Aprendizado de Máquina

Aula 8: Algoritmos baseados em proximidade (parte 2)

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

Tópicos

- Aprendizado baseado em proximidade
  - Aprendizado baseado em instâncias
- Proximidade
  - Similaridade e dissimilaridade (distância)
- 1-vizinho mais próximo
- Distância de Minkowski e suas variações
- K-vizinhos mais próximos
  - Propriedades de medidas de similaridade e de distância
- Variações
- Conclusão

Tópicos

- Aprendizado baseado em proximidade
  - Aprendizado baseado em instâncias
- Proximidade
  - Similaridade e dissimilaridade (distância)
- 1-vizinho mais próximo
- Distância de Minkowski e suas variações
- K-vizinhos mais próximos
  - Propriedades de medidas de similaridade e de distância
- Variações
- Conclusão

Medidas de dissimilaridade (distância)

- Devem satisfazer as seguintes propriedades (axiomas):
  - Seja $d(p, q)$ a distância entre dois objetos p e q
    - $d(p, q) \geq 0 \ \forall \ p$ e q e $d(p, q) = 0$ se somente se $p = q$ (definida positiva)
    - $d(p, q) = d(q, p) \ \forall \ p$ e q (simetria)
  - Para uma medida de dissimilaridade virar uma métrica, deve satisfazer também:
    - $d(p, q) \leq d(p, r) + d(r, q,) \ \forall \ p, q$ e r (desigualdade triangular)
  - Medidas de dissimilaridade = medidas de distância

Medidas de similaridade

- Medidas de similaridade também têm propriedades bem definidas:
  - Seja *s(p, q)* a similaridade entre dois objetos *p* e *q*
    - *s(p, q) = 1* (similaridade máxima) apenas se *p = q*
    - *s(p, q) = s(q, p)* ∀ *p* e *q* (simetria)
  - Como o conceito de desigualdade triangular não faz muito sentido para similaridade, conceito de métrica de similaridade não está bem definido
  - Matematicamente, conceito complementar ou inverso ao de medida de distância

Funções de edição

- Classe de funções de distância criadas para comparar sequências
  - Em geral biológicas
  - Número de operações de edição necessárias para transformar uma sequência em outra
  - Uma das mais usadas é a distância de Levenshtein
  - Permite três operações de edição
    - Deleção (remover um símbolo da sequência)
    - Inserção (inserir um símbolo na sequência)
    - Substituição (substituir um símbolo da sequência por outro símbolo)

Exemplo

- Qual a distância entre as palavras abaixo?
  - Casa
  - Brisa

Exemplo

- Qual a distância entre as palavras abaixo?
  - Casa
  - Brisa
  - Alterando a palavra casa:
    - Troca “C” por “B” ⇒ Basa
    - Trocar “a” por “r” ⇒ Brsa
    - Inserir “i” depois de “r” ⇒ Brisa
  - Número de operações (distância) = 3

Diversas variações
Ex.: cada operação pode ter um peso diferente

Medidas de similaridade

- Algumas vezes, objetos p e q têm apenas valores binários
  - Ex.: 0110 e 1100
- Medidas de similaridade são também chamadas de coeficientes de similaridade
- Similaridades podem ser computadas usando:
  - M_{01} = número de atributos em p e q em que $p_i = 0$ e $q_i = 1$
  - M_{10} = número de atributos em p e q em que $p_i = 1$ e $q_i = 0$
  - M_{00} = número de atributos em p e q em que $p_i = 0$ e $q_i = 0$
  - M_{11} = número de atributos em p e q em que $p_i = 1$ e $q_i = 1$

Similaridade entre vetores binários

- Coeficiente de Casamento Simples

$$CCS = (M_{11} + M_{00}) / (M_{01} + M_{10} + M_{11} + M_{00})$$

- Coeficiente Jaccard

$$J = (M_{11}) / (M_{01} + M_{10} + M_{11})$$

- Muito usadas em agrupamento de dados

Similaridade cosseno

- Muito usada quando objetos são textos
  - *Bag of words* (palavras que aparecem nos textos)
    - Grande número de atributos
    - Esparsos
- Sejam *p* e *q* vetores representando objetos (textos)
  - cos(*p*, *q*) = (*p*· *q*) / ||*p*|| ||*q*||
    - · : produto interno entre vetores
    - || *x* || : tamanho (norma) do vetor *x*

$$Similaridade_{\text{cosseno}} = \frac{\sum_{k=1}^{d} p_k . q_k}{\sum_{k=1}^{d} p_k^2 . \sum_{k=1}^{d} q_k^2}$$

Similaridade de Pearson

- Coeficiente de correlação de Pearson
- Muito usada em bioinformática e séries temporais
    - Mede correlação linear entre dois vetores

