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

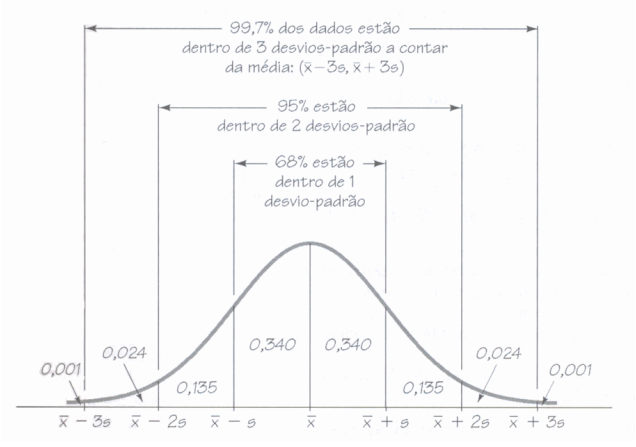
- **Classes de técnicas:**
  - **Semi-automático:**
    - **Visualização**
  - **Automático**
    - **Baseados em Clustering**
    - **Baseado em Estatística**
    - **Baseado em Desvio**
    - **Baseado em Distância**
- **Características desejáveis**
  - Escalável para alta dimensionalidade
  - Interpretabilidade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

6

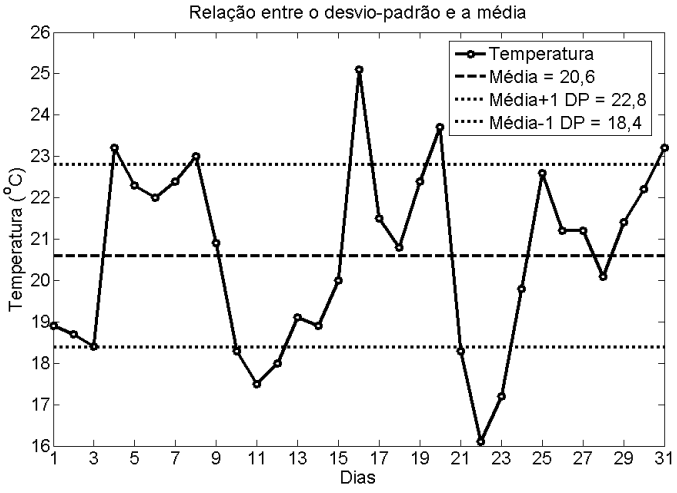
Outliers

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

Outliers



Outliers

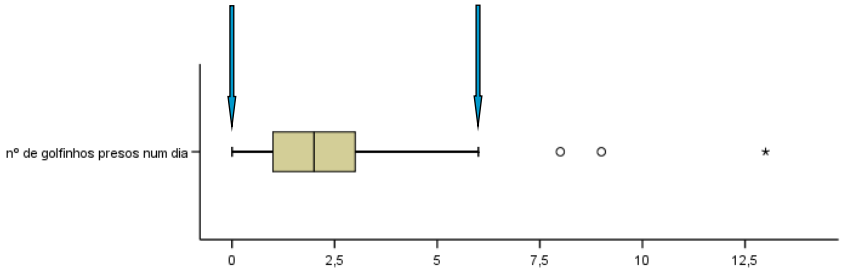


9

Boxplot

Mínimo da amostra mas não menos de Q_1-1.5(Q_3-Q_1)

Máximo da amostra mas não mais de Q_3+1.5(Q_3-Q_1)

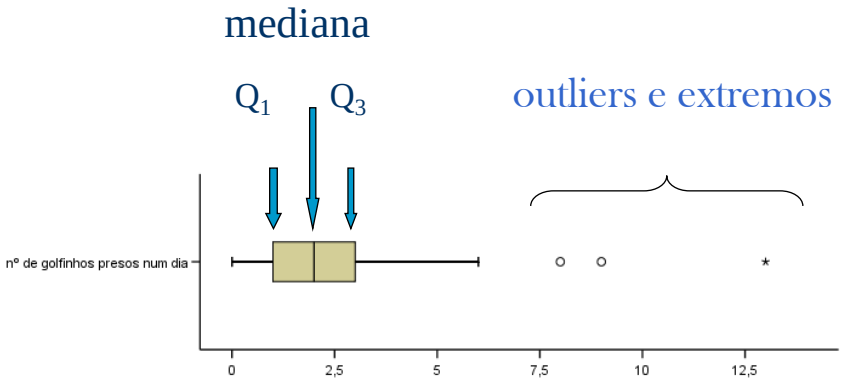


10

10

Boxplot

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



11

11

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

- **Vantagens:**
  - **Pode ser avaliado o nível de significância de uma exceção**
  - **Usa métodos estatístico consolidados ao longo dos tempos**
- **Limitações:**
  - **O modelo escolhido influencia a identificação dos Outliers**
  - **Testa aberração ao longo de apenas uma única dimensão**
  - **Dificuldade na escolha de uma distribuição**

