k-Nearest Neighbor (k-NN)

Diego Bertolini Gonçalves

diegob@utfpr.edu.br

08/05/2014

Estrutura da Aula

1. Revisão Aula Anterior
2. Algoritmo k-NN

Introdução
Formas de Aprendizagem

- **Aprendizagem Supervisionada**
  - Fornecemos a "boa resposta" durante o treinamento ;
  - k-NN ;
- **Aprendizagem Não-Supervisionada**
  - Em geral buscam encontrar aglomerados de conjuntos de dados semelhantes entre si (clusters) ;
- **Aprendizagem por Reforço**
  - Não damos a "boa resposta". O sistema elabora uma hipótese e o recompensamos ou punimos ;

Introdução
Aprendizagem Supervisionada

Características

- Na Aprendizagem Supervisionada, as classes são conhecidas *a priori* ;
- É possível ajustar os pesos em função das respostas corretas ;
- O desafio é capacitar o sistema a atuar de acordo com o padrão observado nos exemplos de entrada e saída ;
- Qual a probabilidade que um cliente com um perfil compre determinado produto?

Introdução
Extração de Características

Características

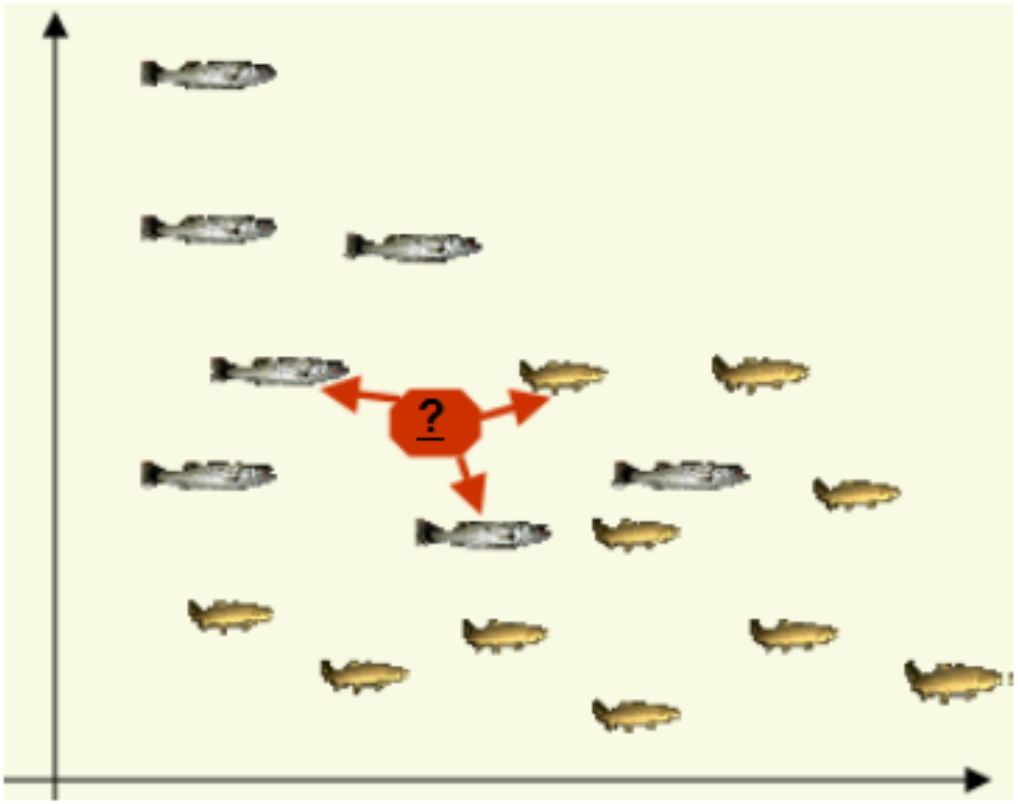
- Escolha das características;
- Uso de características numéricas;
- Seleção de Características;

k-NN
Características

- k-NN ou k Vizinhos mais Próximos ;
- Algoritmo de classificação supervisionado ;
- Clássico e muito Simples ;
- Pode ser implementado facilmente ;
- Baseia-se na analogia de vizinhos mais próximos ;
- Assume que todas as instâncias correspondem a pontos em um espaço n-dimensional.

k-NN
Exemplo

O clássico problema de classificar **Salmão** e **Robalos**.



