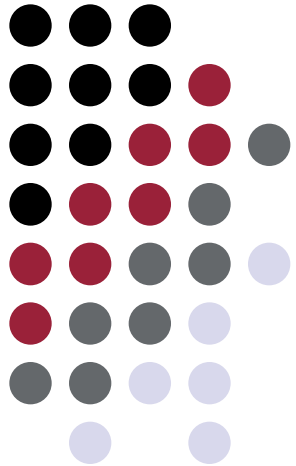
# Mineração de Dados: Detecção de Anomalias


## CEA462 – Sistemas de Apoio à Decisão

Prof. MSc. George H. G. Fonseca
Universidade Federal de Ouro Preto



# Introdução

- Detecção de Anomalias é a tarefa de identificar registros que possuem características demasiadamente diferentes dos demais

- Exemplos de aplicações
  - Detectar mensagens de spam
  - Detectar transações fraudulentas em cartões de crédito

# Introdução

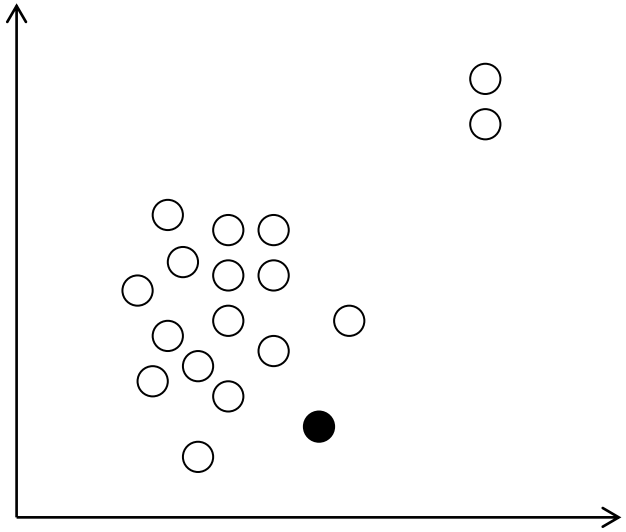
- Podem ser empregadas várias abordagens
  - Abordagem estatística
  - Baseada em densidade
  - **Baseada em Vizinhança**

# Detecção de Anomalias com o K-NN

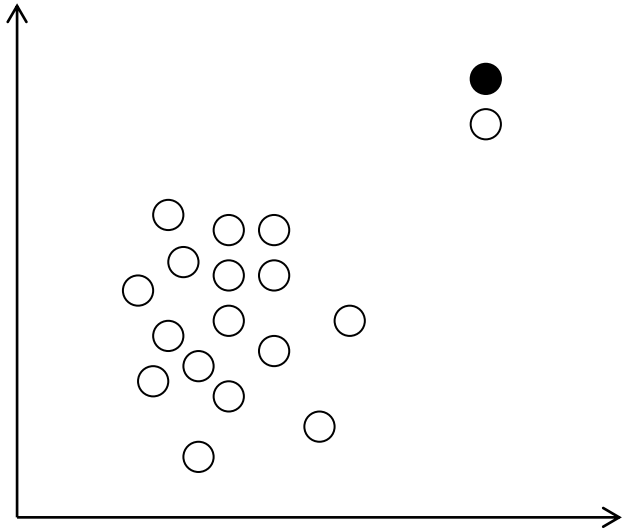
- Um elemento é considerado anômalo se for o mais distante dos K vizinhos mais próximos
  - K é um parâmetro
  - A distância pode ser medida de vários modos, sendo a Distância Euclidiana o mais aplicado
  - A cada elemento é atribuído um grau de estranheza
    - Média das distâncias euclidianas dos K vizinhos mais próximos
  - Considera-se estranhos os n indivíduos com o maior grau de estranheza

# Detecção de Anomalias com o K-NN

- O valor K tem um papel importante no resultado do algoritmo!!



