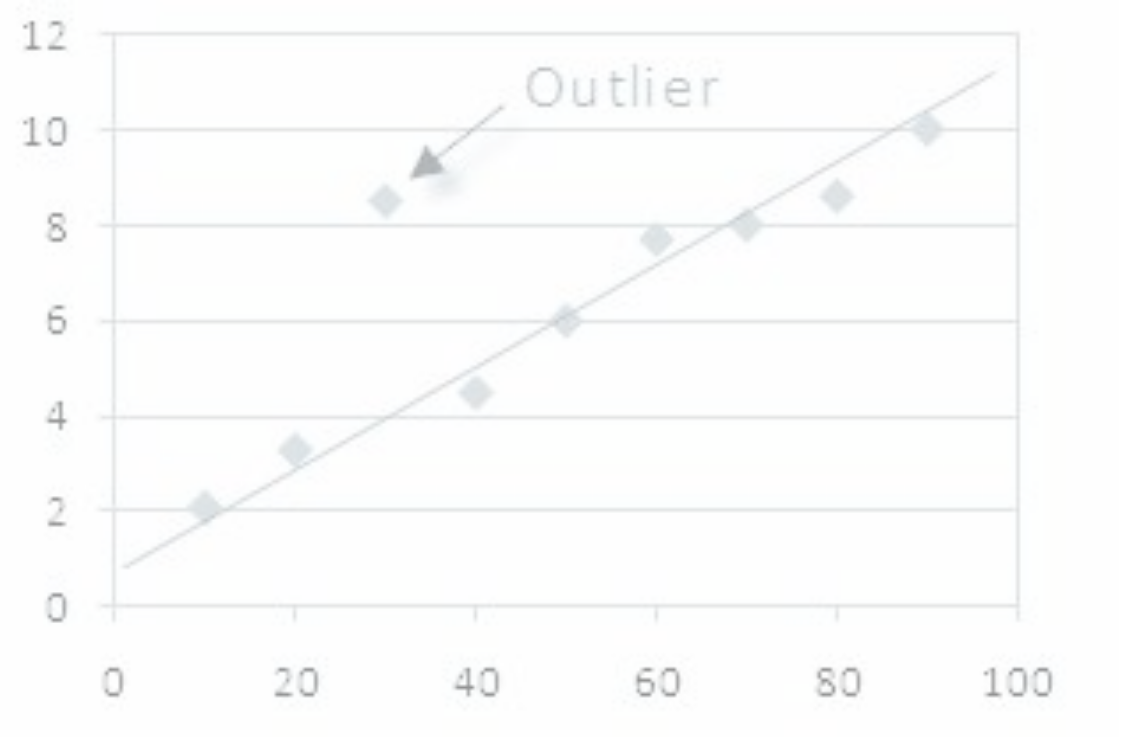


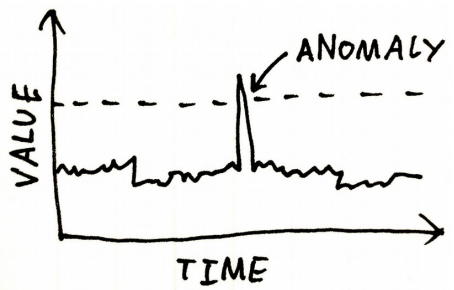
Aula 10

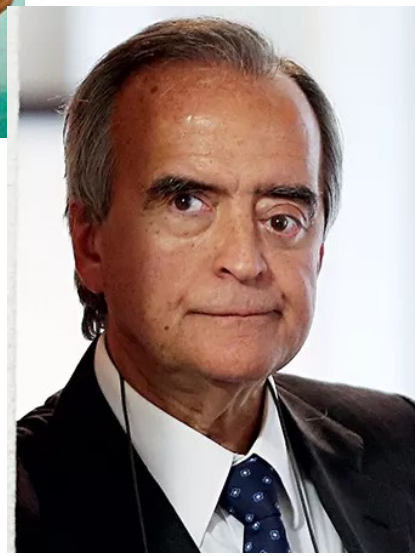
Detecção de Anomalias

Introdução

- Definições do termo "anomalia"
  - **Dicio**: "*Particularidade ou condição do que é anômalo, fora do comum.*"
  - **Priberam**: *"O que se desvia da norma, da generalidade."*
  - **Michaelis**: "*Estado ou qualidade do que é anômalo; anormalidade, defeito, irregularidade.*"
  - **Informal**: "*É, em um sentido amplo, tudo aquilo que se desvia de um padrão de normalidade.*"

Introdução





BRASIL : **Brasileirão Serie A**

#	Time	Pts
1	Corinthians	47
2	Gremio	39
3	Santos	35
4	Palmeiras	32
5	Flamengo	29
6	Sport	28
7	Cruzeiro	27
8	Atletico PR	26
9	Coritiba	25
10	Fluminense	25
11	Botafogo	25
12	Vasco da Gama	24
13	Bahia	23
14	Atletico MG	23
15	Ponte Preta	22
16	Chapecoense	22
17	Sao Paulo	19
18	Vitoria	19
19	Avai	18
20	Atletico GO	12

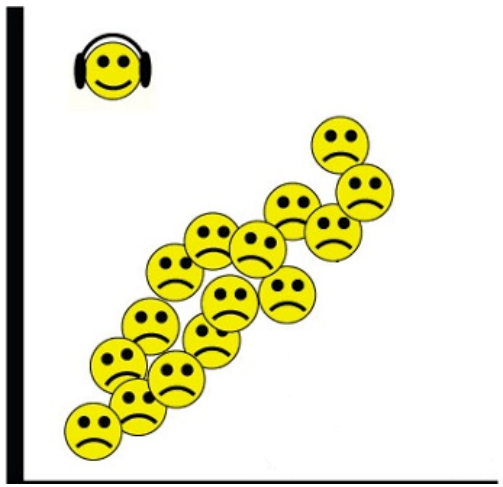
Introdução

- Na **detecção de anomalias**, a tarefa é **encontrar objetos que são diferentes dos demais objetos**

https://postimg.org/image/rlxhwmcq1/

Introdução

- Esses objetos diferentes também são comumente referenciados por *outliers*
- A detecção de anomalias também é conhecida por:
  - Análise/detecção de *outliers*
  - Detecção de desvios
  - Mineração de exceções



