

## Outliers

- A existência de observações discordantes com as restantes é de relativamente fácil determinação em amostras univariadas.
- Por observação dos valores que constituem a amostra ou pela análise de alguns gráficos, é fácil identificar as observações que se afastam da maioria

1

## Outliers

- Em dados multidimensionais, uma observação é considerada outlier se está "muito" distante das restantes no espaço p-dimensional definido pelas variáveis.

2

- Na utilização de testes formais de *outliers* dividem-se em duas classes:
  - aqueles em que as observações discordantes da amostra são identificadas como sendo outliers, e;
  - aqueles que testam a presença de *outliers* mas não identificam observações particulares como *outliers*.

3

- aqueles que testam a presença de *outliers* mas não identificam observações particulares como *outliers*. –
  - estatística de excesso de dispersão (propagação),
  - estatísticas de amplitude/dispersão,
  - estatística de desvio/dispersão,
  - estatísticas de "soma de quadrados",
  - estatísticas dos momentos de ordem superior e
  - estatísticas de localização/extremos

4

## Outliers - 3 fases

- Seleção de outlier
- Verificação dos outlier
- O que fazer com as observações discordantes ?

5

## Técnicas de Mineração de Exceções

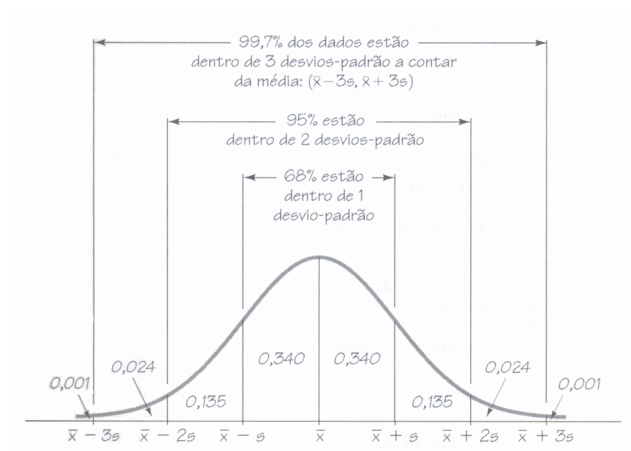
- **Classes de técnicas:**
  - **Semi-automático:**
    - **Visualização**
  - **Automático**
    - **Baseados em Clustering**
    - **Baseado em Estatística**
    - **Baseado em Desvio**
    - **Baseado em Distância**
- **Características desejáveis**
  - Escalável para alta dimensionalidade
  - Interpretabilidade dos resultados
  - Computacionalmente eficiente
  - Dá importância ao comportamento local dos dados
  - Ordenação dos outliers

6

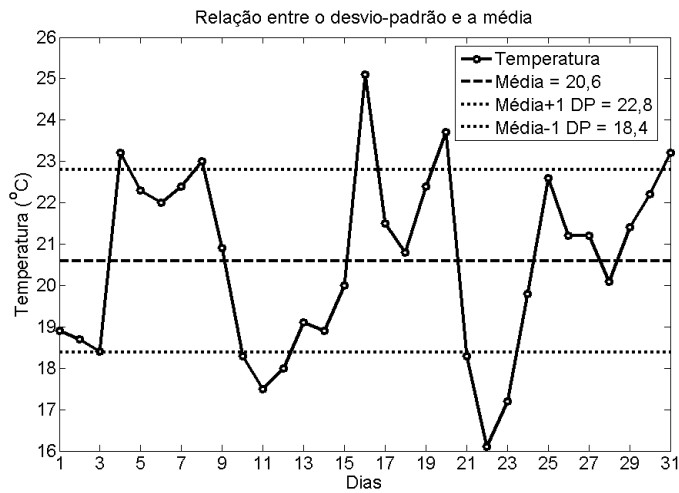
## Outliers

- Métodos baseados em Estatística
- Métodos baseados em Distância
- Métodos baseados em Agrupamento
- Métodos Baseados em Desvio

