

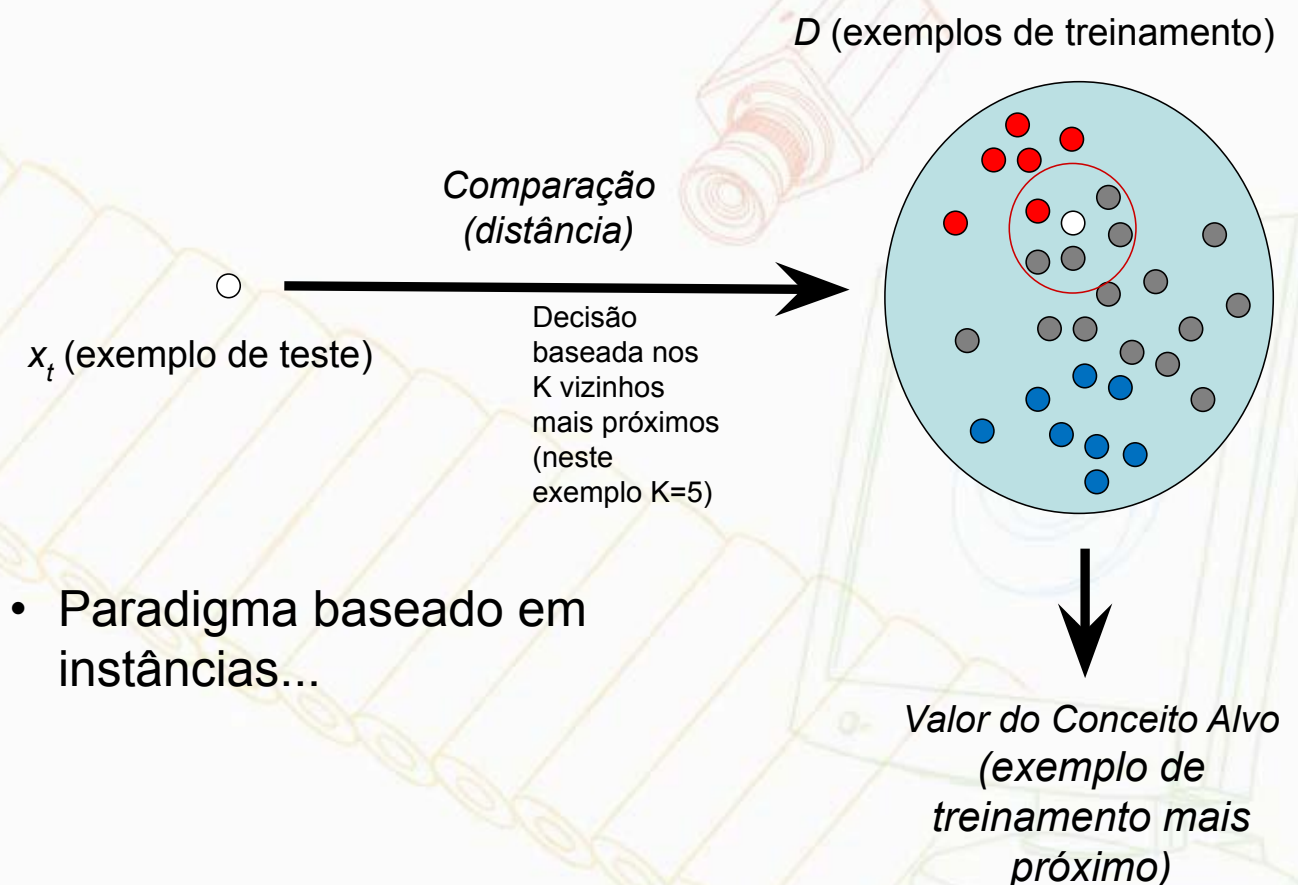
Aprendizagem de Máquina

Alceu S. Britto Jr.

Programa de Pós-Graduação em Informática
Pontifícia Universidade Católica do Paraná (PUCPR)

Aprendizagem Baseada em Instâncias

KNN (K-Nearest Neighbors)



Introdução

- Métodos de aprendizagem baseados em instâncias assumem que as instâncias podem ser representadas como pontos em um espaço Euclidiano.

Introdução

- Os métodos de aprendizagem baseados em instâncias são métodos não paramétricos.
- **Métodos não paramétricos:** podem ser usados com distribuições arbitrárias e sem a suposição de que a forma das densidades são conhecidas.