Aplicações

- A detecção de outlier podem ser aplicados em muitos cenários:
  - **Detecção de Fraudes**
    - O comportamento de compras de um ladrão é provavelmente diferente do proprietário do cartão
    - As companhias de cartões de crédito visam detectar um padrão de alteração do compartamento típico
    - O mesmo conceito pode ser usado para outros tipos de fraudes

https://under-linux.org/attachment.php?attachmentid=54598&d=1410380628

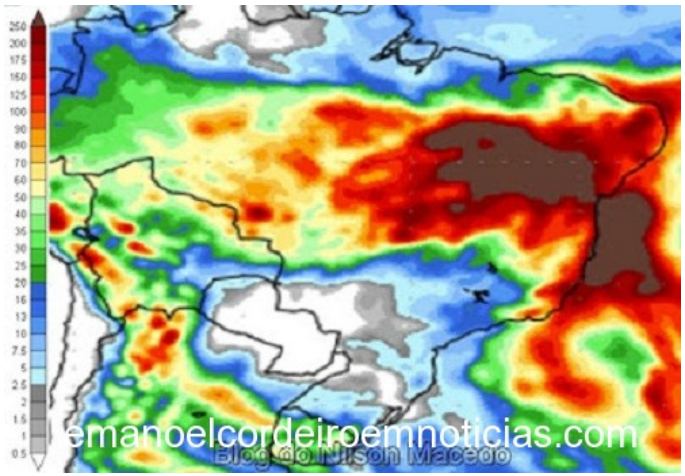
Aplicações

- A detecção de outlier podem ser aplicados em muitos cenários:
  - **Detecção de Intrusão**
    - Muitas das intrusões só podem ser detectadas por meio de monitoramento de sistemas e redes na busca por comportamento não comuns
    - Tráfego de rede incomum
    - Forma de digitar a senha incomum (*keystroke recognition* ou *keystroke dynamics*)

https://under-linux.org/attachment.php?attachmentid=54598&d=1410380628

Aplicações

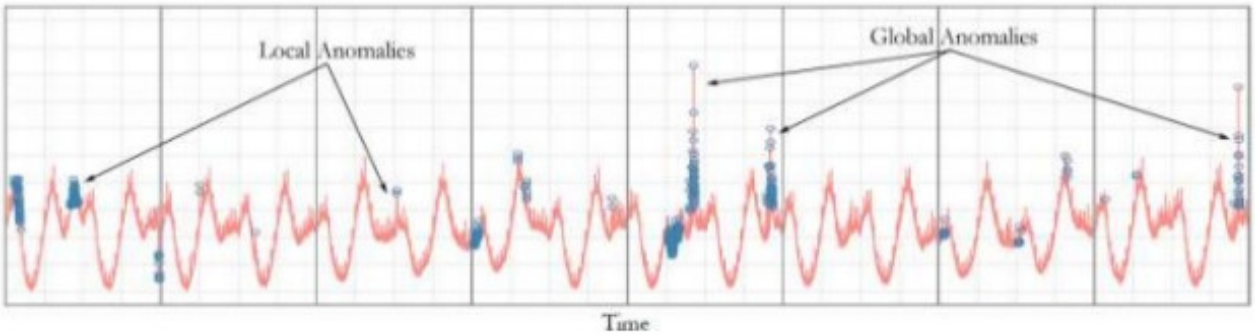
- A detecção de outlier podem ser aplicados em muitos cenários:
  - **Distúrbios nos Ecossistemas**
    - O objetivo é prever a ocorrência de eventos naturais atípicos que podem ter alto impacto na vida do homem, como furacões, enchentes, queimadas, tsunamis, ...



http://emanoelcordeiroemnoticias.com/wp-content/uploads/2016/01/Chuva-1b.jpg

Aplicações

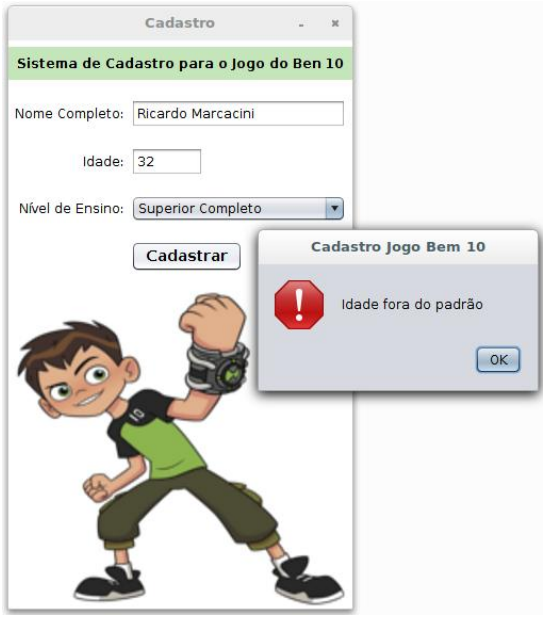
- A detecção de outlier podem ser aplicados em muitos cenários:
  - **Saúde Pública / Medicina**
    - Pode-se detectar problemas em campanhas de vacina ou de prevenção de doenças caso o número de pessoas infectadas com determinada doença estiver acima do normal
    - Para um paciente em particular, sintomas ou resultados de testes incomuns podem indicar potenciais problemas de saúde



https://image.slidesharecdn.com/slnewjerseyhl7fullpresentation-161111155526/95/splunking-hl7-healthcare-data-for-business-value-40-638.jpg?cb=1478879782

Aplicações

- A detecção de outlier podem ser aplicados em muitos cenários:
  - **Sistemas de cadastro**
    - Detectar se o valor informado em um determinado campo está dentro dos padrões dos valores inseridos no sistema anteriormente



Aplicações

- As aplicações apresentadas anteriormente podem fazer parte de um SAD dependendo do tipo de instituição em que o SAD está sendo aplicado
- Entretanto, algumas situações de análise de anomalias são mais comuns em SADs, principalmente naqueles utilizados em instituições privadas
- Basicamente, eles visam analisar algum valor de interesse e verificar se esse valor está dentro do padrão

Aplicações

- Por exemplo pode-se verificar se os seguintes valores estão anormais (geralmente abaixo do padrão) em um determinado período de tempo
  - Número de produtos vendidos
  - Número de peças fabricadas
  - Lucro
  - Valor pago em um determinado produto

http://blogs.uai.com.br/pergunteaopsicologo/wp-content/uploads/sites/70/2016/09/desempregado-300x226.jpg

Causas

- *Outliers* surgem nos conjuntos de dados por diferentes causas
- As causas mais comuns são:
  - **Dados de diferentes classes**
    - Os *outliers* pertencem à uma classe ou padrão diferente daqueles apresentados no conjunto de dados
    - Por exemplo, alguém comentedo uma fraude pertence à uma classe diferente dos usos legítimos do cartão de crédito
    - O mesmo vale para intrusões, resultados de testes anormais, ...

Causas

- As causas mais comuns são:
  - **Variações naturais**
    - Muitos conjuntos de dados podem ser modelados por distribuições estatísticas
    - A mais comum é a distribuição normal (Gaussiana) $\rightarrow$ a probabilidade de um objeto diminui rapidamente conforme o objeto se distancia do centro da distribuição
    - Objetos extremos, muito distantes do centro da distribuição normal, podem ser interessantes em aplicações práticas

