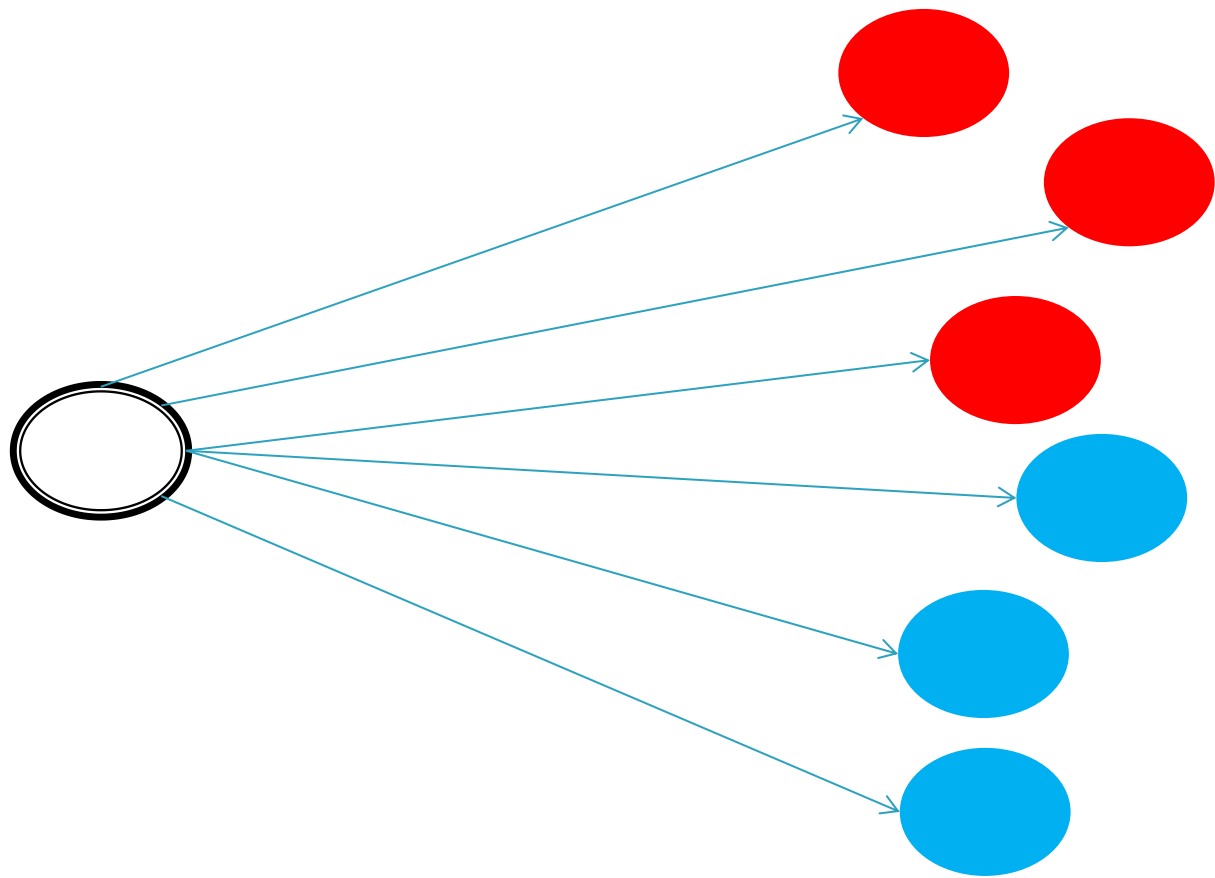




Aprendizagem baseada em instâncias (kNN)

Jones Granatyr





Introdução

- ▶ Muitos métodos de aprendizagem constroem uma descrição geral e explícita da função alvo a partir de exemplos de treinamento (os dados são descartados após a geração do modelo)
- ▶ Os métodos de aprendizagem baseados em instâncias simplesmente armazenam os exemplos de treinamento
- ▶ A generalização/previsão é feita somente quando uma nova instância tiver que ser classificada (lazy)