Aprendizagem Baseada em Instâncias

- A aprendizagem consiste somente em armazenar os exemplos de treinamento $\langle x_1, c_1 \rangle$, $\langle x_2, c_2 \rangle \dots \langle x_n, c_n \rangle$.

vetor de atributos ou características

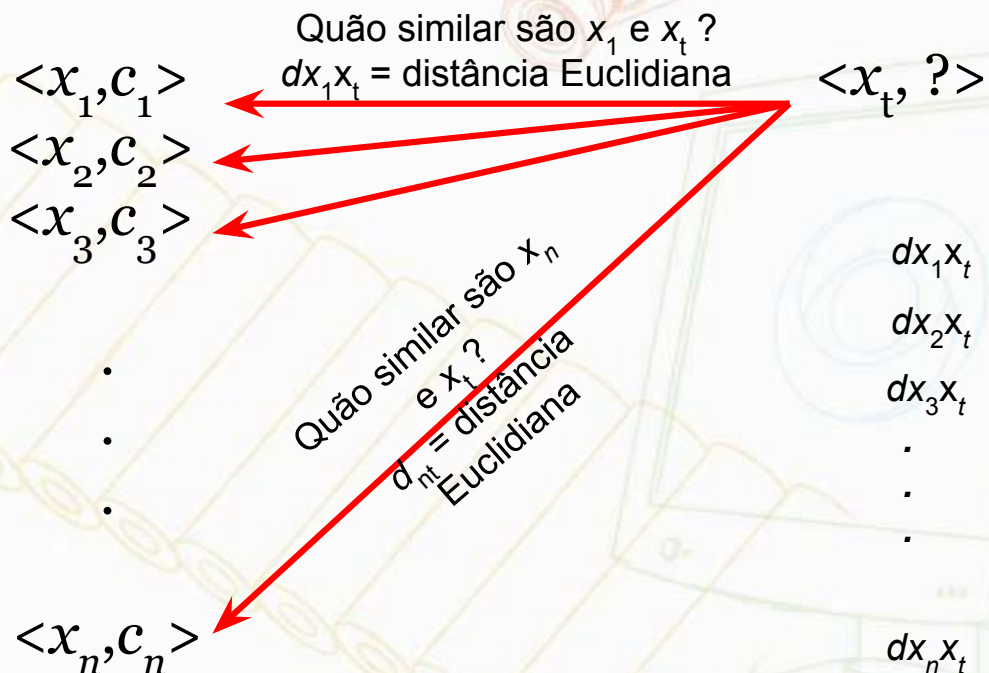
valor do conceito alvo ou classe

- Após a aprendizagem, para encontrar o valor do conceito alvo (classe) associado a uma instância de testes $\langle x_t, ? \rangle$, um conjunto de instâncias similares são buscadas na memória e utilizadas para classificar a nova instância.

Aprendizagem Baseada em Instâncias

Exemplos de Treinamento
<atributos, conceito alvo>

Exemplo a Classificar
<atributos, ????? >



Aprendizagem Baseada em Instâncias

- No final teremos um conjunto de distâncias (medida de similaridade) entre a instância de teste x_t e todas as instâncias de treinamento x_1, x_2, \dots, x_n
- Qual valor de conceito alvo (classe) atribuímos a instância x_t ?

O conceito alvo associado ao exemplo de treinamento mais similar !!

Aprendizagem k -NN

- k -NN = k Nearest Neighbor = k Vizinhos mais Próximos
- O algoritmo k -NN é o método de aprendizagem baseado em instâncias mais elementar.
- O algoritmo k -NN assume que todas as instâncias correspondem a pontos em um espaço n -dimensional.
- Os “vizinhos mais próximos” de uma instância são definidos em termos da distância Euclidiana.

\mathcal{R}^n

Regra k -NN

- A regra dos vizinhos mais próximos:

Meta: Classificar x_t atribuindo a ele o rótulo representado mais frequentemente dentre as k amostras mais próximas e utilizando um esquema de votação.

Aprendizagem k -NN

- Vamos considerar uma instância arbitrária x que é descrita pelo vetor de características:

$$x = \langle a_1(x), a_2(x), \dots, a_n(x) \rangle$$

onde $a_r(x)$ representa o valor do r -ésimo atributo da instância x .