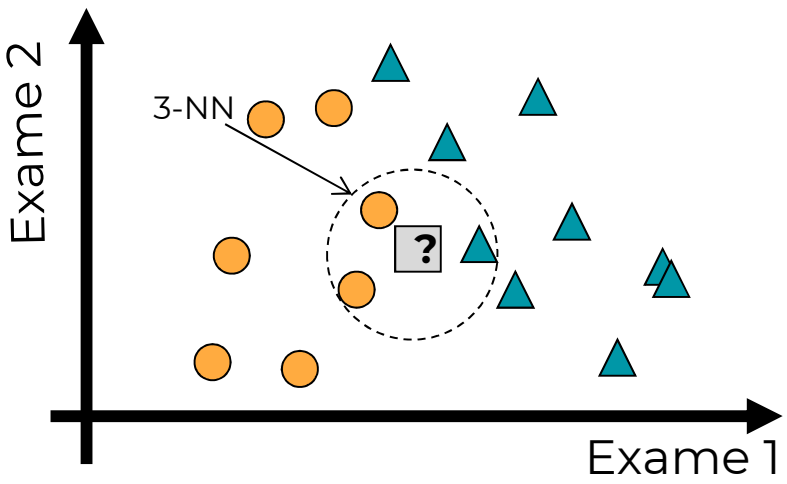
$$Similaridade_{\text{Pearson}} = \frac{\sum_{k=1}^{d}(p_k - \bar{p}).(q_k - \bar{q})}{\sqrt{\sum_{k=1}^{d}(p_k - \bar{p})^2 . \sum_{k=1}^{d}(q_k - \bar{q})^2}}$$

    - Correlação não linear: correlação de Spearman

K-vizinhos mais próximos

- Generalização do 1-vizinho mais próximo
- Um dos algoritmos mais simples de aprendizado de máquina
- Número de vizinhos (valor de k) pode ser definido pelo usuário

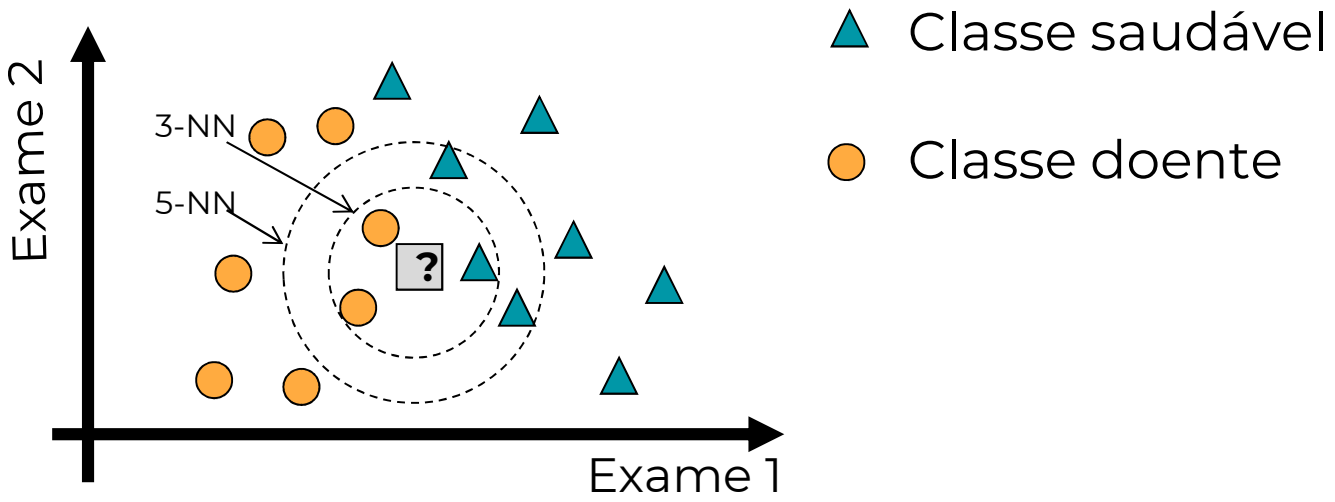
Quantos vizinhos?



Classe saudável

Classe doente

Quantos vizinhos?



Algoritmo k-Vizinhos mais próximos (classificação)

Seja k o número de vizinhos mais próximos
Para cada novo exemplo x
Retornar a classe dos k exemplos (vizinhos) mais próximos a x
Classificar x na classe majoritária dentre as classes retornadas

K-vizinhos mais próximos

- Abordagem local
- Classificação de novos exemplos pode ser lenta
    - Alternativas para reduzir lentidão
        - Seleção de atributos
        - Remoção de exemplos
            - Guardar conjunto de protótipos para cada classe
            - Algoritmos iterativos
                - Remoção sequencial
                - Inserção sequencial

K-vizinhos mais próximos

- Algoritmos iterativos para eliminação
  - Selecionam objetos para serem protótipos
  - Remoção sequencial
    - Conjunto inicial inclui todos os exemplos
    - Descarta exemplos corretamente classificados pelos protótipos
  - Inserção sequencial
    - Conjunto inicial inclui apenas os protótipos
    - Acrescenta exemplos incorretamente classificados pelos protótipos

Variações do K-vizinhos mais próximos

- Normalizar atributos
- Ponderar atributos pela importância
- Ponderar votos dos rótulos pela distância entre exemplos
- Regressão

