Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

Introdução aos Modelos Preditivos

Prof. Tiago A. Almeida

Modelos preditivos

- Algoritmo de AM preditivo: função que, dado um conjunto de exemplos rotulados, constrói um estimador

Classificação

- Rótulos nominais (conjunto discreto e não ordenado de valores)
Ex. {doente, saudável}, {bom pagador, mau pagador}
- *Estimador é chamado classificador*

Regressão

- Rótulos contínuos (conjunto infinito ordenado de valores)
Ex. peso, temperatura, vazão de água
- Estimador é chamado regressor

Estimadores podem ser vistos como funções

Modelos preditivos

- Definição formal: dado conjunto de observações
- $\mathbf{D} = \{(\mathbf{x}^{(i)}, f(\mathbf{x}^{(i)})), i = 1, ..., m\}$
  - f representa uma função desconhecida: função objetivo
    - Mapeia entradas em saídas correspondentes
  - Algoritmo preditivo aprende aproximação $\hat{f}$
    - Que permite estimar valor de f para novos objetos $\mathbf{x}$

Classificação

$y^{(i)} = f(\mathbf{x}^{(i)}) \in \{c_1, ..., c_k\}$

Regressão

$y^{(i)} = f(\mathbf{x}^{(i)}) \in \Re$

Modelos preditivos

- Exemplos de conjuntos de dados:

Conjunto de dados `iris`

TamP	LargP	TamS	LargS	Espécie
5,1	3,5	1,4	0,2	Setosa
4,9	3	1,4	0,2	Setosa
7	3,2	4,7	1,4	Versicolor
6,4	3,2	4,5	1,5	Versicolor
6,3	3,3	6	2,5	Virgínica
5,8	2,7	5,1	1,9	Virgínica

Classificação

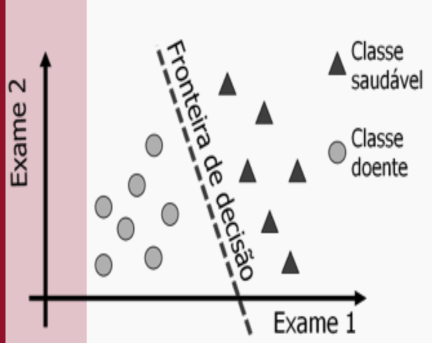
Conjunto de dados `swiss`

Fertilidade	Agricultura	Educação	Renda	Mortalidade
80,2	17	12	9,9	22,2
83,1	45,1	9	84,8	22,2
92,5	39,7	5	93,4	20,2
85,8	36,5	7	33,7	20,3
76,9	43,5	15	5,2	20,6

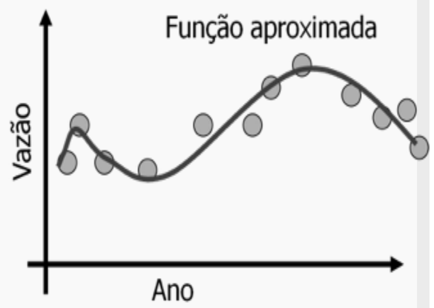
Regressão

Modelos preditivos

- Ilustração das tarefas:



Classificação



Regressão

Modelos preditivos

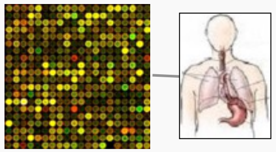
- Outros conceitos:
  - Atributo alvo é também designado como variável dependente ou objetivo
    - Em classificação: classe
  - Atributos restantes são designados como de entrada, preditivos ou variáveis independentes
    - Utilizados como entrada para fazer a predição

Modelos preditivos: classificação

- Classificação:
  - Meta: encontrar fronteira de decisão que separe classes
  - Diferentes algoritmos de AM podem encontrar diferentes fronteiras
  - Mesmo algoritmo pode também encontrar fronteiras diferentes
    - Diferenças nos dados de treinamento
    - Variações na ordem de apresentação dos exemplos
    - Processos estocásticos internos

Classificação

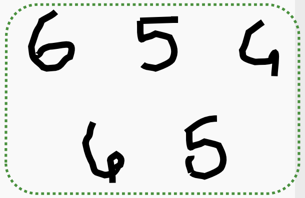
- Exemplos:
  - Diagnóstico de doenças
    - Paciente é doente ou não?



