida em áreas que antes sequer poderíamos imaginar uma aplicação prática, pois eram modelos de negócios em que os dados não se encontravam armazenados digitalmente. Um exemplo é análise de sentimento utilizado por empresas para avaliar a reputação da empresa em redes sociais. Além das tradicionais aplicações em

marketing, hoje a mineração de dados e o aprendizado de máquina encontram aplicações na medicina, educação, processamento de linguagem natural, bioinformática, detecção de fraude, reconhecimento de fala, finanças, robótica, sistemas de recomendação, mineração de texto entre muitos outros.

A seguir, alguns exemplos de aplicações em algumas áreas de negócio:

Marketing:

- Quais clientes irão responder a quais promoções?
- Quais combinações de produtos mais vendem?
- Quais clientes irão comprar mesmo sem ofertas?
- Identificação de consumidores alfa[1]
- Churn analysis: Quais clientes tendem a abandonar a empresa?

Educação:

- Quais alunos irão abandonar o curso e por quê?
- Quais alunos são mais fiéis?
- Quais alunos têm maior probabilidade de voltar a fazer novos cursos?
- Quais cursos serão mais rentáveis?
- Quais cursos, com quais características, atraem mais alunos?

Recursos Humanos:

- Qual o perfil de talentos é mais adequado para quais vagas?
- Qual o perfil de funcionários que abandonarão o emprego e quando?
- Quais ações são efetivas na produtividade?
- Quais funcionários serão mais bem-sucedidos?

Finanças/Contabilidade

- Prever o desempenho financeiro da organização
- Mitigação de riscos futuros
- Apoio na escolha de investimentos e parceiros

Ferramentas de Mineração de Dados

Existe uma grande quantidade de ferramentas de mineração de dados, tanto comerciais como open source. Nas comerciais, alguns dos maiores fornecedores são: Microsoft, SAS, IBM e Oracle. Nas ferramentas Open Source, algumas mais populares são Weka, R e Orange.

[1] Consumidor alfa é aquele com potencial para atrair outros consumidores

Neste curso optamos por usar o Weka, os motivos para sua escolha são muitos, aqui citamos os principais:

- Por ser open source, pode facilmente ser baixado e utilizado, sem custo de aquisição
- É uma ferramenta madura, produzida desde o início dos anos 90, que possui um conjunto muito grande de algoritmos capazes de executar as mais diferentes tarefas de aprendizado de máquina
- Possui uma GUI, uma interface gráfica onde o usuário não precisa digitar códigos, tornando o aprendizado muito mais fácil e intuitivo

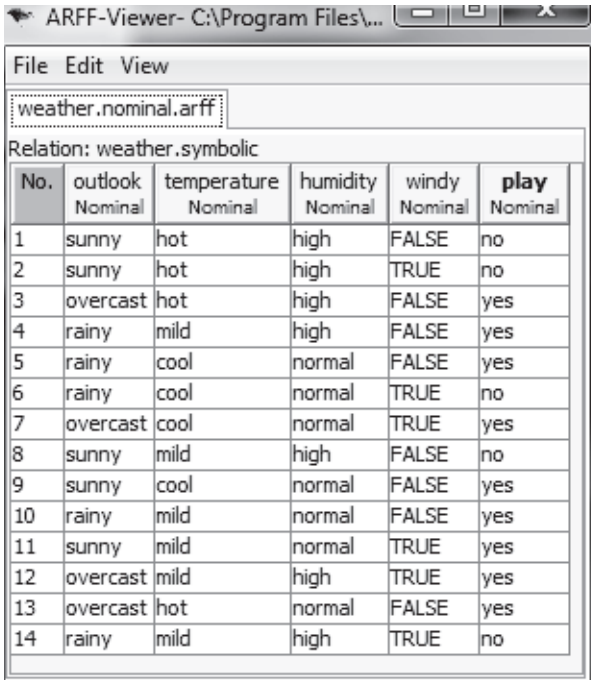
É importante destacar que os conceitos e técnicas que serão aprendidos aqui podem ser aplicados em qualquer ferramenta. O aprendizado de máquina é uma disciplina universal.

Estrutura de Dados

O aprendizado de máquina tem uma nomenclatura própria para se referir aos dados e suas estruturas. A melhor forma de conhecer esta nomenclatura é fazendo uma analogia com uma tabela de um banco de dados. Olhando logicamente, para que uma tabela de banco de dados represente um fato de negócio é preciso percorrer uma série de outras tabelas num processo de desnormalização. Por exemplo, uma venda estará representada em uma tabela, porém o vendedor, o cliente, o fornecedor e os produtos estarão em outras tabelas relacionadas. Porém, quando mineramos dados, a desnormalização e o tratamento de dados já deve ter ocorrido e o negócio já está representado por uma única forma tabular de dados. Voltando ao exemplo de vendas, para cada venda se espera encontrar na mesma linha o produto, o fornecedor, o cliente e o vendedor. Não podemos minerar uma estrutura de banco de dados, por exemplo, na terceira forma normal. Por isso, o que em banco de dados é uma tabela, em aprendizado de máquina é chamado de relação. A relação contém todas as características do negócio.

Uma tabela em banco de dados é composta por linhas e colunas. No aprendizado de máquina, cada coluna é um atributo, e cada linha é uma instância. Podemos fazer uma analogia de atributo com uma característica do negócio: por exemplo, o vendedor é uma característica da venda. Já a instância pode ser comparada pela ocorrência de um fato do negócio, ou seja, cada linha na relação é uma venda efetivada.