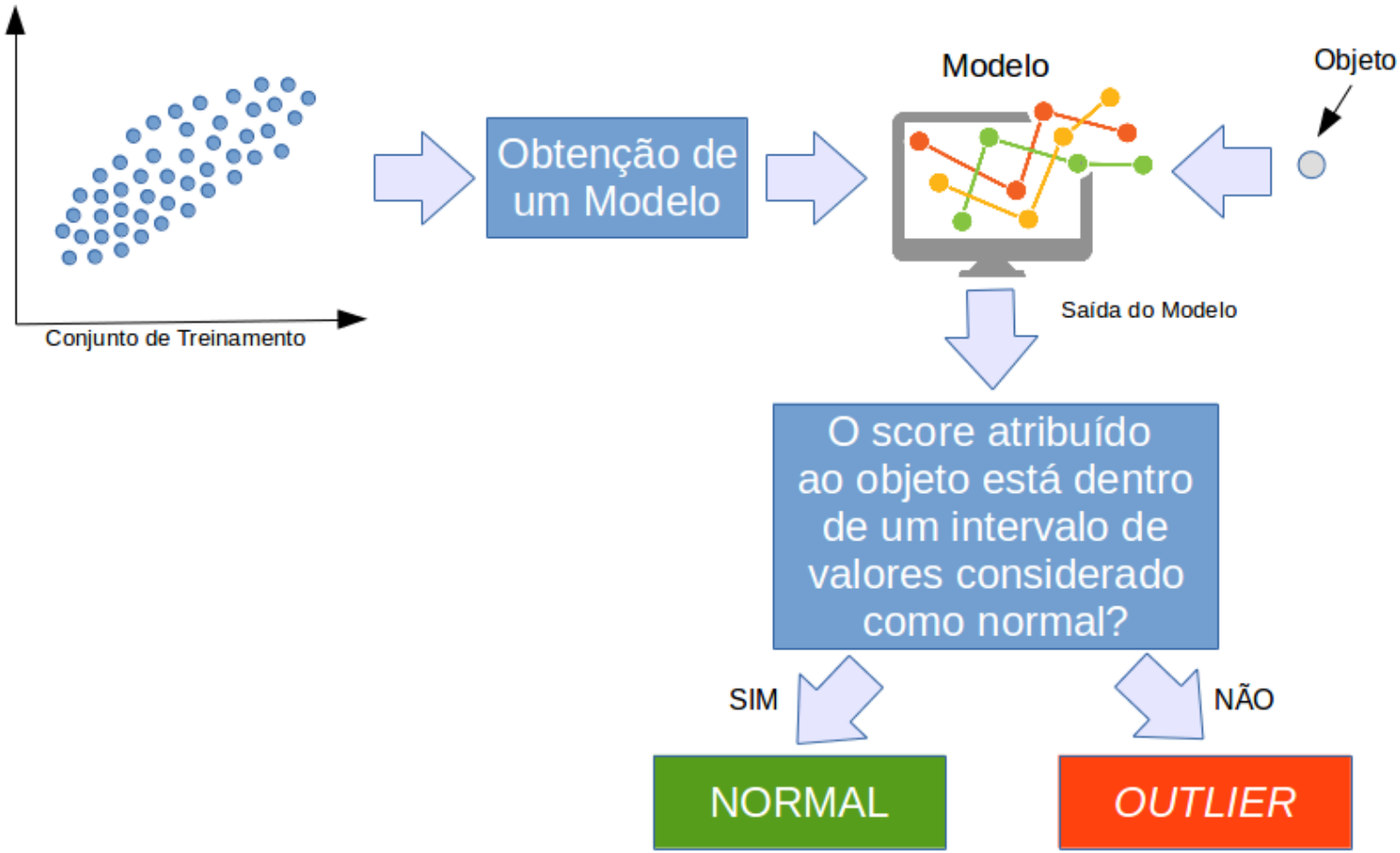
Causas

- As causas mais comuns são:
  - **Erros na coleta de dados**
    - Ao coletar dados por medições automáticas, pode haver problemas no aparelho de medição ou mesmo no processo de gravação dos dados
    - Tais valores coletados erroneamente podem alterar os resultados de processos de extração de padrões
    - A detecção/remoção de tais valores é normalmente aplicada em passos de pré-processamento de dados

Abordagens - Características dos Métodos

- Existem algumas características que podem ser utilizadas para categorizar os métodos para detecção de *outliers*
- Vale ressaltar que um método pode pertencer a mais de uma categoria
- As categorias mais comuns são:
  - **Métodos baseados em modelos**
    - Primeiro é construído um modelo dos dados
    - As anomalias serão aqueles objetos que não pertencem ao modelo dos dados

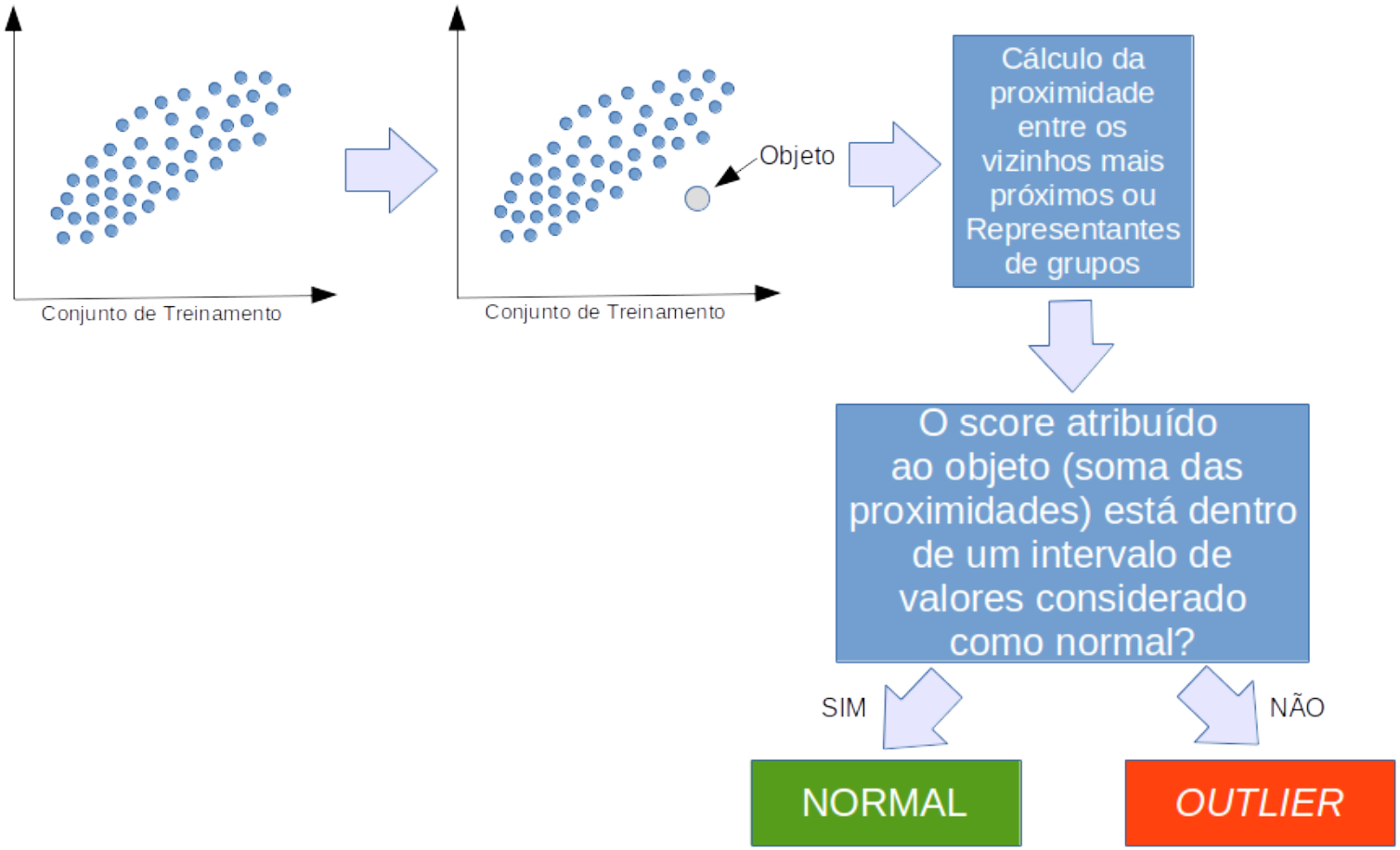
Métodos baseados em modelos



Abordagens - Características dos Métodos

- As categorias mais comuns são:
  - **Métodos baseados em proximidade**
    - Como visto nas aulas anteriores, é possível definir medidas de proximidades entre os objetos
    - Neste caso, objetos anômalos são aqueles distantes da maioria dos objetos não anômalos
    - Quando os dados podem ser mostrados em um scatter-plot bi- ou tri-dimensional, a detecção baseada em distância pode ser feita visualmente