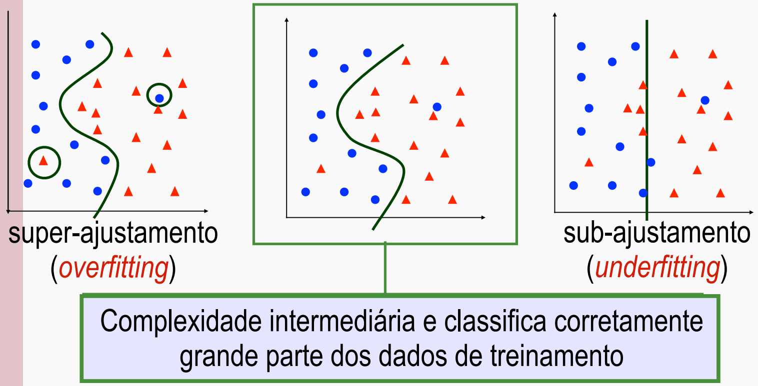
  - Distribuição geográfica de espécies
    - Espécie está presente na região?

Classificação

- Exemplos:
  - Categorização de textos
    - De que assuntos uma notícia trata?
  - Classificação de imagens
    - Qual é o dígito escrito em uma imagem digitalizada?
    - Imagem médica contém um tumor?
  - Detecção de fraudes
    - O cliente será um bom pagador?
  - Filtro de spam
    - A mensagem é um spam ou não?



Classificação



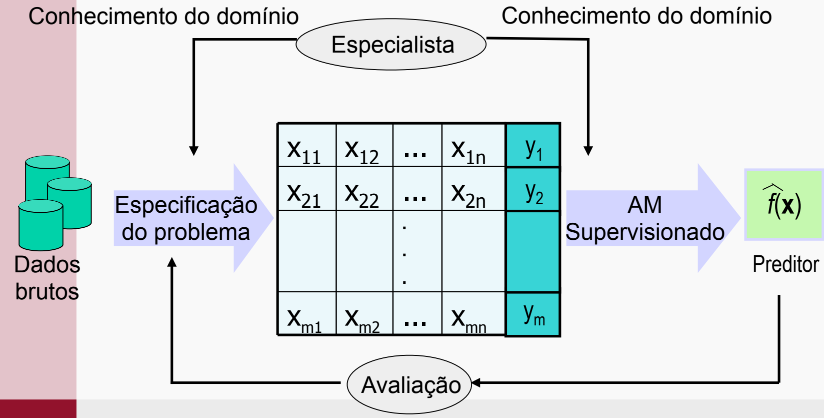
Classificação

- Técnicas de AM:
  - K-vizinhos mais próximos
  - Regressão Logística
  - Árvores de Decisão (C4.5)
  - Conjuntos de regras
  - Redes Neurais Artificiais
  - Máquinas de Vetores de Suporte
  - Redes Bayesianas
  - Etc.

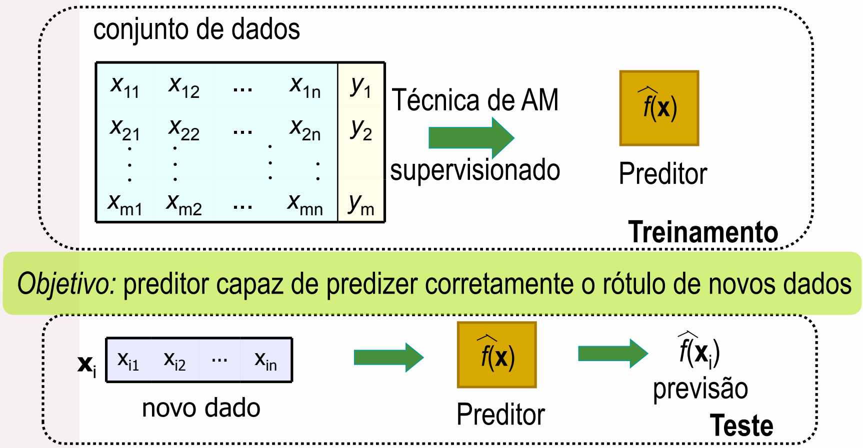
> Cada técnica de AM utiliza um mecanismo distinto na escolha da hipótese