k-NN
Requisitos

A partir de um elemento desconhecido **x** o qual queremos classificar usando o k-NN, necessitamos:

1. Um conjunto para treinamento ;
2. Uma métrica para calcular a distância entre **x** e as demais amostras ;
3. Definir um valor para k, ou seja, quantos vizinhos iremos considerar?

k-NN
Algoritmo Clássico

Assim, para classificarmos um exemplo desconhecido **x**, temos basicamente três passos:

1. Inicialmente, calcula-se a distância entre o exemplo desconhecido **x** e todos os exemplos do conjunto de treinamento ;
2. Identifica-se os k vizinhos mais próximos ;
3. A classificação é feita associando o exemplo desconhecido **x** à classe que for mais frequente, entre os k exemplos mais próximos de **x** ;
    - Utiliza o **voto majoritário** para definir a classe mais frequente.

k-NN
Distância entre dois Pontos

Existem várias métricas diferentes para calcular a distância entre dois pontos, como: City Block, Minkowsky, entre outras. Porém a mais comum é a **Distância Euclidiana**.

- **Distância Euclidiana**:

$$d(x_i, x_j) = \sqrt{\sum_{i=1}^{n}(x_i - x_j)^2} \quad (1)$$

k-NN
Valor de k

Após possuirmos as distâncias entre o elemento desconhecido **x** e todas as amostras de treinamento, temos que:

- Considerar o voto majoritário avaliando os rótulos de classe dos k vizinhos mais próximos ;
- Mas como escolher o valor de k?

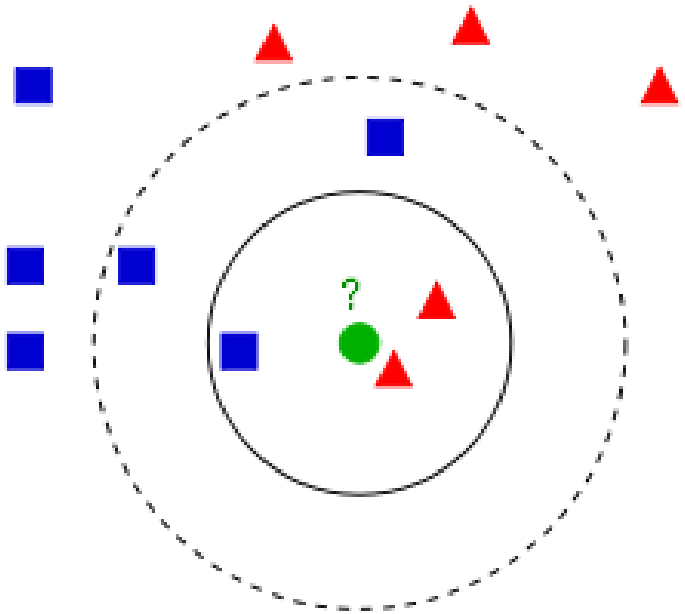
k-NN
Valor de k

Algumas considerações quanto ao valor de k.

- Se k for muito pequeno (k = 1), a classificação pode ser sensível a *ruídos* ;
- Se k for muito grande, podemos estar incluindo elementos de outras classes ;
- É comum avaliarmos diferentes valores para k ;
- A utilização de valores ímpares para k garante que não haverá empate.