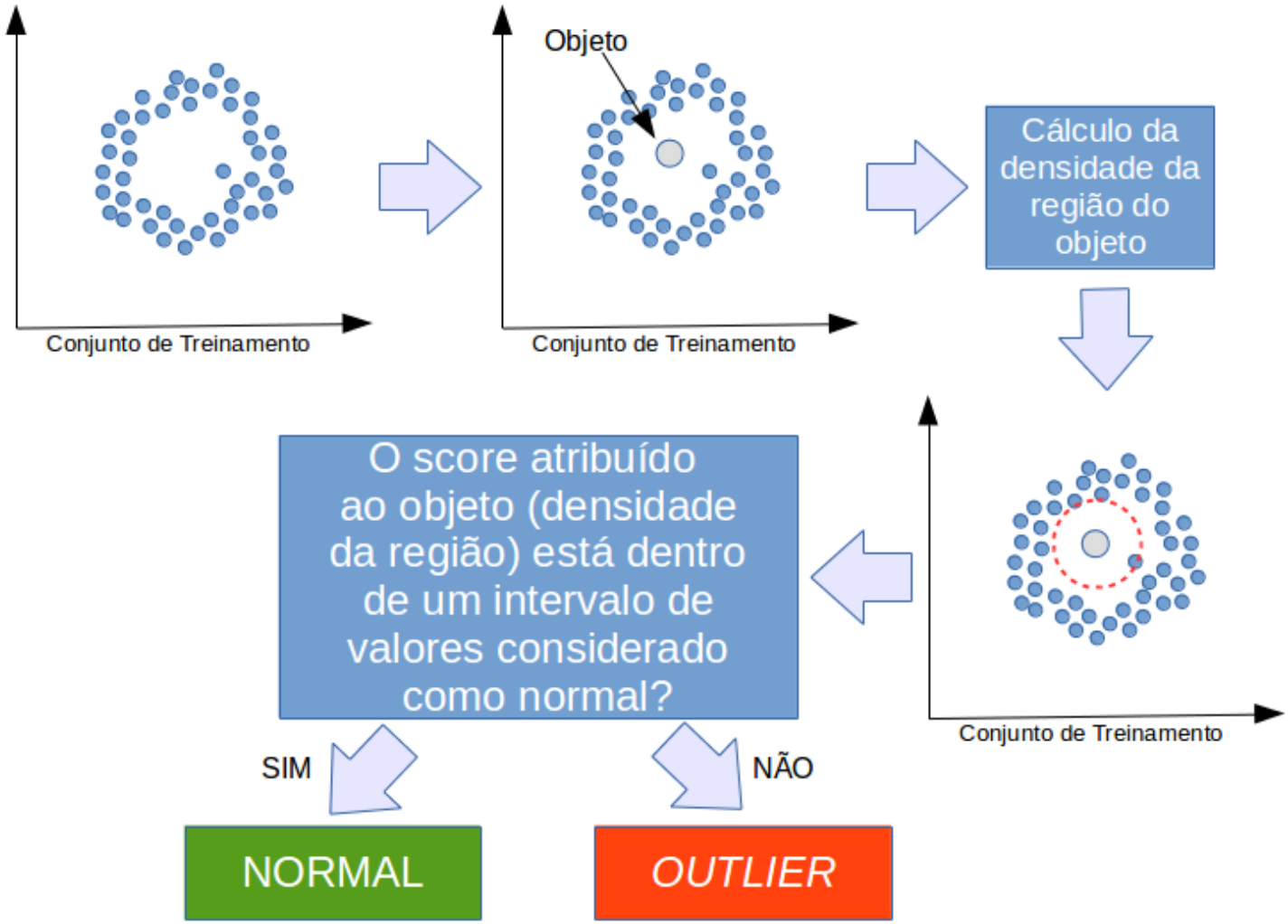
Métodos baseados em proximidade



Abordagens - Características dos Métodos

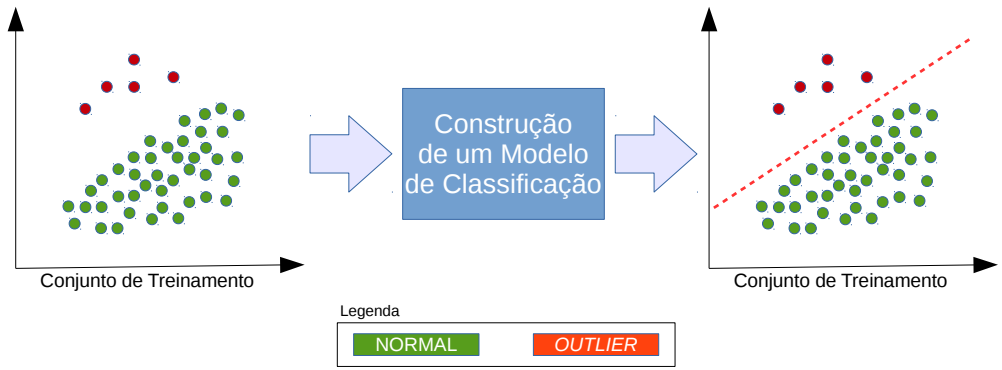
- As categorias mais comuns são:
  - **Métodos baseados em densidade**
    - Objetos que estão em regiões de baixa densidade estão relativamente distante de seus vizinhos e, portanto, podem ser considerados anômalos

Métodos baseados em densidade



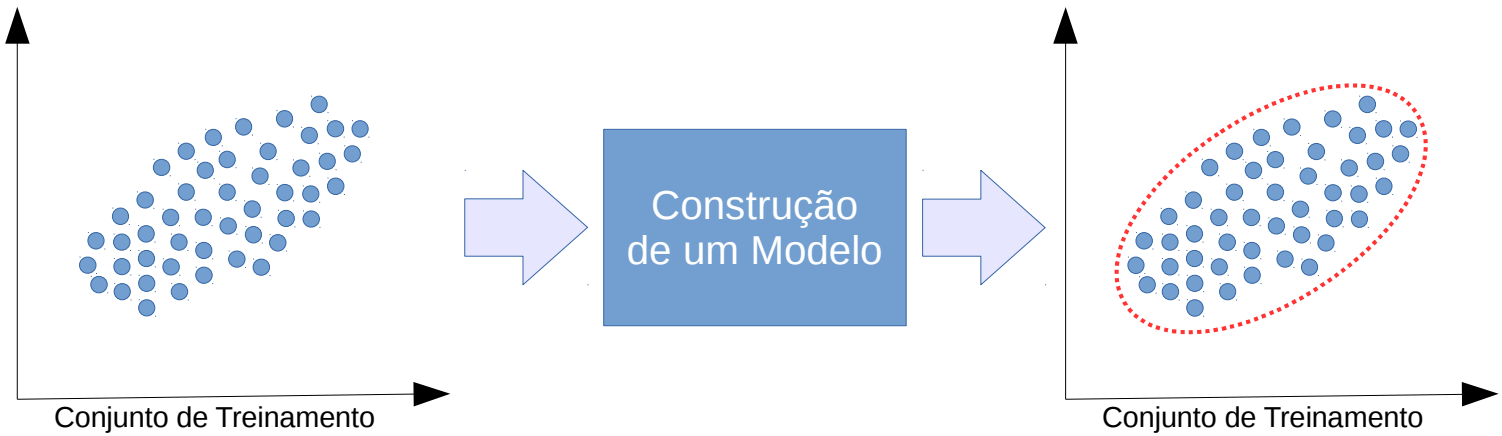
Abordagens - Presença de Exemplos Rotulados