Modelos preditivos

- Esquema geral:



Modelos preditivos



Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

Métodos Baseados em Distâncias

Métodos baseados em distâncias

- Técnicas de AM que consideram proximidade entre os dados para realizar predições

Hipótese: dados similares tendem a estar concentrados em uma mesma região do espaço de entradas

- E dados que não são similares estarão distantes entre si

Proximidade

- Medida de proximidade entre pares de objetos pode ser de:

Similaridade

- Mede o quanto dois objetos são parecidos
- Quanto mais parecidos ⇒ maior o valor

Dissimilaridade

- Mede o quanto dois objetos são diferentes
- Quanto mais diferentes ⇒ maior o valor

Escolha da medida deve considerar tipos e escalas dos atributos, além de propriedades dos dados que se deseja focar

Similaridade/dissimilaridade

- Normalmente as medidas satisfazem algumas propriedades, tais como:
  - Os objetos não são diferentes de si próprios
    - $d(\mathbf{x}_i, \mathbf{x}_i) = 0$
    - Em similaridade, objetos são similares a si próprios
      - $s(\mathbf{x}_i, \mathbf{x}_i) = 1$
  - Simetria
    - $d(\mathbf{x}_i, \mathbf{x}_j) = d(\mathbf{x}_j, \mathbf{x}_i)$
  - Positividade
    - $d(\mathbf{x}_i, \mathbf{x}_j) \geq 0$

Todas medidas de distância (medem dissimilaridade) satisfazem essas propriedades

Distância Mahattan

- Distância de Manhattan:
  - Também chamada distância bloco-cidade
    - Equivalente a Hamming para atributos binários
  - Norma-1 ou distância L_1: || p-q ||

$$d(p,q) = \sum_{i=1}^{m} |p_i - q_i|$$

Distância Euclidiana

- Distância Euclidiana:
  - Medida de distância mais popular
  - Norma-2 ou distância L_2: $|| p-q ||^2$

$$d(p,q) = \sqrt{\sum_{i=1}^{m} (p_i - q_i)^2}$$

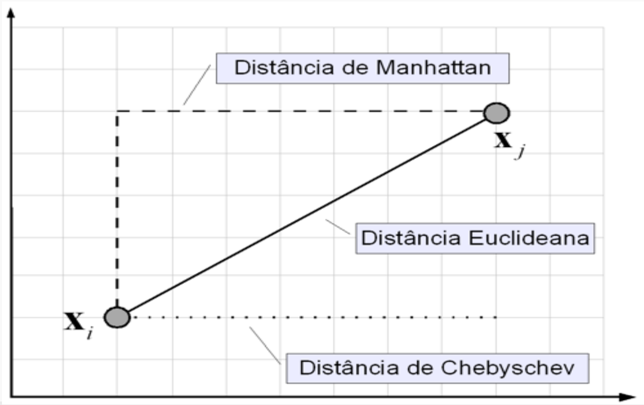
Distância de Chebyshev

- Distância de Chebyshev:
  - Também chamada de distância Supremum
    - Diferença absoluta máxima
  - Norma-∞ ou distância L_∞: $|| p-q ||^\infty$

$$d(p,q) = max_i(|p_i - q_i|)$$

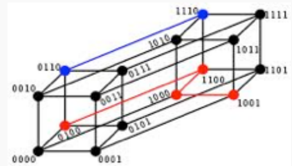
Medidas de distância

- Interpretação das medidas de distância



Medidas para atributos qualitativos

- Medidas obtidas pela soma das contribuições individuais
  - Ex. Distância de Hamming
    - Conta número de atributos categóricos com valores diferentes nos dois objetos
    - Varia em [0, *d*]
      - Valor 0 significa maior similaridade