K = 1

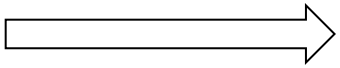


K = 2

# Detecção de Anomalias com o K-NN

- Exemplo
  - Encontrar um elemento anômalo dentre as transações de uso de cartão de crédito de um cliente nos últimos 10 dias
  - Considere K = 3

| Dia | Valor | # Itens |
|---|---|---|
| 1 | 200 | 2 |
| 2 | 70 | 1 |
| 3 | 50 | 2 |
| 4 | 130 | 3 |
| 5 | 890 | 12 |
| 6 | 300 | 4 |
| 7 | 0 | 0 |
| 8 | 25 | 1 |
| 9 | 100 | 3 |
| 10 | 0 | 0 |



| Dia | Valor | # Itens |
|---|---|---|
| 1 | 0,22 | 0,17 |
| 2 | 0,08 | 0,08 |
| 3 | 0,06 | 0,17 |
| 4 | 0,15 | 0,25 |
| 5 | 1,00 | 1,00 |
| 6 | 0,34 | 0,33 |
| 7 | 0,00 | 0,00 |
| 8 | 0,03 | 0,08 |
| 9 | 0,11 | 0,25 |
| 10 | 0,00 | 0,00 |

# Detecção de Anomalias com o K-NN

- Exemplo

| d(x, y) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Grau de Estranheza |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0,00 | **0,17** | 0,17 | **0,11** | 1,14 | 0,20 | 0,28 | 0,21 | **0,14** | 0,28 | 0,14 |
| **2** | 0,17 | 0,00 | **0,09** | 0,18 | 1,30 | 0,36 | 0,11 | **0,05** | 0,17 | **0,11** | 0,08 |
| **3** | 0,17 | **0,09** | 0,00 | 0,12 | 1,26 | 0,33 | 0,18 | **0,09** | **0,10** | 0,18 | 0,09 |
| **4** | **0,11** | 0,18 | **0,12** | 0,00 | 1,14 | 0,21 | 0,29 | 0,20 | **0,03** | 0,29 | 0,09 |
| **5** | **1,14** | 1,30 | 1,26 | **1,14** | 0,00 | **0,94** | 1,41 | 1,34 | 1,16 | 1,41 | **1,07** |
| **6** | **0,20** | 0,36 | 0,33 | **0,21** | 0,94 | 0,00 | 0,47 | 0,40 | **0,24** | 0,47 | 0,22 |
| **7** | 0,28 | **0,11** | 0,18 | 0,29 | 1,41 | 0,47 | 0,00 | **0,09** | 0,27 | **0,00** | 0,07 |
| **8** | 0,21 | **0,05** | **0,09** | 0,20 | 1,34 | 0,40 | **0,09** | 0,00 | 0,19 | 0,09 | 0,08 |
| **9** | **0,14** | 0,17 | **0,10** | **0,03** | 1,16 | 0,24 | 0,27 | 0,19 | 0,00 | 0,27 | 0,09 |
| **10** | 0,28 | **0,11** | 0,18 | 0,29 | 1,41 | 0,47 | **0,00** | **0,09** | 0,27 | 0,00 | 0,07 |