- Considerando a presença (ou não) de exemplos rotulados, têm-se 3 abordagens básicas para a detecção de *outliers*
  - **Detecção de anomalias supervisionada**
    - Requerem a existência de um conjunto de treinamento com objetos normais e anormais (esse último é opcional)
    - Tomar cuidado com tal abordagem pois a classe anormal é relativamente rara
    - Aprendizado com classes desbalanceadas tende a classificar a maior dos exemplos como pertencente à classe majoritária



Abordagens - Presença de Exemplos Rotulados

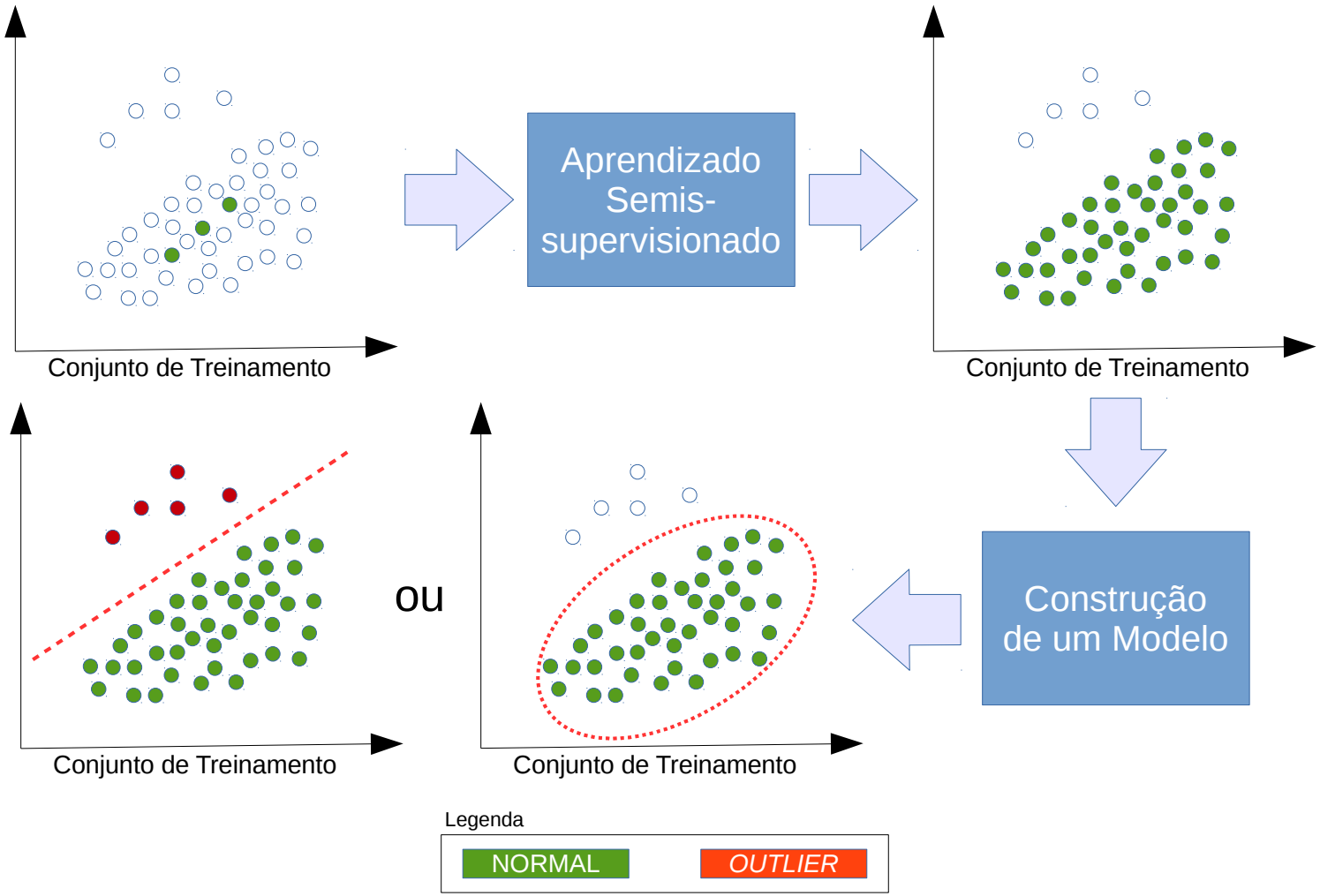
- **Detecção de anomalias não supervisionada**
  - Em muitas situações práticas, os rótulo dos exemplos não estão disponíveis
  - Neste caso, o objetivo é atribuir um score para um objeto que reflete o grau ao qual o objeto é anômalo
  - Normalmente considera-se que todos os objetos do conjunto de dados são objetos normais



Abordagens - Presença de Exemplos Rotulados

- **Detecção de anomalias semissupervisionada**
  - Em algumas situações, o conjunto de dados de treinamento contém exemplos de objetos rotulados como normais, porém, não possuem objetos rotulados como anormais
  - No cenário semissupervisionado, o objetivo é encontrar um conjunto de objetos normais e anômalos dentro dos objetos não rotulados
  - Com isso, pode-se tanto ampliar o número de exemplos rotulados para utilizados em um modelo qualquer para detecção de anomalias, ou ainda considerar os exemplos classificação como anômalos para induzir um modelo de classificação

Abordagens - Presença de Exemplos Rotulados



Abordagem Estatística

- As abordagens estatísticas são abordagens baseadas em modelos
- A maioria das abordagens estatísticas são baseadas na construção de um modelo de distribuição de probabilidades
- Um objeto é considerado como sendo uma anomalia (ou não) de acordo com a sua pertinência (normalmente probabilidade) de pertencer ao modelo
- **Definição probabilística de anomalia**: uma anomalia é um objeto que tem baixa probabilidade com respeito ao modelo de distribuição de probabilidade dos dados

Abordagem Estatística

- Um modelo de distribuição de probabilidade é criado utilizando o conjunto de dados para estimar os parâmetros de uma distribuição especificada pelo usuário

- **Ex:** se for assumido que os dados possuem uma distribuição normal, a média e o desvio padrão da função Gaussiana são obtidos analisando os dados