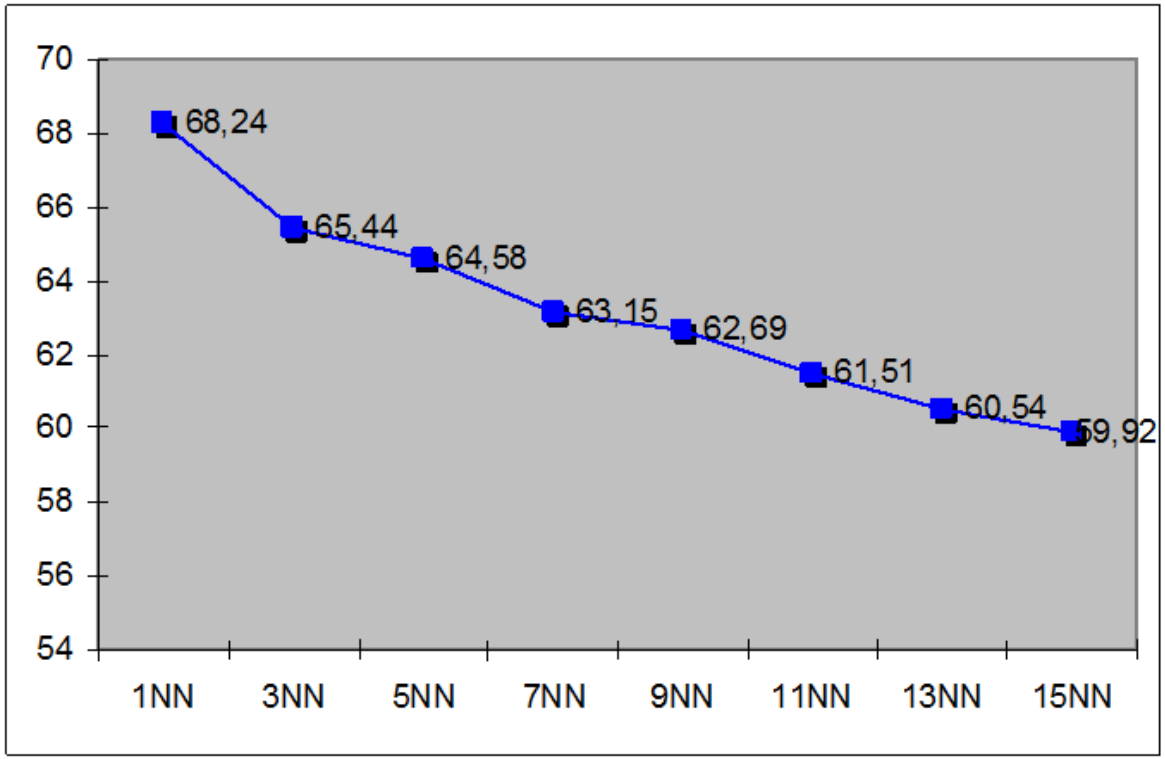
k-NN
Valor de k

- $k = 1$
- $k = 3$
- $k = 5$



k-NN
Exemplo Variando k

Trabalho avaliando diferentes valores de k (base de meses do ano).



Lembrando que tais resultados referem-se a um determinado problema
Outros conjuntos podem apresentar resultados diferentes.

k-NN
Normalização

É importante lembrar da necessidade de normalizarmos os dados, já que podemos estar trabalhando com características que variam bastante, como:

- Peso (40 - 150 kg) ;
- Altura (1,00 - 2,10m) ;
- Salário (550 - 50.000) ;

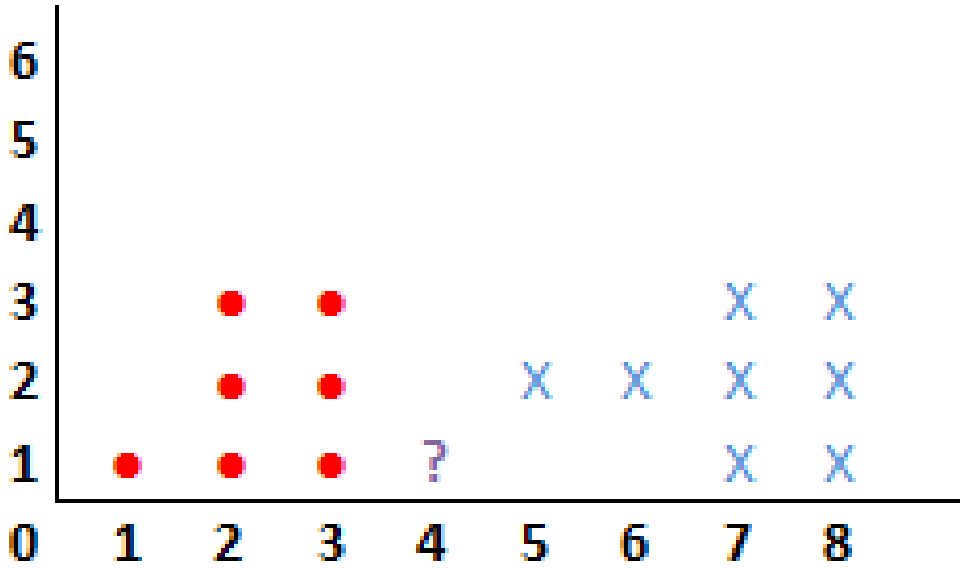
A técnica mais simples consiste em dividir cada característica pelo somatório de todas as características.