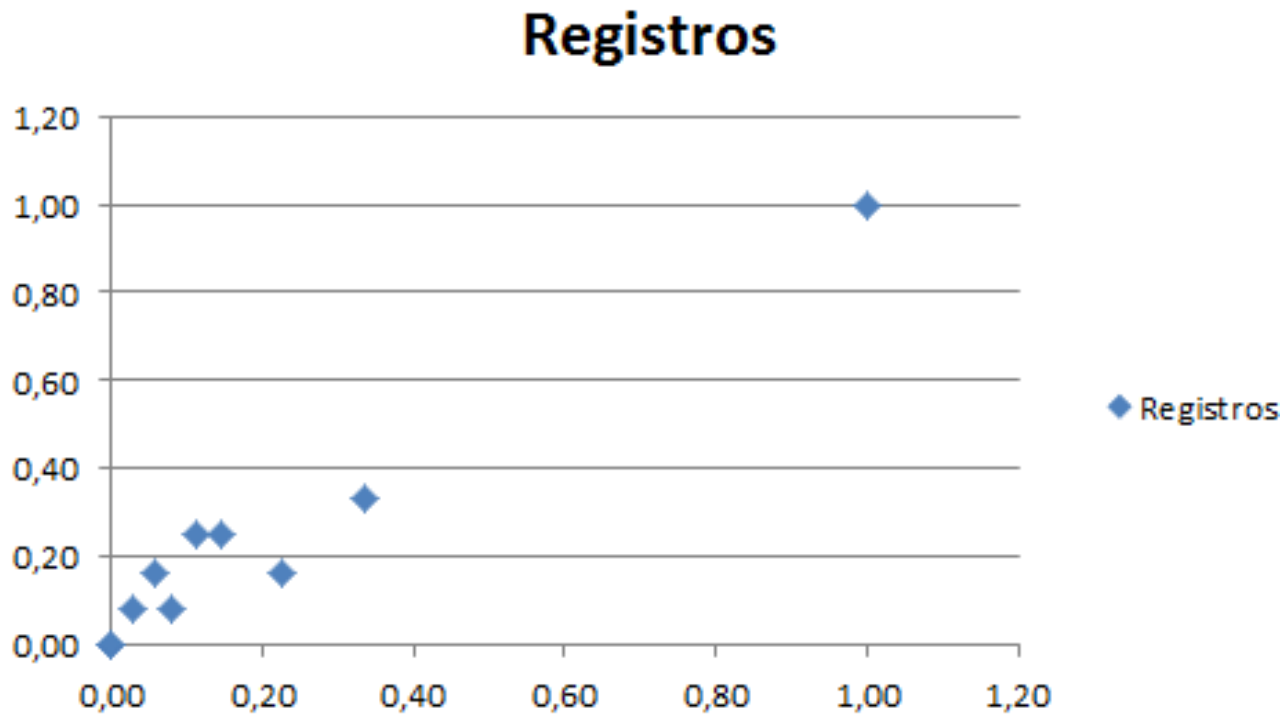
# Detecção de Anomalias com o K-NN

- Exemplo
  - Pela análise dos graus de estranheza, o **registro 5** é o mais anômalo

| d(x, y) | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | Grau de Estranheza |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **1** | 0,00 | **0,17** | 0,17 | **0,11** | 1,14 | 0,20 | 0,28 | 0,21 | **0,14** | 0,28 | 0,14 |
| **2** | 0,17 | 0,00 | **0,09** | 0,18 | 1,30 | 0,36 | 0,11 | **0,05** | 0,17 | **0,11** | 0,08 |
| **3** | 0,17 | **0,09** | 0,00 | 0,12 | 1,26 | 0,33 | 0,18 | **0,09** | **0,10** | 0,18 | 0,09 |
| **4** | **0,11** | 0,18 | **0,12** | 0,00 | 1,14 | 0,21 | 0,29 | 0,20 | **0,03** | 0,29 | 0,09 |
| **5** | **1,14** | 1,30 | 1,26 | **1,14** | 0,00 | **0,94** | 1,41 | 1,34 | 1,16 | 1,41 | **1,07** |
| **6** | **0,20** | 0,36 | 0,33 | **0,21** | 0,94 | 0,00 | 0,47 | 0,40 | **0,24** | 0,47 | 0,22 |
| **7** | 0,28 | **0,11** | 0,18 | 0,29 | 1,41 | 0,47 | 0,00 | **0,09** | 0,27 | **0,00** | 0,07 |
| **8** | 0,21 | **0,05** | **0,09** | 0,20 | 1,34 | 0,40 | **0,09** | 0,00 | 0,19 | 0,09 | 0,08 |
| **9** | **0,14** | 0,17 | **0,10** | **0,03** | 1,16 | 0,24 | 0,27 | 0,19 | 0,00 | 0,27 | 0,09 |
| **10** | 0,28 | **0,11** | 0,18 | 0,29 | 1,41 | 0,47 | **0,00** | **0,09** | 0,27 | 0,00 | 0,07 |

# Detecção de Anomalias com o K-NN

- Exemplo
  - Para confirmar, veja o gráfico



# Detecção de Anomalias com o K-NN

- Base de dados para o trabalho
  - Identificação plantas Iris anômalas (diferentes das demais)
  - http://archive.ics.uci.edu/ml/datasets/Iris

  - Baseado no tamanho da pétala e da sépala, identificar plantas Iris anômalas em relação às demais

# Bibliografia


- *Introdução ao Data Mining*. Steinbach, Michael; Kumar, Vipin; Tan, Pang-ning, Rio de Janeiro: Ed. Ciência Moderna, 2009. Capítulo 10.