Exercício

- Considere dois vetores:
  - $x^{(1)}$ = [3 0 1 3 8]
  - $x^{(2)}$ = [4 1 6 3 3]
- Calcule as seguintes distâncias entre $x^{(1)}$ e $x^{(2)}$:
  - Manhattan
  - Euclides
  - Chebyshev
  - Hamming

Exercício

- Considere dois vetores:
  - $x^{(1)}$ = [3 0 1 3 8]
  - $x^{(2)}$ = [4 1 6 3 3]
- Calcule as seguintes distâncias entre $x^{(1)}$ e $x^{(2)}$:
  - Manhattan = (|3-4|+|0-1|+|1-6|+|3-3|+|8-3|) = **12**
  - Euclides
  - Chebyshev
  - Hamming

Exercício

- Considere dois vetores:
  - $x^{(1)} = [3\ 0\ 1\ 3\ 8]$
  - $x^{(2)} = [4\ 1\ 6\ 3\ 3]$
- Calcule as seguintes distâncias entre $x^{(1)}$ e $x^{(2)}$:
  - Manhattan = (|3-4|+|0-1|+|1-6|+|3-3|+|8-3|) = **12**
  - Euclides = $\sqrt{((3-4)^2+(0-1)^2+(1-6)^2+(3-3)^2+(8-3)^2)}$ = **7,2**
  - Chebyshev
  - Hamming

Exercício

- Considere dois vetores:
  - $x^{(1)} = [3\ 0\ 1\ 3\ 8]$
  - $x^{(2)} = [4\ 1\ 6\ 3\ 3]$
- Calcule as seguintes distâncias entre $x^{(1)}$ e $x^{(2)}$:
  - Manhattan = (|3-4|+|0-1|+|1-6|+|3-3|+|8-3|) = **12**
  - Euclides = $\sqrt{((3-4)^2+(0-1)^2+(1-6)^2+(3-3)^2+(8-3)^2)}$ = **7,2**
  - Chebyshev = max(|3-4|,|0-1|,|1-6|,|3-3|,|8-3|) = **5**
  - Hamming

Exercício

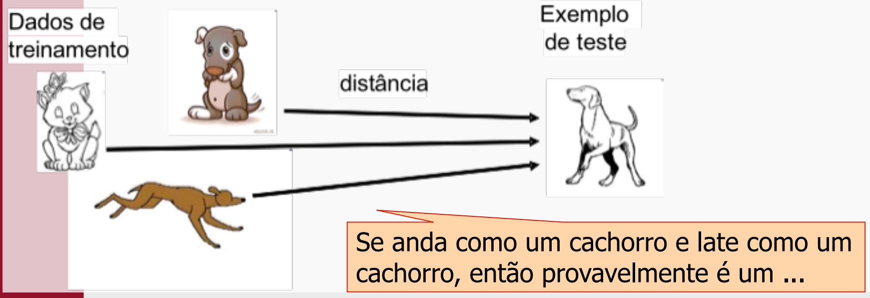
- Considere dois vetores:
  - $x^{(1)} = [3\ 0\ 1\ 3\ 8]$
  - $x^{(2)} = [4\ 1\ 6\ 3\ 3]$
- Calcule as seguintes distâncias entre $x^{(1)}$ e $x^{(2)}$:
  - Manhattan = (|3-4|+|0-1|+|1-6|+|3-3|+|8-3|) = **12**
  - Euclides = $\sqrt{((3-4)^2+(0-1)^2+(1-6)^2+(3-3)^2+(8-3)^2)}$ = **7,2**
  - Chebyshev = max(|3-4|,|0-1|,|1-6|,|3-3|,|8-3|) = **5**
  - Hamming = **4**

