MATB16

Laboratório de Inteligência Artificial

Aprendizado Baseado em Instância

Tatiane Nogueira Rios
Ricardo Araújo Rios

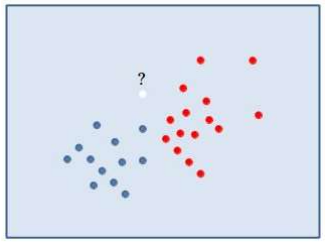
LabIA
Instituto de Computação - UFBA

Objetivo

- Apresentar conceitos necessários para classificação de dados usando Aprendizado Baseado em Instância (Instance-Based Learning - IBL).

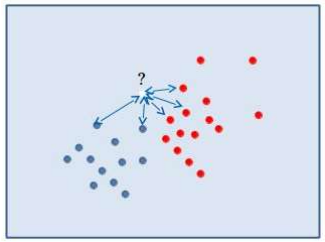
Métodos Baseados em Distância

- Proximidade entre os dados é considerada na realização de predições



Métodos Baseados em Distância

- Proximidade entre os dados é considerada na realização de predições



Métodos Baseados em Distância

- K-Nearest Neighbors (KNN)
  - Algoritmos dos vizinhos mais próximos
  - Objetos relacionados ao mesmo conceito são semelhantes entre si
  - Algoritmo preguiçoso (lazy): não gera modelo, comparação com objetos usados no treinamento

Métodos Baseados em Distância

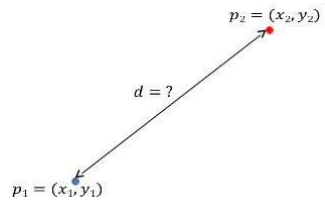
- K-Nearest Neighbors (KNN)
  - Este algoritmo pode ser utilizado para classificação e regressão
  - Variações definidas pelo número de vizinhos

Algoritmo 1-NN

- Espaço de entrada
  - Cada objeto representa um ponto em um espaço definido pelos atributos
- Calcular as distâncias entres dois pontos
  - Métrica mais usual: distância euclidiana

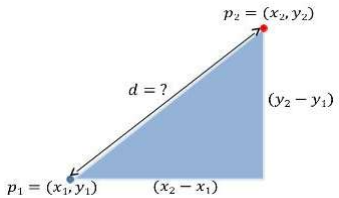
Algoritmo 1-NN

- Distância euclidiana



Algoritmo 1-NN

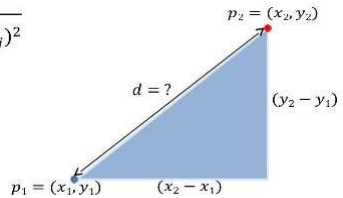
- Distância euclidiana



Algoritmo 1-NN

- Distância euclidiana

$$d(p_i,p_j)=\sqrt{\sum_{l=1}^{d}(p^l{}_i - p^l{}_j)^2}$$

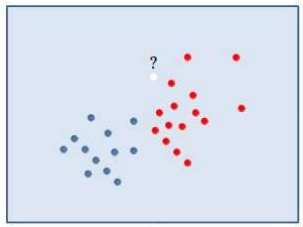


Algoritmo 1-NN

- Fase de treinamento: memorização dos exemplos rotulados do conjunto de treinamento.
- Classificação: cálculo da distância entre o vetor de valores do exemplo não rotulado e cada exemplo armazenado na memória.
- Resultado: o rótulo do novo exemplo será o mesmo rótulo do vizinho mais próximo.

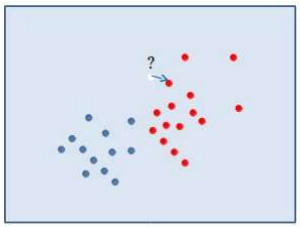
Algoritmo 1-NN

- Qual o rótulo do novo exemplo?



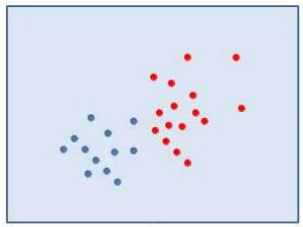
Algoritmo 1-NN

- Qual o rótulo do novo exemplo?



Algoritmo 1-NN

- Qual o rótulo do novo exemplo?



IBL

- Dados correspondem a pontos no espaço d-dimensional, ou seja, seus atributos são numéricos.
- Medidas de distância são afetadas pela escala. Solução: normalizar os atributos.
- E se os dados forem categóricos?

IBL

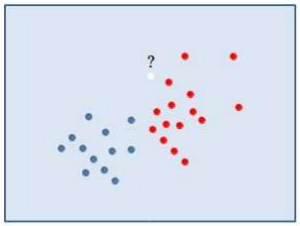
- Dados forem categóricos:
  - Distância de Hamming

$$d_{hamming}(p_i,p_j)=\sum_{r=1}^{n} dist(a_r(p_i),a_r(p_j))$$

$$dist(a_r(p_i),a_r(p_j)) = \begin{cases} 1, se\ a_r(p_i) = a_r(p_j) \\ 0, caso\ contrário \end{cases}$$

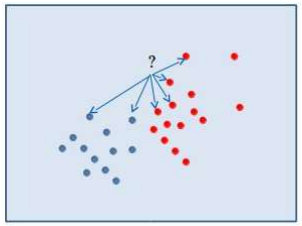
Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste.



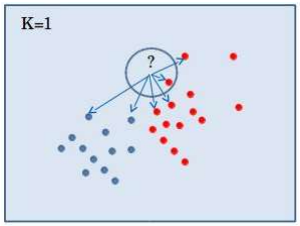
Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste.



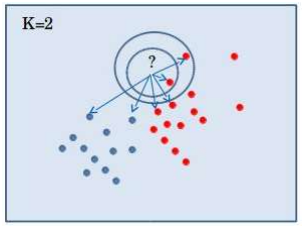
Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste.



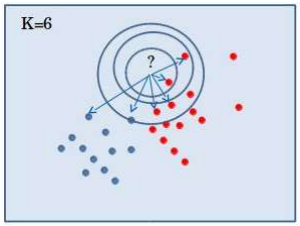
Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste.



Algoritmo K-NN

- Considerar os k objetos do conjunto de treinamento mais próximos do ponto de teste.



Algoritmo K-NN

- Selecione os k exemplos de treinamento mais próximo
- Defina o formato do atributo de saída:
  - Categórico (Classificação)
    - Ex.: Maior número de votos
  - Contínuo (Regressão)
    - Ex.: Média dos k exemplos mais próximos

Algoritmo K-NN

- Qual o melhor valor de k?
  - O problema pode fornecer indícios do valor ideal
  - Validação
  - Geralmente, um valor pequeno e ímpar
    - Valores pares podem gerar empates

Pratique usando KNN sobre dados reais:

- UCI (https://archive.ics.uci.edu/ml/datasets.html);
- UCI KDD Archive (http://kdd.ics.uci.edu/summary.data.application.html);
- Statlib (http://lib.stat.cmu.edu/);
- Delve (http://www.cs.utoronto.ca/~delve/);
- LETOR (http://research.microsoft.com/en-us/um/beijing/projects/letor/index.html);

Referências

- Faceli et al., Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina, LTC, 2015.
- Smola, A. and Vishwanathan, S.V.N., Introduction to Machine Learning, Cambridge University Press, 2008
- Witten et al., Data Mining – Practical Machine Learning Tools and Techniques, 3d edition, Elsevier, 2011.
- J. Han; M. Kamber. Data Mining: Concepts and Techniques, Morgan Kaufmann, 2000