Aprendizagem k -NN

- Então a distância Euclidiana entre duas instâncias x_i e x_j é definida como $d(x_i, x_j)$, onde:

$$d(x_i, x_j) \equiv \sqrt{\sum_{r=1}^n (a_r(x_i) - a_r(x_j))^2}$$

Algoritmo k -NN

Algoritmo de treinamento

- Para cada exemplo de treinamento $\langle x, c \rangle$, adicione o exemplo à lista *training_examples*

Algoritmo de classificação

- Dada uma instância x_t a ser classificada,
 - Faça x_1, \dots, x_k representar as k instâncias de *training_examples* que estão mais próximas de x_t
 - Retorne

$$f(x_t) \leftarrow \arg \max_{c \in C} \sum_{i=1}^k \delta(c, f(x_i))$$

onde $\delta(a, b) = 1$ se $a = b$ e $\delta(a, b) = 0$ caso contrário.

Aprendizagem k -NN

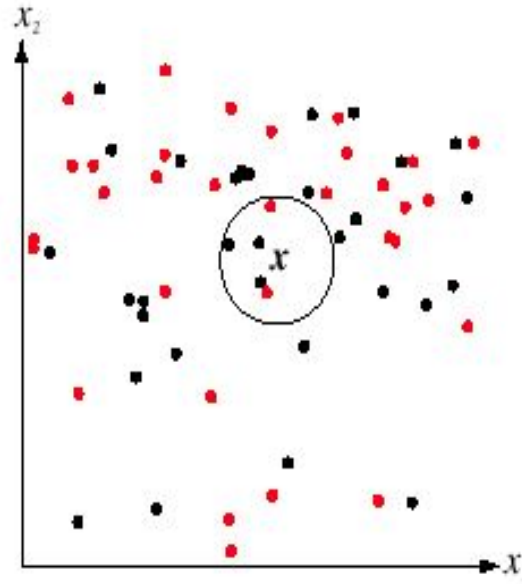
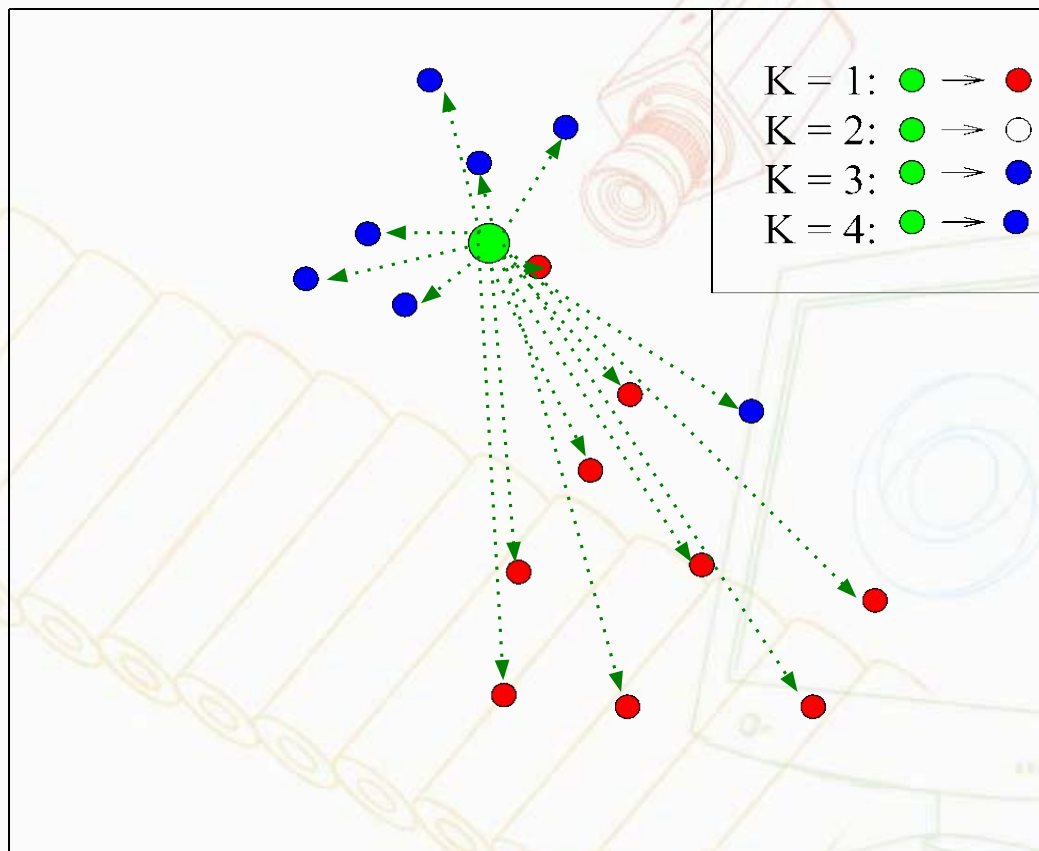
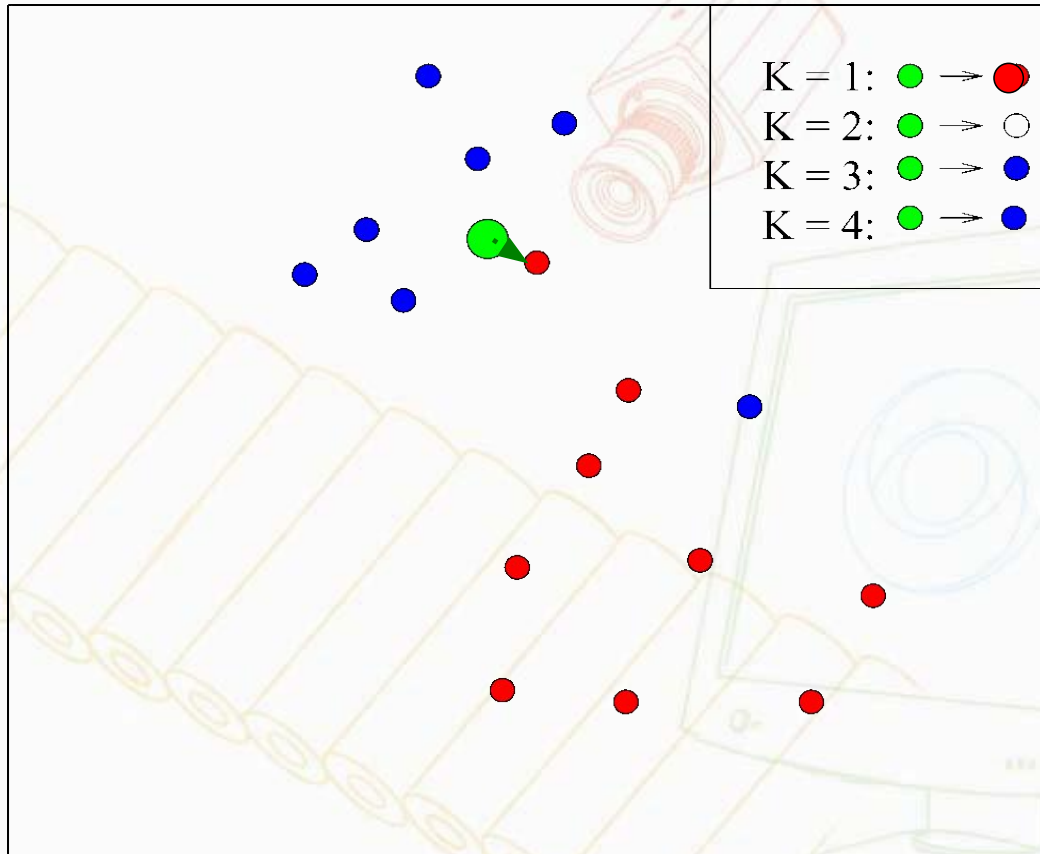