Inteligência Artificial: Uma Abordagem de Aprendizado de Máquina

Método dos Vizinhos mais Próximos

Algoritmo dos vizinhos mais próximos

- Algoritmo de AM mais simples
  - Intuição: *Objetos relacionados ao mesmo conceito são semelhantes entre si*



Algoritmo dos vizinhos mais próximos

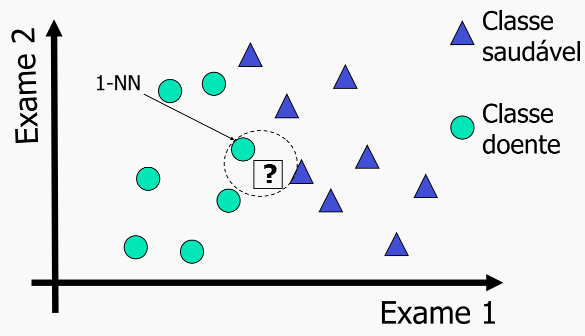
- Rotula novos objetos com base nos exemplos do conjunto de treinamento mais próximos a ele
  - É um algoritmo preguiçoso (*lazy*)
    - Não aprende modelo compacto, memoriza objetos de treinamento
    - Adia computação para a fase de classificação
  - Há variações de acordo com o número de vizinhos mais próximos adotado

Algoritmo 1-vizinho mais próximo

- Variação mais simples: 1-NN
  - 1-*Nearest Neighbor*
  - Cada objeto representa um ponto no espaço de entradas
  - Definindo métrica, é possível calcular distâncias
  - Métrica mais usual: distância euclidiana
  - Treinamento: memoriza exemplos rotulados do conjunto de treinamento
  - Classificação de novo exemplo: classe do exemplo de treinamento mais próximo

Algoritmo 1-vizinho mais próximo

- Ex. 1-NN



Algoritmo 1-vizinho mais próximo

Algoritmo 1-NN

Entrada: conjunto de treinamento $\mathbf{D} = \{(\mathbf{X}_{(m \times n)}, \mathbf{Y}_{(m \times 1)})\}$, objeto de teste a ser classificado $\mathbf{x}_{(1 \times n)}$

Saída: y

$d_{min} \leftarrow \infty$

para cada $i \in 1, \ldots, m$ **faça**

  se $d(\mathbf{x}, \mathbf{X}^{(i)}) < d_{min}$ **então**

    $d_{min} \leftarrow d(\mathbf{x}, \mathbf{X}^{(i)})$

    $idx \leftarrow i$

  fim-se

fim-para

$y = Y^{(idx)}$

Retorna y

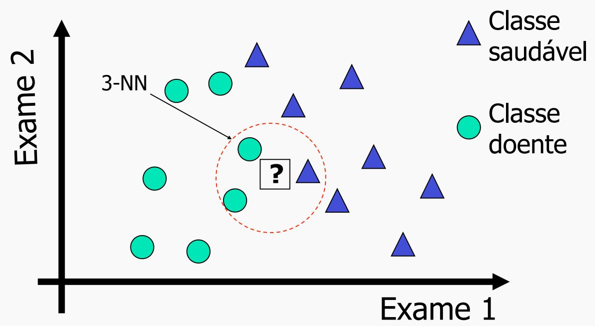
Algoritmo k-vizinhos mais próximos

- Extensão imediata do 1-NN considerando mais vizinhos
  - *k* vizinhos mais próximos
    - *k* é parâmetro do algoritmo
  - Cada vizinho vota em uma classe
    - Previsões são então agregadas

Classificação	Regressão
$\hat{f}(\mathbf{x}_t) \leftarrow \text{moda}(\hat{f}(\mathbf{x}_1), \ldots, \hat{f}(\mathbf{x}_k))$	$\hat{f}(\mathbf{x}_t) \leftarrow \text{média}(\hat{f}(\mathbf{x}_1), \ldots, \hat{f}(\mathbf{x}_k))$ ou $\hat{f}(\mathbf{x}_t) \leftarrow \text{mediana}(\hat{f}(\mathbf{x}_1), \ldots, \hat{f}(\mathbf{x}_k))$