k-NN
Vantagens × Desvantagens

- **Vantagens**
  - Técnica simples e de fácil implementação ;
  - Em alguns casos apresenta ótimos resultados ;
  - Pode ser aplicada a problemas complexos, como: Análise de Crédito, Diagnósticos Médicos, Detecção de Fraudes, entre outros.
- **Desvantagens**
  - O fator tempo ;
  - Ruídos nos dados ou características irrelevantes.

k-NN
Exercício Proposto

Considere duas classes representadas a seguir. Usando o k-NN classifique o ponto (?) em uma das duas classes para k = 3 e k = 5. Utilize a Distância Euclidiana como métrica de distância.



k-NN
Exercício Proposto

Elemento **x** (x,y)	Treinamento (x,y)	Fórmula	Distância
(4,1)	(1,1)	$d(x_i, x_j) = \sqrt{(4-1)^2 + (1-1)^2}$	3
(4,1)	(2,1)	$d(x_i, x_j) = \sqrt{(4-2)^2 + (1-1)^2}$	2
(4,1)	(2,2)	$d(x_i, x_j) = \sqrt{(4-2)^2 + (1-2)^2}$	2,23
(4,1)	(2,3)	$d(x_i, x_j) = \sqrt{(4-2)^2 + (1-3)^2}$	2,82
(4,1)	(3,1)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-1)^2}$	1
(4,1)	(3,2)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-2)^2}$	1,41
(4,1)	(3,3)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-3)^2}$	2,23
(4,1)	(5,2)	$d(x_i, x_j) = \sqrt{(4-5)^2 + (1-2)^2}$	1,41
(4,1)	(6,2)	$d(x_i, x_j) = \sqrt{(4-6)^2 + (1-2)^2}$	2,23
(4,1)	(7,1)	$d(x_i, x_j) = \sqrt{(4-7)^2 + (1-1)^2}$	3
(4,1)	(7,2)	$d(x_i, x_j) = \sqrt{(4-7)^2 + (1-2)^2}$	3,16
(4,1)	(7,3)	$d(x_i, x_j) = \sqrt{(4-7)^2 + (1-3)^2}$	3,60
(4,1)	(8,1)	$d(x_i, x_j) = \sqrt{(4-8)^2 + (1-1)^2}$	4
(4,1)	(8,2)	$d(x_i, x_j) = \sqrt{(4-8)^2 + (1-2)^2}$	4,12
(4,1)	(8,3)	$d(x_i, x_j) = \sqrt{(4-8)^2 + (1-3)^2}$	4,24

k-NN
Exercício Proposto

Ordenando:

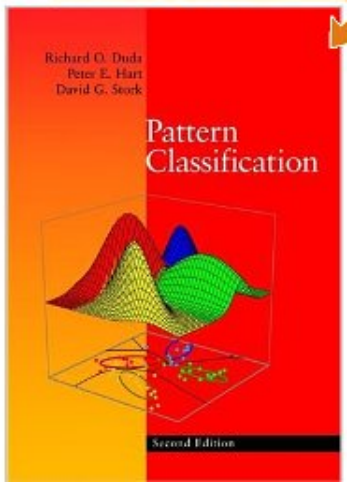
Elemento **x** (x,y)	Treinamento (x,y)	Fórmula	Distância
(4,1)	(3,1)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-1)^2}$	1
(4,1)	(3,2)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-2)^2}$	1,41
(4,1)	(5,2)	$d(x_i, x_j) = \sqrt{(4-5)^2 + (1-2)^2}$	1,41
(4,1)	(2,1)	$d(x_i, x_j) = \sqrt{(4-2)^2 + (1-1)^2}$	2
(4,1)	(3,3)	$d(x_i, x_j) = \sqrt{(4-3)^2 + (1-3)^2}$	2,23

Próximas Aula
Conteúdo

- Aula em Laboratório: Implementação e testes;
- Iremos estudar outro classificador que faz parte da abordagem de Aprendizagem Supervisionada, o classificador **SVM**(*Support Vector Machine*);
- Tal Classificador foi proposto por Vapnik em 1979 é até hoje é largamente utilizado em pesquisas.

Bibliografia
Bibliografia Sugerida

- Duda, R. Hart, P. Stork, D. **Pattern Classification**, John Wiley & Sons, 2000.
- Mitchell, T. **Machine Learning**, McGraw-Hill Science/Engineering/Math, 1997.



Dúvidas?