Mineração de Dados em Biologia Molecular

Métodos baseados em distância

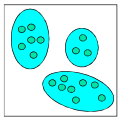
Docente: André C. P. L. F. de Carvalho
PAE: Victor Hugo Barella

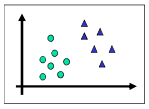
Principais tópicos

- Aprendizado baseado em instâncias
- 1-vizinho mais próximo
- K-vizinhos mais próximos
- Conclusão

AM e Geometria

- Medidas de distância podem ser usadas para
  - Agrupar dados
    - Ex.: K-médias
  - Classificar novos dados
    - Ex.: K-NN
  - Existem várias medidas

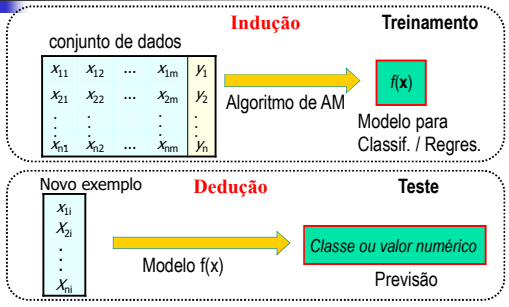




Algoritmos de AM preditivos

- Induzem modelos (funções) preditivas
  - Indução de modelos para parte do conjunto de dados
    - Subconjunto de treinamento
  - Modelo pode ser aplicado a novos dados (predição)
    - Subconjunto de teste
- Principais tarefas:
  - Regressão
  - Classificação

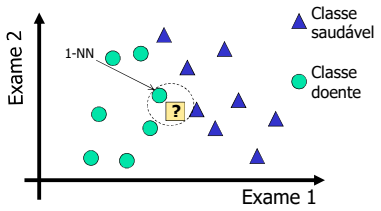
Algoritmos de AM preditivos



1-vizinho mais próximo

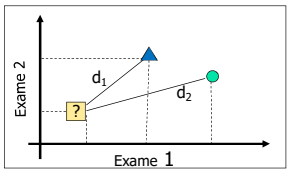
- Versão simples do algoritmo k-NN
  - Geralmente usado para classificação
- Algoritmo *lazy* (preguiçoso)
  - Acessa os dados de treinamento apenas quando vai classificar um novo objeto
  - Não constrói um modelo explicitamente
    - Diferente de algoritmos *eager*
      - Induzem um modelo
        - Ex.: ADs, RNs e SVMs

1-vizinho mais próximo



Métodos baseados em distância

- Consideram proximidade entre dados
  - Similaridade
  - Dissimilaridade



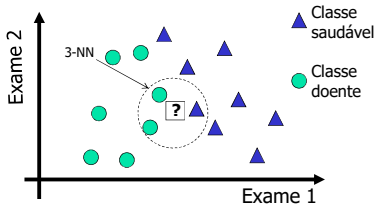
1-vizinho mais próximo

- Algoritmo de AM baseado distância muito simples
  - Mesma classe do vizinho mais próximo
  - É seguro confiar em apenas um vizinho?
- Número de vizinhos pode variar
  - K-vizinhos mais próximos

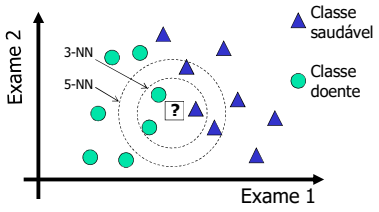
Quantos vizinhos?

- K muito grande
  - Vizinhos podem ser muito diferentes
  - Predição tendenciosa para classe majoritária
- K muito pequeno
  - Apenas os objetos muito parecidos serão considerados
    - Informação pode não ser suficiente (confiável) para uma boa classificação
  - Previsão pode ser instável
    - Influenciada pela presença de ruído

Quantos vizinhos?



Quantos vizinhos?



K-Vizinhos mais próximos

Seja k o número de vizinhos mais próximos
Para cada novo exemplo x
    Definir a classe dos k exemplos (vizinhos) mais próximos
    Classificar x na classe majoritária entre seus k vizinhos mais próximos

K-vizinhos mais próximos

- Abordagem local
- Classificação de novos exemplos pode ser lenta
  - Alternativas para maior rapidez:
    - Seleção de atributos
    - Eliminação de exemplos
      - Guardar apenas protótipos das classes
      - Utilizar algoritmos iterativos para seleção de exemplos

K-vizinhos mais próximos

- Algoritmos iterativos para seleção
  - Separam um protótipo para cada classe
  - Eliminação sequencial
    - Conjunto inicial com todos os exemplos
    - Descarta exemplos corretamente classificados pelos protótipos
  - Inserção sequencial
    - Conjunto inicial vazio
    - Acrescenta exemplos incorretamente classificados pelos protótipos

K-vizinhos mais próximos

- Normalizar atributos
- Ponderar atributos
- Ponderar voto por distância entre exemplos
- Regressão
- Naturalmente incremental

Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício 1

- Usar K-NN e os exemplos anteriores para definir as classes dos exemplos de teste
  - Usar k = 1, 3 e 5
- Exemplos de teste
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Exercício 2

- Dada a tabela abaixo, com k =1 e 3, definir a classe dos exemplos:
  - (RJ, Médio, 178, 2000)
  - (SP, Superior, 200, 800)

Estado	Escolaridade	Altura	Salário	Classe
SP	Médio	180	3000	A
RJ	Superior	174	7000	B
RS	Médio	180	600	B
RJ	Superior	100	2000	A
SP	Fundam.	178	5000	A
RJ	Fundam.	188	1800	A

Conclusão

- Aprendizado de máquina preditivo
- Aprendizado baseado em distância
- K-vizinhos mais próximos
- Variações
- Tarefas de classificação e regressão
- Exemplos

Perguntas?