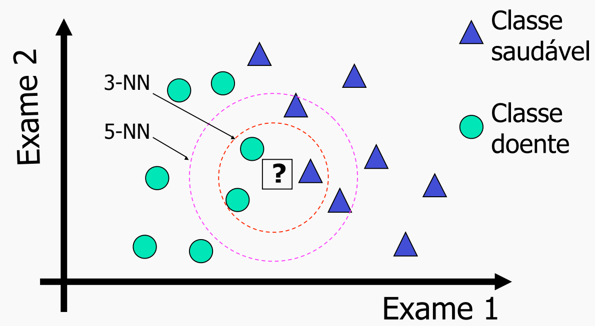
Algoritmo k-vizinhos mais próximos

- Ex.



Algoritmo k-vizinhos mais próximos

- Ex.



Quantos vizinhos?

- *k* muito grande
  - Vizinhos podem ser muito diferentes
  - Predição tendenciosa para classe majoritária
- *k* muito pequeno
  - Não usar informação suficiente
  - Previsão pode ser instável

Frequentemente usa k pequeno e ímpar (3, 5, ...). Valores pares não são usuais em classificação por poderem levar a empates

Quantos vizinhos?

- Algumas estratégias para definir *k*:
  - Fazer busca em grid
  - Associar peso à contribuição de cada vizinho
    - De forma inversamente proporcional à distância

Análise do algoritmo

- Vantagens:
  - O algoritmo de treinamento é simples
    - Armazenar os objetos
  - É aplicável mesmo em problemas complexos
  - É um algoritmo naturalmente incremental
    - Novos exemplos $\Rightarrow$ basta armazená-los na memória

Análise do algoritmo

- Desvantagens:
  - Não obtém uma representação compacta dos dados
    - Não se tem modelo explícito a partir dos dados
  - Predição pode ser custosa
    - Requer calcular distâncias a todos os objetos de treinamento
  - É afetado pela presença de atributos redundantes e irrelevantes
  - Problemas com dimensionalidade elevada
    - Objetos ficam equidistantes

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	saudável

Exercício

- 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)

Atributo contendo nome não é usado

Exercício

- Exemplos de teste:
  - (Luis, não, não, pequenas, sim)
    - $d(\mathbf{x}, \mathbf{x}^{(1)}) = 1 + 1 + 0 + 0 = 2$
    - $d(\mathbf{x}, \mathbf{x}^{(2)}) = 0 + 0 + 1 + 1 = 2$
    - $d(\mathbf{x}, \mathbf{x}^{(3)}) = 1 + 1 + 0 + 1 = 3$
    - $d(\mathbf{x}, \mathbf{x}^{(4)}) = 1 + 0 + 1 + 0 = 2$
    - $d(\mathbf{x}, \mathbf{x}^{(5)}) = 1 + 0 + 0 + 0 = 1$
    - $d(\mathbf{x}, \mathbf{x}^{(6)}) = 0 + 0 + 1 + 0 = 1$

Distância de Hamming

k = 1: saudável
k = 3: saudável
k = 5: saudável

Exercício

- Exemplos de teste:
  - (Laura, sim, sim, grandes, sim)
    - $d(\mathbf{x}, \mathbf{x}^{(1)}) = 0 + 0 + 1 + 0 = 1$
    - $d(\mathbf{x}, \mathbf{x}^{(2)}) = 1 + 1 + 0 + 1 = 3$
    - $d(\mathbf{x}, \mathbf{x}^{(3)}) = 0 + 0 + 1 + 1 = 2$
    - $d(\mathbf{x}, \mathbf{x}^{(4)}) = 0 + 1 + 0 + 0 = 1$
    - $d(\mathbf{x}, \mathbf{x}^{(5)}) = 0 + 1 + 1 + 0 = 2$
    - $d(\mathbf{x}, \mathbf{x}^{(6)}) = 1 + 1 + 0 + 0 = 2$

Distância de Hamming

k = 1: doente
k = 3: doente
k = 5: saudável