K-vizinhos mais próximos: escala

- Classificar um objeto x_{novo} (4,1,9) usando o conjunto de treinamento formado pelos objetos x_1 (3,3,100), da classe A e x_2 (10,9,10), da classe B
- Segundo a distância euclidiana, a distância entre os objetos x_{novo} e x_i é dada por:

$$d(x_{novo},x_1) = \sqrt{(4-3)^2 + (1-3)^2 + (9-100)^2}$$

$$= \sqrt{1 + (-2)^2 + (-91)^2} = \sqrt{1+4+8281} \cong 91{,}03$$

$$d(x_{novo},x_2) = \sqrt{(4-10)^2 + (1-9)^2 + (9-10)^2}$$

$$= \sqrt{(-6)^2 + (-8)^2 + (-1)^2} = \sqrt{101} \cong 10{,}05$$

- Todos os atributos contribuíram igualmente no cálculo do valor de distância para a classificação do objeto novo?

K-vizinhos mais próximos: escala

- Alguns atributos assumem uma faixa de valores maior que outros
  - Têm maior influência no cálculo da distância
    - Para evitar isso, os valores dos atributos podem ser escalados
    - Faz com que todos os atributos tenham a mesma faixa de valores
      - Contribuam com o mesmo peso no cálculo da distância

K-vizinhos mais próximos: escala

- Normalizar o terceiro atributo preditivo
- Classificar um objeto x_{novo} (4, 1, 0,9) usando o conjunto de treinamento formado pelos objetos x_1 (3,3,10), da classe A e x_2 (10,9,1), da classe B
- Segundo a distância euclidiana, a distância entre os objetos x_{novo} e x_i é dada por:

$$d(x_{novo},x_1) = \sqrt{(4-3)^2 + (1-3)^2 + (0.9-10)^2}$$

$$= \sqrt{1 + (-2)^2 + (-9{,}1)^2} = \sqrt{1 + 4 + 82{,}81} \cong 9{,}37$$

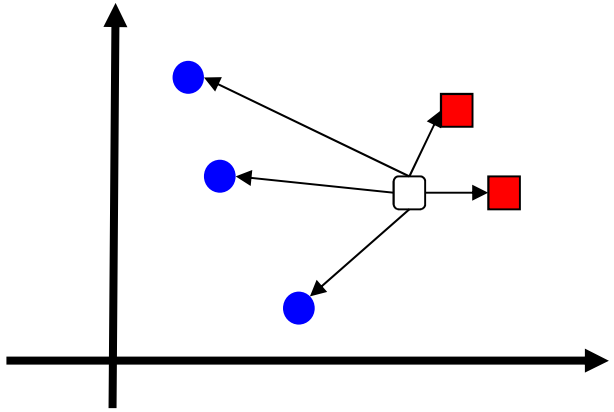
$$d(x_{novo},x_2) = \sqrt{(4-10)^2 + (1-9)^2 + (0{,}9-1)^2}$$

$$= \sqrt{(-6)^2 + (-8)^2 + (-0{,}1)^2} = \sqrt{100{,}01} \cong 10{,}00$$

- Ao normalizar o terceiro atributo preditivo, mudou a classificação do objeto novo

K-vizinhos mais próximos: pesos

- Uma variação popular algoritmo k-NN é associar um peso a cada vizinho proporcional à sua distância ao objeto a ser classificado
- Associa um peso a cada vizinho
    - Proporcional à sua distância ao objeto a ser classificado



K-vizinhos mais próximos: pesos

- Recomendação: ajustar o peso do voto de cada vizinho pela equação:

$$w_i = \frac{1}{d(x_{novo}, x_i)^2}$$

- Quanto mais distante o vizinho mais próximo (x_i) estiver do exemplo a ser classificado (x_{novo}), menor o seu peso
    - Permite, ao invés de apenas os k-vizinhos mais próximos, usar todo o conjunto de treinamento
    - Exemplos muito distantes terão pouca influência na classificação do novo exemplo

K-vizinhos mais próximos: regressão

- O algoritmo k-NN pode ser adaptado para tarefas de regressão
  - Rótulos dos objetos são valores contínuos
- Valor predito: a média dos valores dos rótulos dos k-vizinhos mais próximos
- Exemplo: aplicação que precisa fornecer o salário de um funcionário dadas suas características
  - Formação
  - Experiência
  - Cargo atual
  - ...

Vantagens

- Simples
- Boa capacidade preditiva em várias aplicações
- Tempo de treinamento baixo (inexistente)
- Inerentemente incremental

Desvantagens

- Tempo de processamento na fase de teste pode ser elevado
- Usa apenas informação local para prever o valor do rótulo
- Sensível a atributos irrelevantes
- Atributos quantitativos precisam ser escalados
- Sensível a presença de outliers
- Por não ter modelo, não é “interpretável”

Conclusão

- Aprendizado baseado em distância
- Conceitos básicos
- Medidas de distância
- Algoritmo k-vizinhos mais próximos
- Variações
- Exemplos

Fim da apresentação