Distribuição Normal Univariada

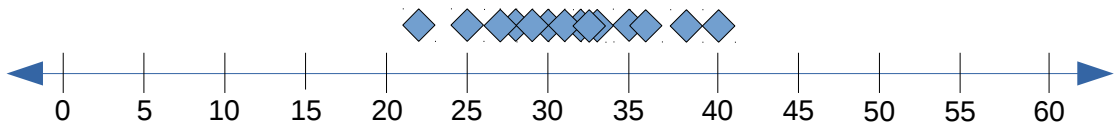
- A distribuição Gaussiana (normal) é uma das mais distribuições mais utilizadas na literatura e possui 2 parâmetros:
  - μ: centro da distribuição
  - σ: desvio padrão da distribuição
- Função Gaussiana

$$f(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

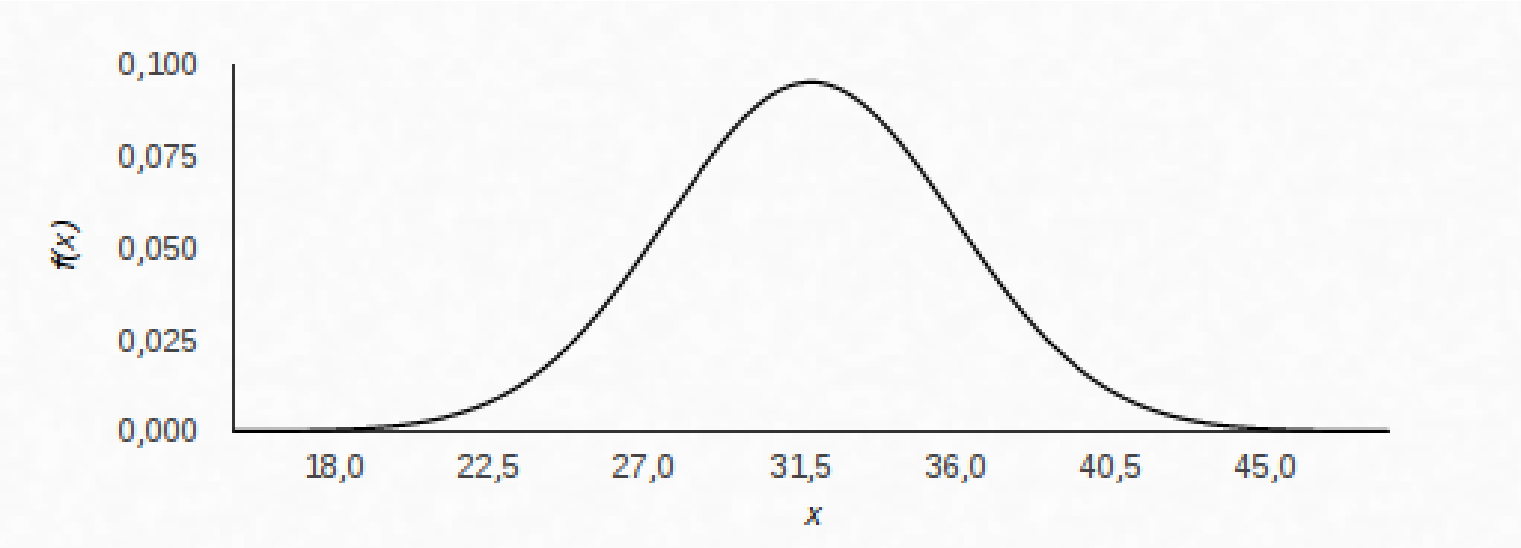
- Após obtida a distribuição, um objeto é classificado como *outlier* se a sua probabilidade de pertencer ao modelo for menor que um limiar *l* definido pelo usuário

Distribuição Normal Univariada

ID	Valor
1	25,0
2	28,0
3	35,0
4	30,0
5	32,0
6	29,0
7	31,0
8	33,0
9	34,0
10	27,0
11	30,0
12	36,0
13	40,0
14	38,0
15	22,0



Distribuição Normal Univariada



VER: `https://planetcalc.com/4986/`

Exercício

Considere como conjunto de exemplos normais os exemplos apresentados na tabela abaixo. Assumindo que os mesmos pertencem à uma distribuição normal, Verifique se os valores $R\$11,50$ e $R\$20,00$ são normais ou *outliers*. Considere um exemplo normal, aquele cuja probabilidade de pertencer à distribuição é maior que 0,15.

ID	Gastos (R$)
1	10
2	12
3	11

Distribuição Normal Multivariada

- Para considerar um conjunto de dados multivariados, isto é, compostos por mais de um atributo, a abordagem será semelhante à distribuição univariada
- Porém, agora termos que considerar a distância de um objeto, o qual é composto por um conjunto de atributos, para o centro da distribuição normal
- O passo seguinte, para definir se um objeto é um *outlier* (ou não) permanece o mesmo

Distribuição Normal Multivariada

- A parte crítica, portanto, é o cálculo da distância entre um objeto e o centro da distribuição
- Como cada atributo pode ter sua própria variância, precisamos de uma medida para considerar tal fator no cálculo da distância
- Para isso, o mais aconselhável é utilizar a distância de Mahalanobis

Distribuição Normal Multivariada

- Distância de Mahalanobis

$$mahalanobis(\mathbf{x}, \overline{\mathbf{x}}) = (\mathbf{x} - \overline{\mathbf{x}})\mathbf{S}^{-1}(\mathbf{x} - \overline{\mathbf{x}})^T$$

na qual $\mathbf{S}$ é a matriz de covariância dos dados $(m \times m)$ e m é o número de atributos

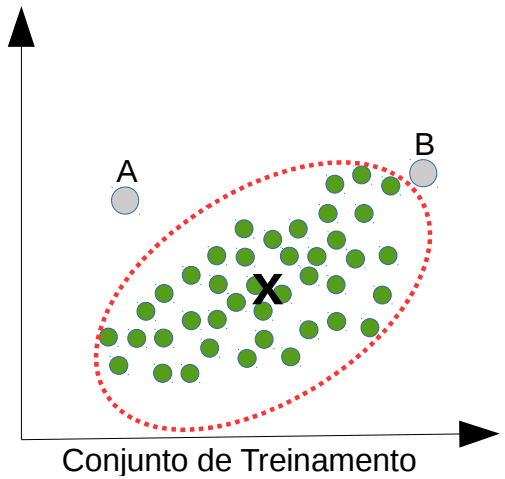
- A distância de Mahalanobis de um ponto para a média dos pontos é diretamente relacionada com a probabilidade do ponto pertencer a distribuição
- Mais precisamente, a distância de Mahalanobis é igual ao log da densidade de probabilidade do ponto mais uma constante

Distribuição Normal Multivariada

- A função de distribuição normal multivariada considerando a distância de Mahalanobis é dada por:

$$f(\mathbf{x}, \overline{\mathbf{x}}, \mathbf{S}) = \frac{1}{\sqrt{|\mathbf{S}|}(2\pi)^{m/2}} e^{-\frac{1}{2}(Mahalanobis(\mathbf{x},\overline{\mathbf{x}},\mathbf{S})^2}$$

na qual $|\mathbf{S}|$ é a determinante da matriz de covariância, e m é o número de dimensões do conjunto de dados