Outliers

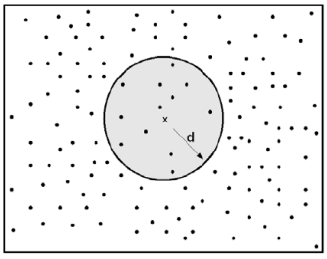
- Métodos baseados em Estatística
- Métodos baseados em Distância
- Métodos baseados em Agrupamento
- Métodos Baseados em Desvio

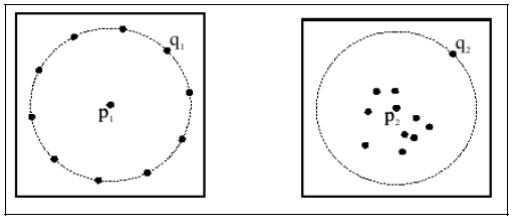
Mineração de Outliers Baseada em Distância:

- 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, ás vezes, convergir para os métodos estatísticos
- Desvantagem
  - Não é escalável para mais que 5 dimensões

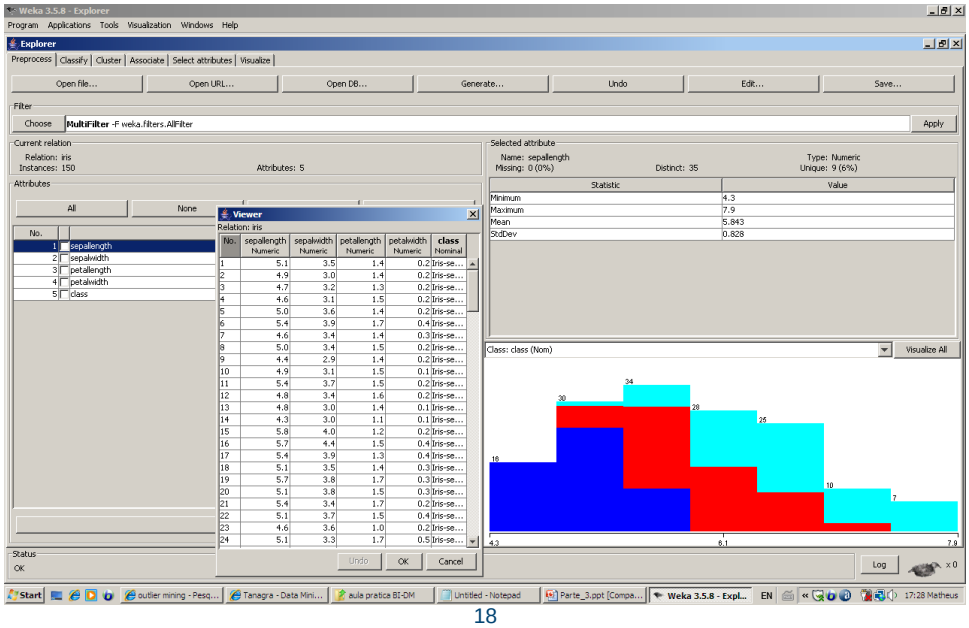
Detecção de Outliers Baseada em Distâncias: $D^k(p)$