Observe a Figura 1.2, temos aqui a relação weather (tempo), que será utilizada diversas vezes ao longo do nosso estudo. Outlook, temperature, humidity, windy e play são atributos, ou características coletadas do tempo em determinado momento. Temos 14 instâncias (linhas), ou seja, dados sobre o aspecto do tempo foram coletados 14 vezes.


ARFF-Viewer- C:\Program Files\...

File Edit View

weather.nominal.arff

Relation: weather.symbolic

No.	outlook Nominal	temperature Nominal	humidity Nominal	windy Nominal	**play** Nominal
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Figura 1.2: Tabela de Tempo

Importante destacar que, às vezes, os atributos são referidos como dimensões de uma relação.

Classe, um Atributo Especial

A tarefa mais comum de aprendizado de máquina é a classificação. Vamos estudar classificação com mais detalhes em vários capítulos e seções seguintes, mas o que você precisa entender no momento, é que na classificação existe um atributo especial que é chamado classe: o objetivo é usar todos os atributos que compõem a relação para tentar prever a classe. Nos dados da Figura 1.2, play é a classe. Coletamos dados do tempo como aparência, temperatura, umidade e vento para tentar prever se naquele dia vamos poder ou não jogar.

Normalmente a classe é o último atributo. O Weka vai considerar por padrão que a classe está nesta posição, porém, é possível informar se ela estiver em uma posição diferente.

Tipos de Dados

Em mineração de dados existem dois grandes grupos principais de dados: contínuos, como números reais, e nominais, que podem ser uma descrição, um nome ou uma categoria, por isso podem ser denominados também dados categóricos. Dados discretos são dados finitos, normalmente valores inteiros.

Tarefas de Aprendizado de Máquina

O aprendizado de máquina pode ser dividido em quatro grandes grupos: classificação, regressão, agrupamentos e regras de associação.

Na classificação, que já foi brevemente descrita em seções anteriores, queremos descrever ou prever um atributo especial chamado classe.

Observe a Figura 1.3: Classificação: dentro dos retângulos, as formas já estão classificadas de acordo com seu tipo: triângulo, quadrado ou círculo. A forma à esquerda com uma interrogação ao centro ainda não teve seu tipo identificado. O aprendizado de máquina deve analisar quais são as características que definem cada forma e atribuir ela a um dos grupos. Usamos classificação para prever uma fraude, descobrir a qual espécie um animal pertence, prever uma doença ou classificar um tipo de fungo.

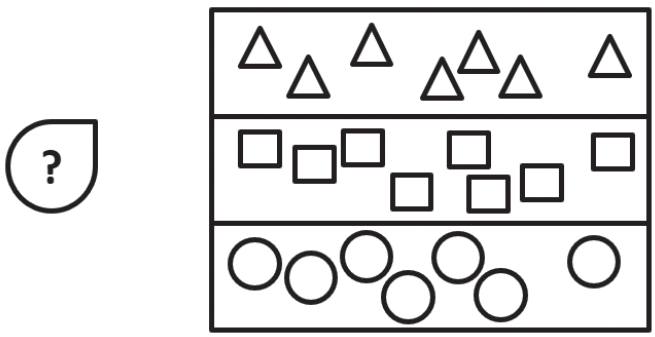


Figura 1.3: Classificação

A regressão é um tipo de classificação: enquanto na classificação a classe é um tipo de dado nominal ou categórico, na regressão a classe é numérica. Prever a altura de uma pessoa a partir do peso é um exemplo de tarefa de regressão.

Em agrupamentos, não existe classe. O objetivo é criar grupos e atribuir instâncias a estes grupos a partir das características, ou atributos destas instâncias.

Observe agora a Figura 1.4: desta vez não existem grupos de formas com os quais podemos comparar as características de uma forma desconhecida para ver onde eles se encaixam melhor. O agrupamento vai buscar semelhança entre as características dos próprios elementos e atribuir grupos a eles. Dependendo do tipo de agrupamento que utilizarmos, um elemento pode pertencer a mais de um grupo ou não ser agrupado, ou seja, ser considerado ruído. Exemplos de uso: identificar grupos de clientes e direcionar campanhas de marketing específicas; identificar tentativas de acesso à rede; categorizar uma nova espécie entre outros.

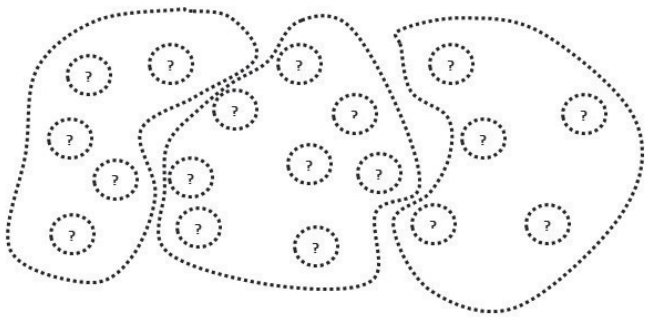


Figura 1.4: Agrupamentos

Regras de associação buscam a relação entre itens.

Na Figura 1.5, cada círculo pontilhado por um determinado número de formas. Algoritmos de regras de associação podem gerar muitas regras neste cenário, uma delas seria que a chance de existir um círculo havendo um triângulo é de 50%.