Considerações Gerais

- As abordagens estatísticas possuem uma base sólida baseadas em técnicas comuns da estatística, tais como estimar os parâmetros de uma distribuição
- Quando há conhecimento suficiente dos dados, pode-se aplicar testes para se verificar a correta distribuição dos dados
- Há uma grande quantidade de testes para distribuições univariadas
- Entretanto, há poucos testes para distribuições multivariadas e estes testes não costumam funcionar corretamente para conjuntos de dados com altas dimensionalidades

Abordagem Baseada em Proximidade

- Apesar de haver vários métodos baseados na ideia de detecção de anomalias considerando medidas de proximidade, a noção básica da abordagem é bem simples
- Esta é uma abordagem mais geral e mais facilmente aplicada uma vez que é mais fácil determinar a proximidade entre objetos do que sua real distribuição
- Uma das maneiras mais simples de medir se um objeto está distante da maioria dos pontos é considerar sua distância para os k-vizinhos mais próximos

k-Vizinhos Mais Próximos

- O score que será utilizado para determinar se um objeto é um *outlier* (ou não) é a média das proximidades dos seus k-vizinhos mais próximos
- Lembrando que se for utilizada uma medida de distância/dissimilaridade, um objeto será classificado como *outlier* se seu score for $> \alpha$
- Caso contrário, se for utilizada uma medida de similaridade, um objeto será classificado como *outlier* se seu score for $< \alpha$

k-Vizinhos Mais Próximos

- Pode ser sensível ao valor em relação ao valor de k
- Se o valor de k é muito pequeno, por exemplo $k = 1$, um pequeno número de vizinhos será usado para definir se um ponto é um *outlier*, e isso por causar classificações errôneas
- Por outro lado, um valor de k muito grande, pode fazer com que ponto muito distantes sejam considerados como vizinhos, e isso faça com que um objeto normal seja classificado como *outlier*
- Essa abordagem tem custo computacional maior que das outras abordagens, pois é preciso comparar o novo exemplo com todos os exemplos armazenados

Exercício

Considere $k = 1$ eu um limiar de distância $> 0,2$ para um exemplo ser considerado um *outlier*. Considere também os dados de compra de um usuário apresentadas abaixo. Considere agora que um cliente fez uma compra em uma quinta-feira no valor de R$150,00 reais e em uma sexta-feira no valor de R$210,00. Qual das duas será considerada um *outlier*?

ID	Dia da Semana	Gastos (R$)
1	Segunda	50
2	Terça	30
3	Quarta	30
4	Quinta	50
5	Sexta	250

Agrupamento de Dados

- Uma outra abordagem é primeiro agrupar os objetos e então calcular o grau com que um objeto pertence à um grupo
- Para algoritmos de agrupamento baseados em protótipos, a distância de um objeto para o centro do seu grupo pode ser usada para medir o grau ao qual um objeto pertence à um grupo $\rightarrow$ *score*
- Assim como no agrupamento baseado em protótipos, definir o número de grupos em tal abordagem é essencial
  - k pequeno $\rightarrow$ grupos muito grandes e vários objetos podem ficar distante dos centros dos grupos, isto é, serem considerados *outliers*
  - k grandes $\rightarrow$ grande número pequenos grupos, fazendo com que cada ponto fique muito próximo do centro do seu grupo

Agrupamento de Dados

- O algoritmo *k*-Means possui complexidade de tempo e espaço lineares, e portanto, é computacionalmente eficiente
- A definição de grupos e *outliers* são complementares $\rightarrow$ é possível achar ambos ao mesmo tempo
- Vale relembrar que algoritmos baseados em protótipos podem ter os representantes de grupos afetados por *outliers*

Exercício *k*-Means

Considere o conjunto de dados normais apresentados abaixo. Obtenha o centroide desse grupo. Depois, calcule a distância de uma nova entrada {Qtd. Peças Produzidas = 110; Custo R$ = 6500}. Considere que um *outlier* é aquele cuja distancia (considerando os atributos normalizado pela técnica *Max*). Considere um *outlier* aquele cuja distância euclidiana para o centro do grupo seja maior ou igual a 0,1.

ID	Qtd. Peças Produzidas	Custos (R$)
1	100	6050
2	125	6180
3	120	6250
4	115	6000
5	130	6100

Abordagem Baseada em Densidade

- Em um ponto de vista baseado em densidade, um `outlier` é um objeto que está em regiões de baixa densidade
- O *score* de um objeto é o inverso da densidade ao redor do objeto
- Neste caso, um objeto será considerado um *outlier* se seu *score* for maior que um limiar, isto é, $score > \alpha$

Inverso da Densidade

- A detecção de anomalias baseada em densidade está relacionada com a abordagem baseada em proximidades, uma vez que a densidade é normalmente definida em termos de proximidade entre os objetos
- Uma abordagem comum é definir a densidade como a média da distância dos k vizinhos mais próximos
- Se a distância é pequena entre os vizinhos mais próximos, a densidade é alta
- As mesmas considerações para o método baseado no k-NN apresentadas anteriormente devem ser consideradas aqui

Inverso da Densidade

$$density(\mathbf{x}, k) = \left(\frac{\mathbf{y} \in \mathcal{N}(\mathbf{x}, k) distance(\mathbf{x}, \mathbf{y})}{|\mathcal{N}(\mathbf{x}, k)|} \right)^{-1}$$

na qual $N(\mathbf{x}, k)$ é o conjunto contendo os k vizinhos mais próximos de $\mathbf{x}$, $|N(\mathbf{x}, k)|$ é o tamanho do conjunto, e $\mathbf{y}$ é o vizinho mais próximo

Densidade Relativa

- Outra abordagem é considerar também a densidade dos pontos vizinhos para definir a densidade da região

- Com isso, pode-se verificar se o ponto está em uma região de densidade semelhante à densidade da região de seus vizinhos

Densidade Relativa

$$average\_relative\_density(\mathbf{x}, k) = \frac{density(\mathbf{x}, k)}{\sum_{\mathbf{y} \in \mathcal{N}(\mathbf{x},k)} density(y, k)/|\mathcal{N}(\mathbf{x}, k)|}$$

A complexidade de tempo dessa abordagem também é $n^2 m$

Avaliação

- Assim como em outras situações, é interessante avaliar o **poder preditivo** do algoritmo de detecção de anomalias
- O poder preditivo corresponde a **mensurar o quanto o algoritmo acerta o que é uma anomalia ou o que não é uma anomalia**
- Além disso, se conseguirmos medir o poder preditivo dos algoritmos, podemos compará-los e escolher o algoritmo mais adequado em uma determinada situação

Avaliação

- O mais adequado na avaliação de outlier é o uso de **medidas externas** [Aggarwal, 2015]
- Medidas externas são aquelas que **comparam o resultado da predição com um conjunto verdade**
- Portanto, para se fazer uso dessas duas medidas há duas possibilidades:
  1. Gerar exemplos anômalos artificialmente
  2. Considerar um conjunto de exemplos que possuam exemplos normais e exemplos anômalos
- **OBSERVAÇÃO:** não se deve utilizar na avaliação exemplos que foram utilizados no treinamento do modelo

Avaliação

- Vamos considerar nesse cenário que exemplos positivos são os outliers e negativos são os exemplos normais
- Com isso, podemos utilizar medidas derivadas de uma matriz de confusão assim como na classificação