## Outliers

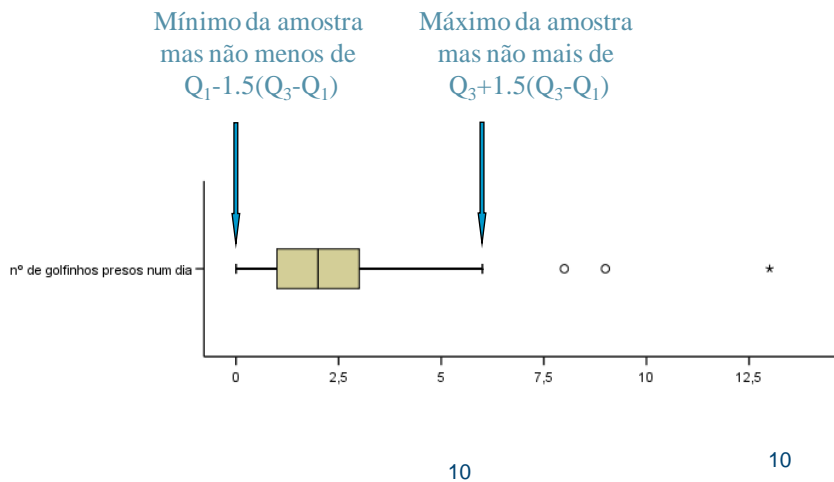


## Outliers



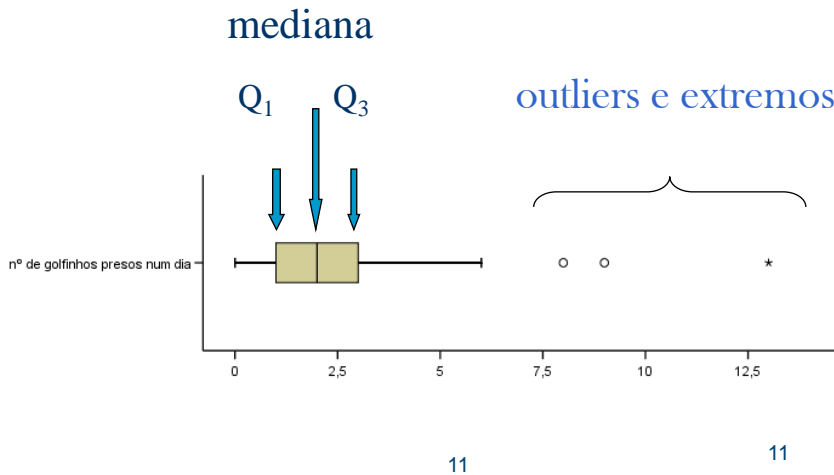
9

## Boxplot



## Boxplot

- Pode ser encarada como a representação gráfica de algumas medidas de localização:



## Mineração de Outliers Baseada em Estatística

- Assume distribuição ou modelo probabilístico para um conjunto de dados
  - Ex: distribuição normal
- Usa Teste de discordância (TD) → identifica os outliers com respeito ao modelo escolhido
  - Se um objeto for significativamente maior ou menor que o modelo escolhido ele é uma exceção
- O TD examina 2 hipóteses:
  - Uma hipótese de trabalho (hipótese nula)
  - Uma hipótese alternativa

## modelo de discordância

- É considerada a hipótese nula, segundo a qual a amostra foi retirada de uma população com distribuição específica que pode ou não ser conhecida e ser especificada completamente ou não, e onde não existem observações "anormais".
- Em oposição, a hipótese alternativa considera que todas as observações ou apenas as "anormais" têm uma distribuição diferente da hipótese nula.
- A hipótese nula será rejeitada em favor da hipótese alternativa se existirem observações aberrantes.

13

## Mineração de Outliers Baseada em Estatística

- |                                                                                                                                                                                                                                                   |                                                                                                                                                                                                                                                                                                       |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| <ul style="list-style-type: none"><li>■ <b>Vantagens:</b><ul style="list-style-type: none"><li>– Pode ser avaliado o nível de significância de uma exceção</li><li>– Usa métodos estatístico consolidados ao longo dos tempos</li></ul></li></ul> | <ul style="list-style-type: none"><li>■ <b>Limitações:</b><ul style="list-style-type: none"><li>– O modelo escolhido influencia a identificação dos Outliers</li><li>– Testa aberração ao longo de apenas uma única dimensão</li><li>– Dificuldade na escolha de uma distribuição</li></ul></li></ul> |
|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|

## Outliers

- Métodos baseados em Estatística
- **Métodos baseados em Distância**
- Métodos baseados em Agrupamento
- Métodos Baseados em Desvio

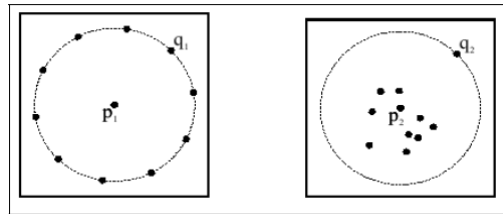
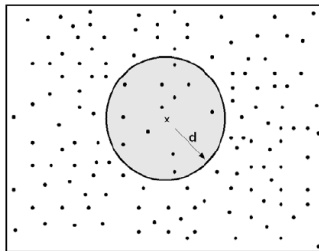
### Mineração de Outliers Baseada em Distância:

- Busca Resolver limitações do estatístico
- Um outlier é determinado baseado na distancia  $D^k(p)$
- $D^k(p)$ = distância de p ao seu k-esimo vizinho
- Evita suposição sobre distribuição dos dados
- Menor custo computacional
- Pode, às vezes, convergir para os métodos estatísticos
- Desvantagem
  - Não é escalável para mais que 5 dimensões



## Detecção de Outliers Baseada em Distâncias: $D^k(p)$

- Para cada ponto  $p$  no conjunto de dados calcula  $D^k(p)$
- Para calcular cada  $D^k(p)$  percorre todos os dados
- Mantém uma lista de  $k$  vizinhos mais próximo para cada ponto  $p$
- Os  $n$  pontos com maior valor de  $D^k(p)$  são os  $n$  outliers
- Para melhorar a eficiência pode-se considerar blocos de pontos ao invés de pontos individuais



### Exemplo: IRIS

Weka 3.5.8 - Explorer

Program Applications Tools Visualization Windows Help

Explorer

Preprocess | Classify | Cluster | Associate | Select attributes | Visualize

Open File... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: MultiFilter - F weka.filters.AllFilter Apply

Current relation: Relation: iris Instances: 150 Attributes: 5

Attributes:

No.	Attribute	Type
1	sepal.length	Numeric
2	sepal.width	Numeric
3	petal.length	Numeric
4	petal.width	Numeric
5	class	Nominal

Selected attribute: Name: sepal.length Type: Numeric Missing: 0 (0%) Distinct: 35 Unique: 9 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.843
StdDev	0.828

Class: class (Nom) Visualize All

Status: OK

Log x0

Windows taskbar: Start, Outlook, Tanagra - Data Mini..., aula pratica BI-OM, Untitled - Notepad, Parte\_3.ppt (Combo...), Weka 3.5.8 - Expl..., EN, 17:28 Mathieu

## Exemplo: IRIS

**Weka 3.5.8 - Explorer**  
Program Applications Tools Visualization Windows Help

Preprocess Classify Cluster Associate Select attributes Visualize

Choose **348 - C 0.25 - M2**

Test options  
 Use training set  
 Supplied test set  
 Cross-validation  
 Percentage split % **66**  
 More options...

(Nom) class  
 Start Stop

Result list (right-click for options)  
 17:27:18 - trees\_348

Classifier output

Size of the tree : 9

Time taken to build model: 0 seconds

=== Evaluation on test split ===  
 === Summary ===

Correctly Classified Instances	49	96.0784 %
Incorrectly Classified Instances	2	3.9216 %
Kappa statistic	0.9408	
Mean absolute error	0.0396	
Root mean squared error	0.1579	
Relative absolute error	8.8979 %	
Root relative squared error	33.4091 %	
Total Number of Instances	51	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Iris-setosa
1	0.063	0.905	1	0.95	0.969	Iris-versicolor
0.882	0	1	0.882	0.938	0.967	Iris-virginica

=== Confusion Matrix ===

```

a b c <- classified as
15 0 0 | a = Iris-setosa
0 19 0 | b = Iris-versicolor
0 2 15 | c = Iris-virginica
    
```

Status: OK

19

## Exemplo: IRIS

**Weka 3.5.8 - Explorer**  
Program Applications Tools Visualization Windows Help

Preprocess Classify Cluster Associate Select attributes Visualize

Open file... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **RemoveMisclassified -W "weka.classifiers.lazy.Ibk -K 5 -W 0 -I -A "weka.core.neighboursearch.LinearNNSearch -A "weka.core.EuclideanDistance -R first-last" -C -I -F 10 -T 0.01 -I 0** Apply

Current relation  
 Relation: iris-weka.filters.unsupervised.instance.RemoveMisclassified-  
 Instances: 143

Attributes

No.	Name
1	sepalwidth
2	sepalwidth
3	petalwidth
4	petalwidth
5	class

weka.gui.GenericObjectEditor  
 weka.filters.unsupervised.instance.RemoveMisclassified  
 About: A filter that removes instances which are incorrectly classified.

weka.gui.GenericObjectEditor  
 weka.classifiers.lazy.Ibk  
 About: K-nearest neighbours classifier.