Cálculo da Distância Euclidiana

$$DE(x, y) = \sqrt{\sum_i^p (x_i - y_i)^2}$$

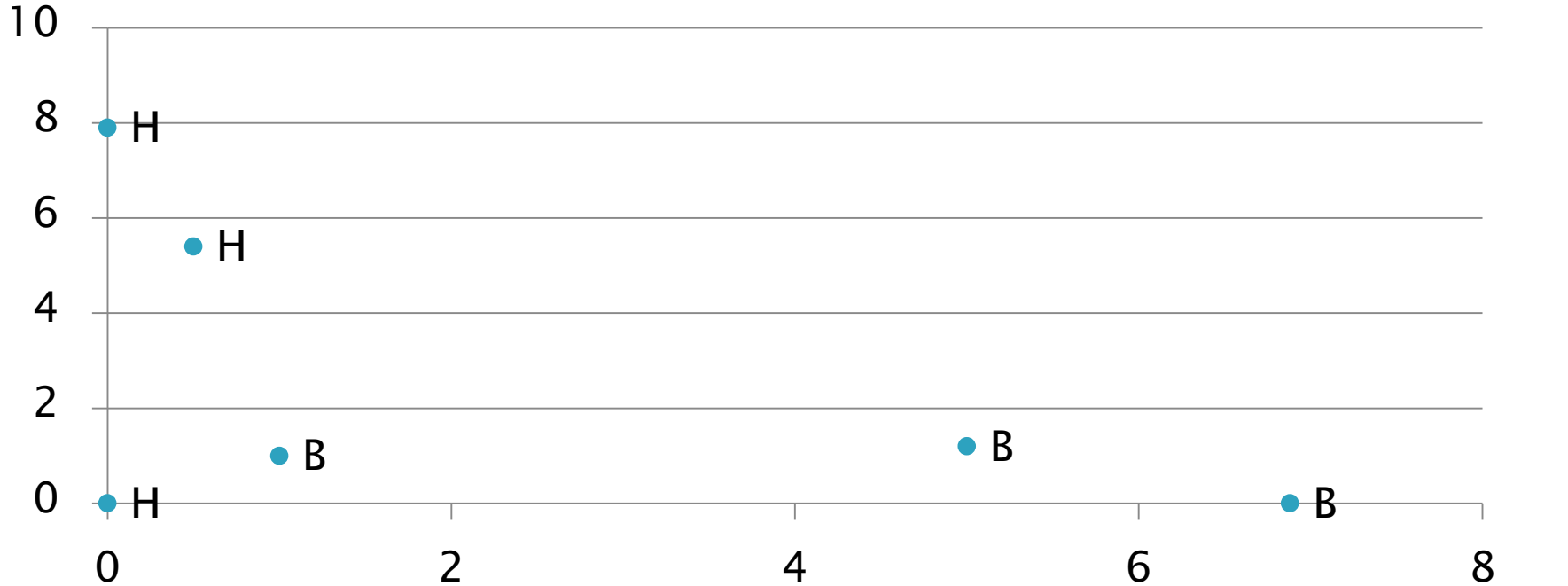
Exemplo

- ▶ $x_i = 5, 7, 9$
- ▶ $y_j = 5, 5, 5$
- ▶ Subtração de cada posição do vetor
 - $5 - 5 = 0$
 - $7 - 5 = 2$
 - $9 - 5 = 4$
- ▶ Elevação ao quadrado
 - $\text{Raiz}(0^2 + 2^2 + 4^2)$
 - $\text{Raiz}(0 + 4 + 16)$
 - $\text{Raiz}(20)$
- ▶ **Distância Euclidiana = 4,47**

	Violência	Romance	Ação	Comédia
Invocação do Mal	0.6	0.0	0.3	0.0
Floresta Maldita	0.9	0.0	0.5	0.1
Meu Passado me Condena	0.1	0.2	0.1	0.9
Vizinhos 2	0.0	0.1	0.2	0.8
Deadpool	0.5	0.0	0.7	0.1
Tirando o atraso	0.0	0.2	0.2	0.8

Azul Homer x Laranja Bart

Azul



Laranja

Laranja	Azul calção	Azul sapato	Marrom boca	Azul calça	Cinza sapato	Classe
6.88	3.49	1.48	0.00	0.00	0.06	Bart
5.00	3.18	1.00	0.00	0.00	0.03	Bart
0.00	0.00	0.00	0.00	0.00	0.07	Bart
0.00	0.00	0.02	3.25	4.48	0.08	Bart
0.00	0.00	0.00	3.19	7.89	1.19	Homer
4.49	3.75	0.89	0.00	0.00	0.05	Homer

Base original

História do crédito	Dívida	Garantias	Renda anual	Risco
Ruim	Alta	Nenhuma	< 15.000	Alto
Desconhecida	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto
Desconhecida	Baixa	Nenhuma	>= 15.000 a <= 35.000	Moderado
Desconhecida	Baixa	Nenhuma	> 35.000	Alto
Desconhecida	Baixa	Nenhuma	> 35.000	Baixo
Desconhecida	Baixa	Adequada	> 35.000	Baixo
Ruim	Baixa	Nenhuma	< 15.000	Alto
Ruim	Baixa	Adequada	> 35.000	Moderado
Boa	Baixa	Nenhuma	> 35.000	Baixo
Boa	Alta	Adequada	> 35.000	Baixo
Boa	Alta	Nenhuma	< 15.000	Alto
Boa	Alta	Nenhuma	>= 15.000 a <= 35.000	Moderado
Boa	Alta	Nenhuma	> 35.0000	Baixo
Ruim	Alta	Nenhuma	>= 15.000 a <= 35.000	Alto

Base KNN

História do crédito	Dívida	Garantias	Renda anual	Risco
3	1	1	1	Alto
2	1	1	2	Alto
2	2	1	2	Moderado
2	2	1	3	Alto
2	2	1	3	Baixo
2	2	2	3	Baixo
3	2	1	1	Alto
3	2	2	3	Moderado
1	2	1	3	Baixo
1	1	2	3	Baixo
1	1	1	1	Alto
1	1	1	2	Moderado
1	1	1	3	Baixo
3	1	1	2	Alto

Outras distâncias


- ▶ Coeficiente de Pearson
 - ▶ Índice de Tanimoto
 - ▶ City Block
- 

Vantagens x desvantagens

▶ Vantagens

- Fácil interpretação
- Técnica *on-line* sem computação adicional

▶ Desvantagens

- Todos os dados precisam estar carregados
 - Pode ser demorado
 - Escalonamento dos dados
- 

Conclusão