		Real	
		Positivo	**Negativo**
Predito	**Positivo**	*VP*	*FP*
	Negativo	*FN*	*VN*

- **VP (Verdadeiros Positivos)**: número de exemplos positivos e que foram corretamente classificados como positivos
- **FP (Falsos Positivos)**: número de exemplos negativos mas que foram erroneamente classificados como positivos
- **FN (Falsos Negativos)**: número de exemplos positivos mas que foram erroneamente classificados como negativos
- **VN (Verdadeiros Negativos)**: número de exemplos negativos e que foram corretamente classificados como negativos

Avaliação

- **1º conjunto de medidas**:

$$Precisão = \frac{TP}{TP + FP}$$

$$Revocação = \frac{TP}{TP + FP}$$

$$F_1 = \frac{2 * \text{Precisão} * \text{Revocação}}{\text{Precisão} + \text{Revocação}}$$

Avaliação

- **2º conjunto de medidas**:
  - $\mathcal{S}(t)$ conjunto de outliers obtidos por um threshold t
  - $\mathcal{G}$ conjunto verdade dos exemplos positivos, i.e., quais são os outliers
  - $\mathcal{D}$ conjunto verdade total
  - **True Positive Rate** (TPR) - porcentagem dos outliers em G que foram preditos como *outliers* dado um limiar t:

$$TRP(t) = 100 * \frac{|\mathcal{S}(t) \bigcap \mathcal{G}|}{|\mathcal{G}|}$$

  - **False Positive Rate** (FPR) - porcentagem de falsos positivos em relação ao conjunto verdade dos negativos

$$FPR(t) = 100 * \frac{|\mathcal{S}(t) - \mathcal{G}|}{|\mathcal{D} - \mathcal{G}|}$$

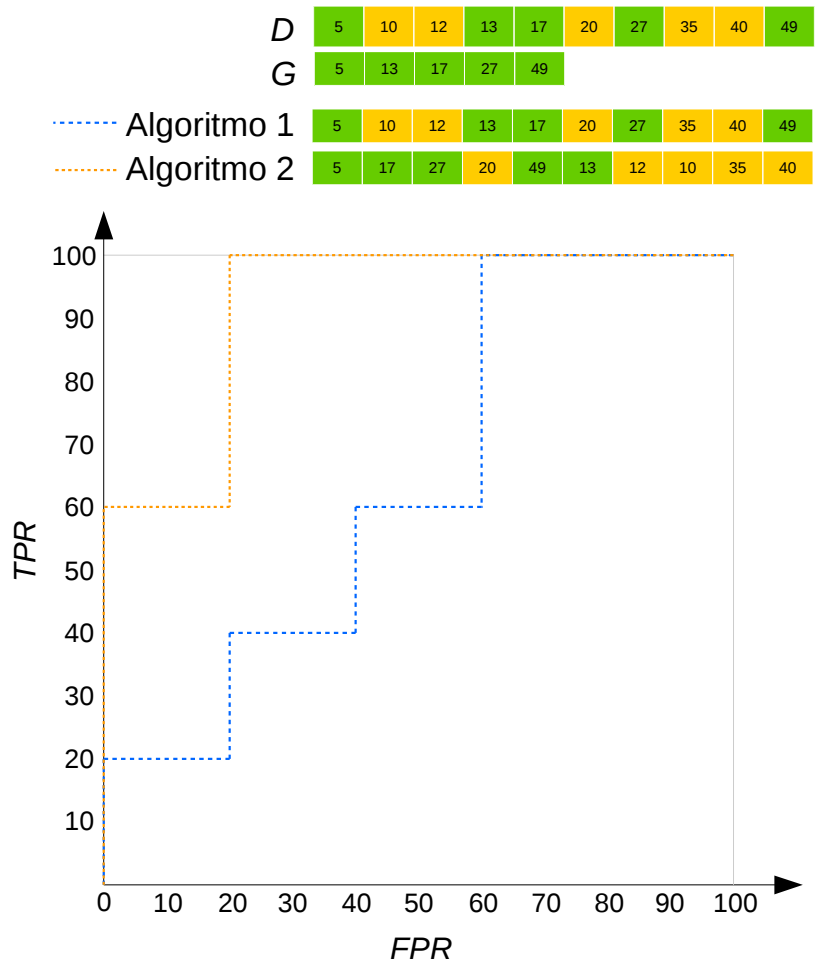
Avaliação

- **2º conjunto de medidas**:
  - **Receiver Operating Characteristic (ROC)**
    - Um limiar muito restritivo irá "perder" muitos "outliers"
    - Um limiar pouco restritivo irá considerar muitos exemplos negativos como outliers
    - Isso gera um trade-off entre Falsos Positivos e Falsos Negativo
    - Pode haver um problema em definir o limiar "correto"

Avaliação

- **2º conjunto de medidas**:
  - **Receiver Operating Characteristic (ROC)**
    - Porém, podemos comparar os algoritmos utilizando um *ranking* de acordo com um *score* atribuído aos exemplos
    - A partir disso, pode-se comparar a *TPR* e a *FPR* considerando os r primeiros exemplos do *ranking*
    - Ao variar o valor de r, desde 1 até o número de exemplos a serem testados, têm-se a curva *ROC*, sendo que o valor de *FPR* é disposto no eixo x e o valor do *TPR* é disposto no eixo y

Avaliação


D
5 10 12 13 17 20 27 35 40 49
G
5 13 17 27 49
Algoritmo 1
5 10 12 13 17 20 27 35 40 49
Algoritmo 2
5 17 27 20 49 13 12 10 35 40
TPR
100 90 80 70 60 50 40 30 20 10
0 10 20 30 40 50 60 70 80 90 100
FPR

Material Complementar

- Normal Distribution Applet

  `http://homepage.divms.uiowa.edu/~mbognar/applets/normal.html`

- Gaussian (Normal) Distribution

  `https://academo.org/demos/gaussian-distribution/`

- Keystroke dynamics

  `https://en.wikipedia.org/wiki/Keystroke_dynamics`

- Matriz de Covariância

  `https://pt.wikipedia.org/wiki/Matriz_de_covari%C3%A2ncia`

Material Complementar

- Distância de Mahalanobis

  https://pt.wikipedia.org/wiki/Dist%C3%A2ncia_de_Mahalanobis

- *F-score*

  https://en.wikipedia.org/wiki/F-score

- Unsupervised Machine Learning Approaches for Outlier Detection in Time Series, using Python

  https://towardsdatascience.com/

  unsupervised-machine-learning-approaches-for-outlier-detection-in-time-series-using-python-5759c6

- Outlier Detection — Theory, Visualizations, and Code

  https:

  //towardsdatascience.com/outlier-detection-theory-visualizations-and-code-a4fd39de540c

Material Complementar

- Deep Learning for Anomaly Detection: A Survey

  `https://arxiv.org/abs/1901.03407`

Imagem do Dia

Inteligência Artificial
http://lives.ufms.br/moodle/

Rafael Geraldeli Rossi
rafael.g.rossi@ufms.br

Slides baseados em [Tan et al., 2013, Aggarwal, 2015]

Referências Bibliográficas I


Aggarwal, C. (2015).
Data Mining: The Textbook.
Springer International Publishing.


Tan, P., Steinbach, M., and Kumar, V. (2013).
Introduction to Data Mining: Pearson New International Edition.
Pearson Education Limited.