KNN: 5  
 crossValidate: False  
 debug: False  
 distanceWeighting: Weight by 1/distance  
 meanSquared: False  
 nearestNeighbourSearchAlgorithm: Choose **LinearNNSearch -A "weka.core.Euc**  
 windowSize: 0

Status: OK

20

## Exemplo: IRIS

**Weka 3.5.8 - Explorer**

Program Applications Tools Visualization Windows Help

Preprocess Classify Cluster Associate Select attributes Visualize

Open File... Open URL... Open DB... Generate... Undo Edit... Save...

Filter: Choose **RemoveMisclassified** -W "\weka.classifiers.lazy.IBK -K 5 -W 0 -1 -A "\weka.core.neighboursearch.LinearNNSearch -A {\{\weka.core.EuclideanDistance -R first-last\}\} -C -1 -F 10 -T 0.01 -1.0" Apply

Current relation: Relation: iris-weka.filters.unsupervised.instance.RemoveMisclassified-weka.classifiers.lazy.IBK -K 5 -W 0 -1 -A "\weka.c... Instances: 143 Attributes: 5

Selected attribute: Name: sepalength Missing: 0 (0%) Distinct: 35 Type: Numeric Unique: 3 (6%)

Statistic	Value
Minimum	4.3
Maximum	7.9
Mean	5.836
StdDev	0.84

Attributes: All None Invert Pattern

No.	Name
1	sepalength
2	sepalwidth
3	petalength
4	petalwidth
5	class

Class: class (Nom) Visualize All

Remove

Status: OK Log

21

## Exemplo: IRIS

**Weka 3.5.8 - Explorer**

Program Applications Tools Visualization Windows Help

Preprocess Classify Cluster Associate Select attributes Visualize

Classifier: Choose **J48** -C 0.25 -M 2

Test options: Use training set, Supplied test set, Cross-validation, Percentage split: % 66

(Nom) class: Start Stop

Result list (right-click for options): 17:24:55 - trees.J48

Classifier output:

Time taken to build model: 0 seconds

=== Evaluation on test split ===

=== Summary ===

Correctly Classified Instances	49	100 %
Incorrectly Classified Instances	0	0 %
Kappa statistic	1	
Mean absolute error	0.0068	
Root mean squared error	0.0146	
Relative absolute error	1.5295 %	
Root relative squared error	3.0905 %	
Total Number of Instances	49	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
1	0	1	1	1	1	Iris-setosa
1	0	1	1	1	1	Iris-versicolor
1	0	1	1	1	1	Iris-virginica

=== Confusion Matrix ===

```

a b c -- classified as
19 0 0 | a = Iris-setosa
0 16 0 | b = Iris-versicolor
0 0 14 | c = Iris-virginica
    
```

Status: OK Log

22

## Detecção de Outliers Baseada em Distâncias: $D^k(p)$

### Algoritmo Baseado em partições

- Detecta os outliers mais fortes
  - Os outliers são ordenados pela distância  $D^k(p)$
- Baseia se na distância dos vizinhos mais próximos
- O conjunto de dados é dividido em partições por meio de algoritmos de agrupamento
- Poda partições que não são candidatas a conter outlier
  - Acelera a identificação pois diminui a quantidade de pontos

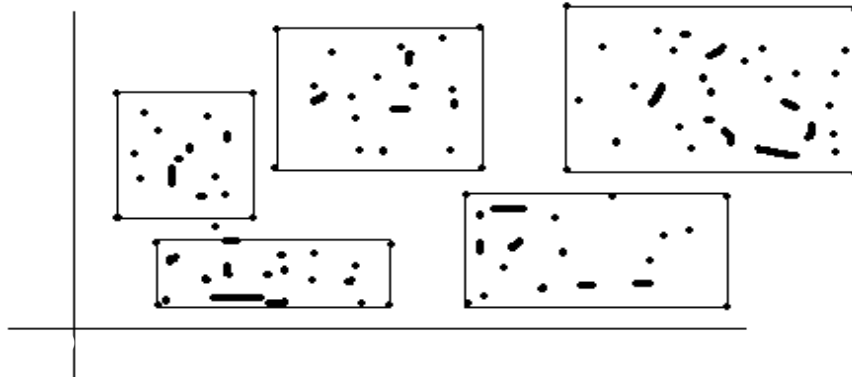
## Detecção de Outliers Baseada em Distâncias : $D^k(p)$