- 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

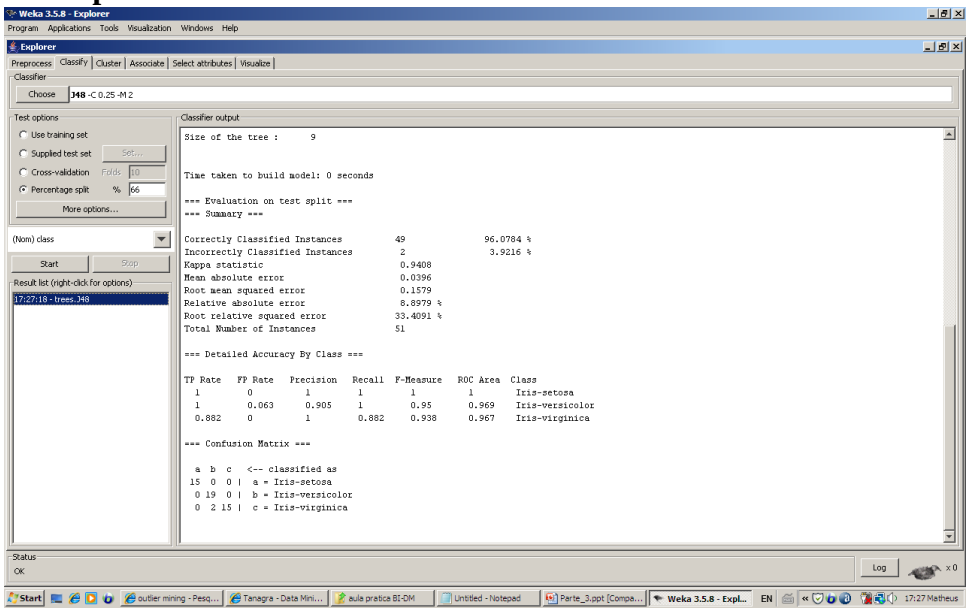




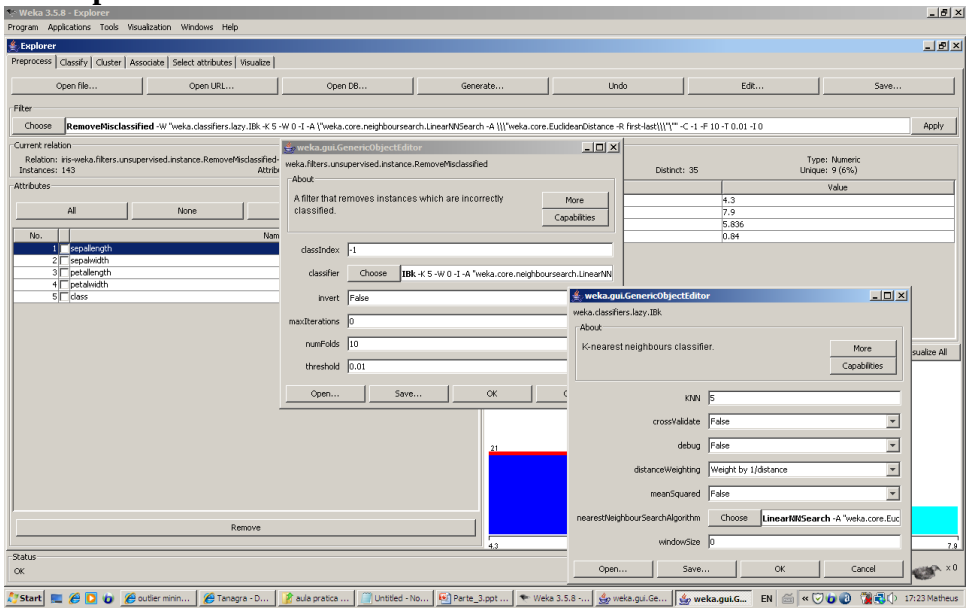
Exemplo: IRIS



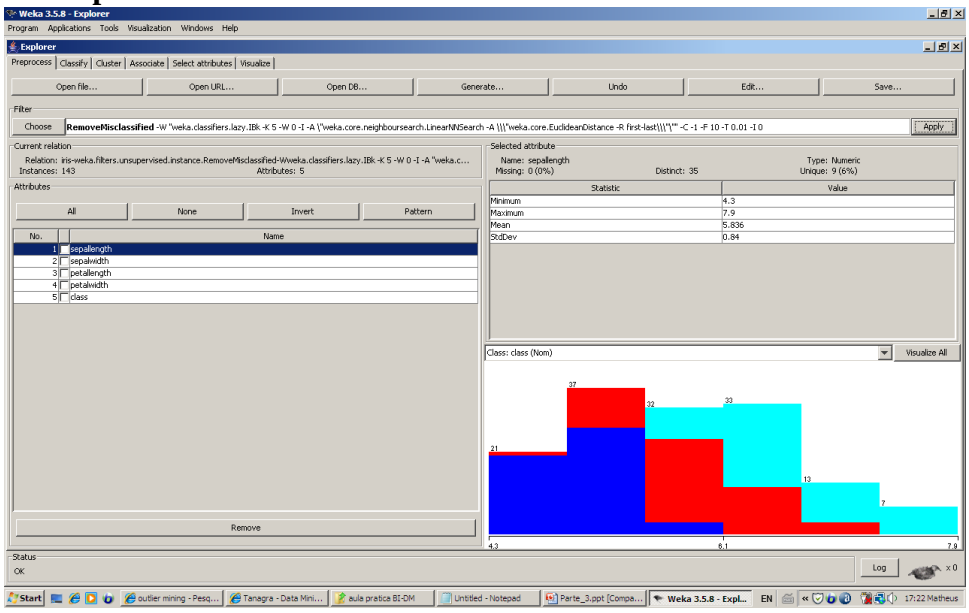
Exemplo: IRIS



Exemplo: IRIS

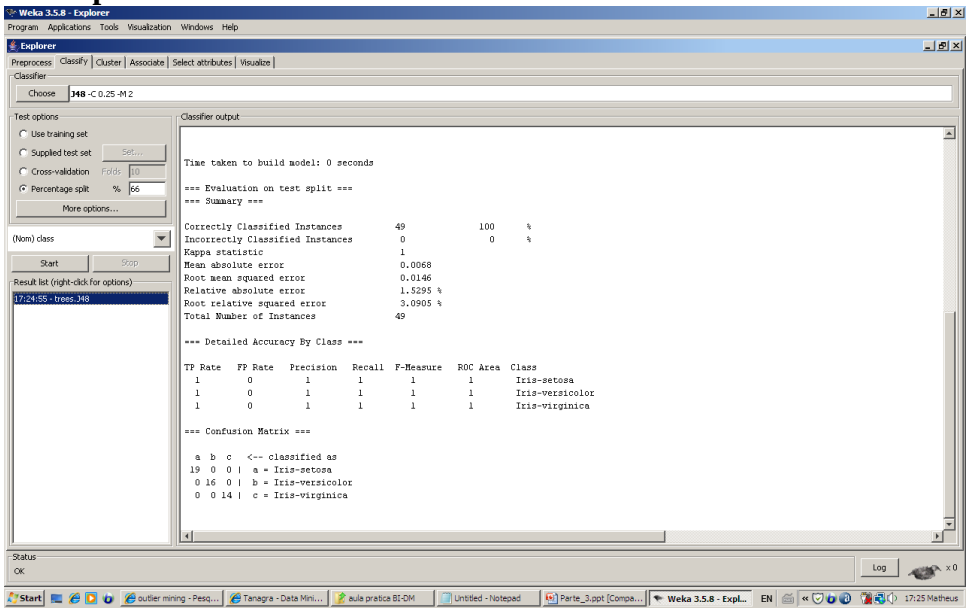


Exemplo: IRIS



21

Exemplo: IRIS



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 divididos 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

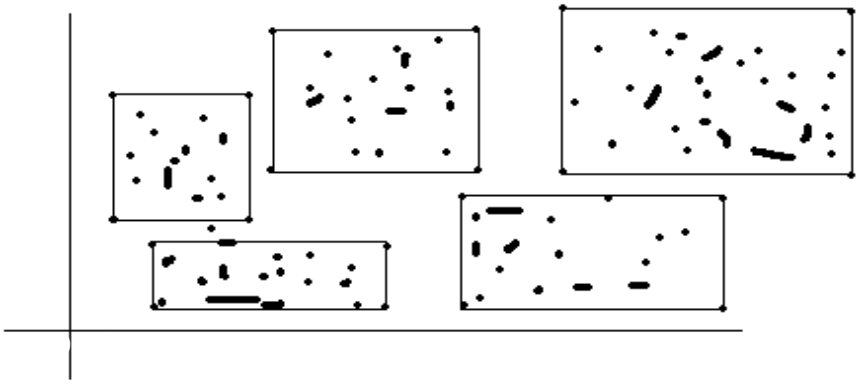
Detecção de Outliers Baseada em Distâncias : $D^k(p)$

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≥minDkDist=min{P_i.lower:1≤ $i \leq l$}
  - P_i.lower>P_j.lower>..>P_l.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)

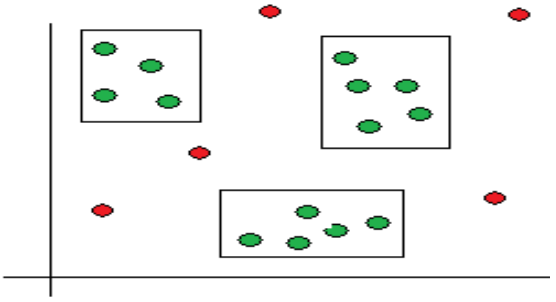


Outliers

- 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

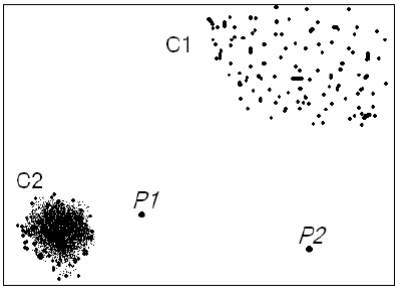
- Métodos Baseados em Estatística
- Métodos baseados em Distância
- Métodos Baseados em Agrupamento
- Métodos Baseados em Desvio
- Métodos Baseados em Densidade

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 vizinhaça.
- A vizinhaça é definida pela distância em relação aos MinPts-th que são os vizinhos mais próximos, onde MinPts é o número mínimo de pontos considerados como vizinhos mais próximos.
- Os passos do processo são:
  - Computam a densidade da vizinhaça 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 p2 não é considerado como *outlier*, enquanto a para a abordagem LOF, p1 e p2 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 Algoritms 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.