### Algoritmo Baseado em partições (passos)

- Gerar partições
  - Através de clustering
- Calcular limites  $D^k$  para os pontos em cada partição
  - $P.upper = \max(D^k)$  e  $P.lower = \min(D^k)$  dos pontos da partição  $P$
- Identificar partições candidatas a conter exceções
  - $P.upper \geq \min D^k Dist = \min\{P_i.lower : 1 \leq i \leq l\}$
  - $P_i.lower > P_j.lower \dots > P_1.lower$  e o número de pontos seja pelo menos  $n$
- Computar exceções com os pontos nas partições candidatas
  - $P.neighbors$  denota as partições vizinhas de  $P$  a uma distância de  $P.upper$

## Detecção de Outliers Baseada em Distâncias : $D^k(p)$

Algoritmo Baseado em partições (passos)

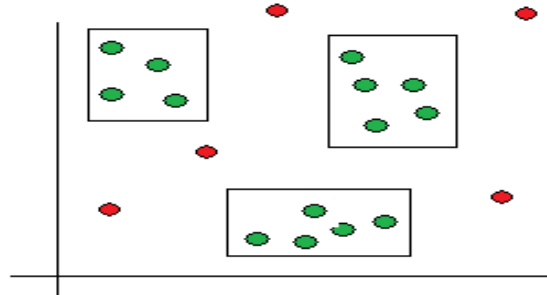


## Outliers

- Métodos Baseados em Estatística
- Métodos baseados em Distância
- **Métodos Baseados em Agrupamento**
- Métodos Baseados em Desvio

## Outliers - baseada em clustering

- Dados que não se ajustam a nenhum grupo são considerados exceções



## Mineração de exceção baseada em agrupamento

- Vantagens
  - Reutiliza vasto leque de métodos de agrupamentos
  - Não requer conhecimento prévio de distribuição
- Limitações
  - O que se busca é otimizar os agrupamentos, não a detecção de exceções
  - O que é exceção para uma configuração pode não ser para outra

## Outliers

- Métodos Baseados em Estatística
- Métodos baseados em Distância
- Métodos Baseados em Agrupamento
- **Métodos Baseados em Desvio**
- Métodos Baseados em Densidade

## Mineração de Outliers Baseada em Desvio

- **Não usa métodos estatísticos nem medidas de distância**
- **Define exceção como pontos cujo valor desviam da maioria de uma, algumas ou todas as dimensões**
- **Exceções são equivalentes a Desvios de comportamento**

## Outliers

- Métodos Baseados em Estatística
- Métodos baseados em Distância
- Métodos Baseados em Agrupamento
- Métodos Baseados em Desvio
- Métodos Baseados em Densidade

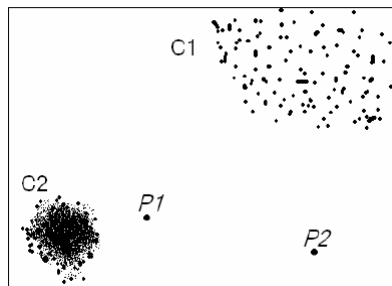
## Mineração de Outliers Baseada em Densidade de Distribuição

- Baseada no *Local Outlier Factor (LOF)* que é a média das densidade do exemplo  $p$  e a densidade dos seus vizinhos mais próximos.
- LOF depende da densidade local da sua vizinhança.
- A vizinhança é definida pela distância em relação aos  $\text{MinPts}$ -th que são os vizinhos mais próximos, onde  $\text{MinPts}$  é o número mínimo de pontos considerados como vizinhos mais próximos.
- Os passos do processo são:
  - Computam a densidade da vizinhança local de cada ponto.
  - Computam LOF.
  - Escolhem exemplos  $p$  com maiores LOF como outliers.



## Mineração de Outliers Baseada em Densidade de Distribuição

- Na abordagem K-vizinhos  $p_2$  não é considerado como *outlier*, enquanto a para a abordagem LOF,  $p_1$  e  $p_2$  são *outliers*.



- Em uso típico, pontos com altos LOF são considerados como outliers.

33

## Referências

- *Data Mining: concepts and techniques*, de **Han, J. & Kamber, M.**, Morgan Kaufmann, 2001
- Efficient Algorithms for Mining Outliers from Data sets. **Sridhar Ramaswamy, Rajeev Ratogi e Kyuseok Shim.** 2000
- Outlier Detection for High Dimensional Data. **Charu C. Aggarwal e Philip S. Yu.** 2001
- Identification of Outliers, **D. Hawkins**, Chapman and Hall, London, 1980.