FIGURE 4.15. The k -nearest-neighbor query starts at the test point x and grows a spherical region until it encloses k training samples, and it labels the test point by a majority vote of these samples. In this $k = 5$ case, the test point x would be labeled the category of the black points. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

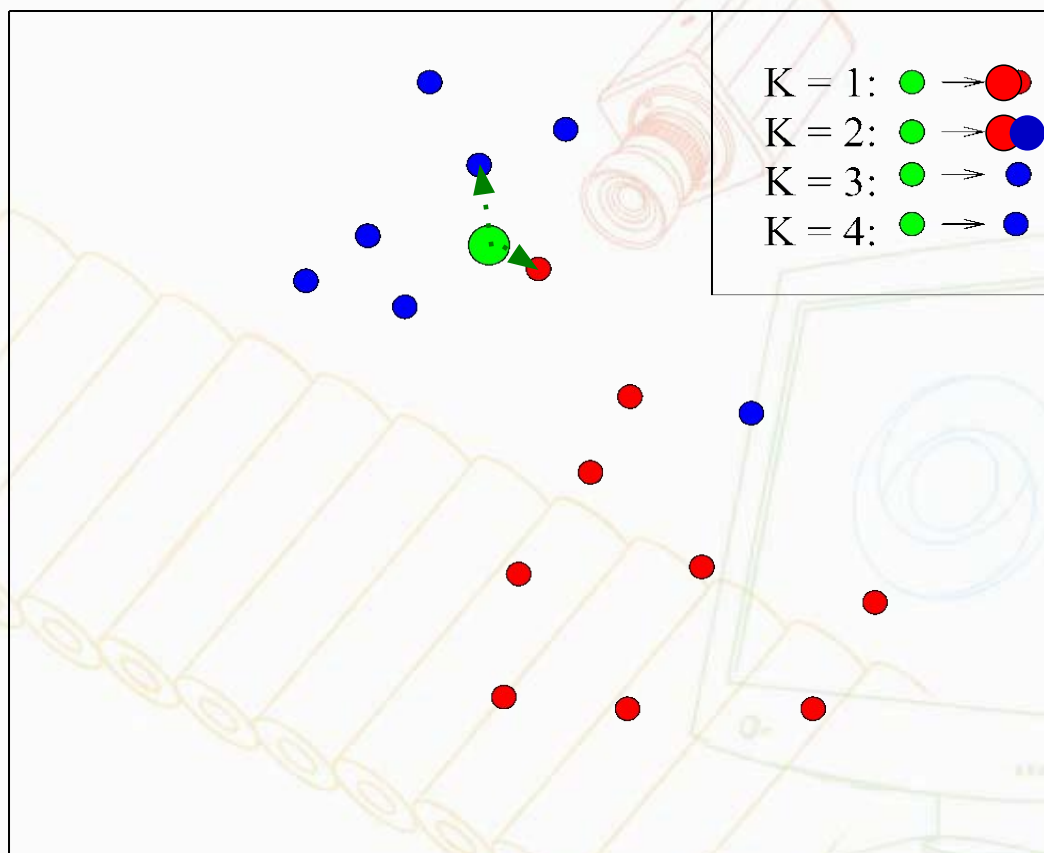
Algoritmo k -NN



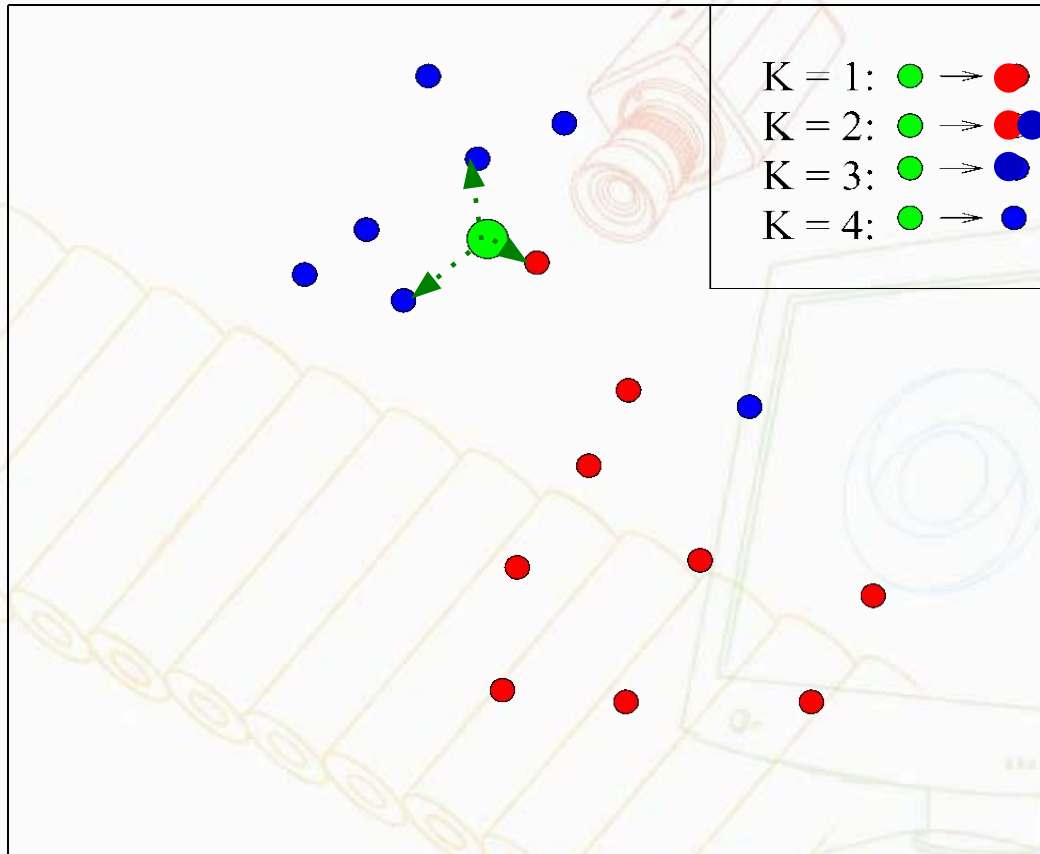
Algoritmo k -NN



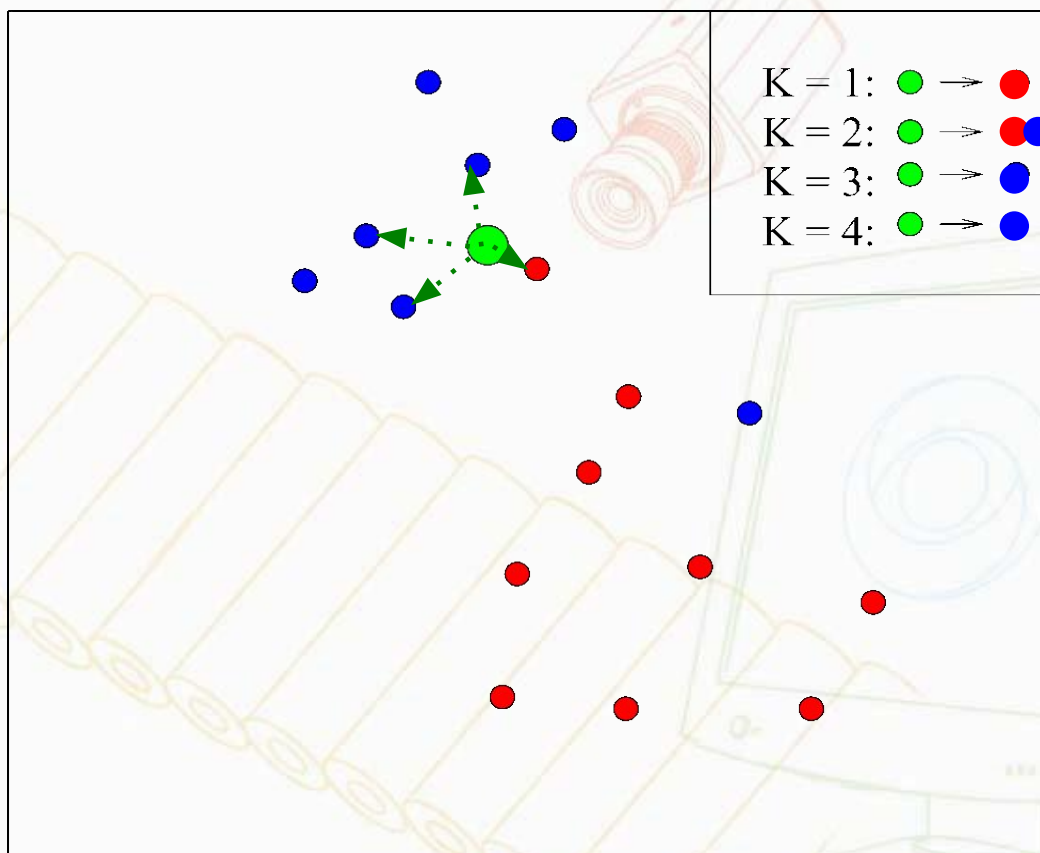
Algoritmo k -NN



Algoritmo k -NN



Algoritmo k -NN



k-NN com Distância Ponderada

- Refinamento do k-NN:
 - Ponderar a contribuição de cada um dos k vizinhos de acordo com suas distâncias até o ponto x_t que queremos classificar, dando maior peso aos vizinhos mais próximos.
- Podemos ponderar o voto de cada vizinho, de acordo com o quadrado do inverso de sua distância de x_t .

$$\hat{f}(x_t) \leftarrow \arg \max_{c \in C} \sum_{i=1}^k w_i \delta(c, f(x_i)) \quad w_i \equiv \frac{1}{d(x_t, x_i)^2}$$

- Porém, se $x_t = x_i$, o denominador $d(x_t, x_i)^2$ torna-se zero. Neste caso fazemos $f(x_t) = f(x_i)$.

Resumo

- Métodos de aprendizagem baseados em instâncias não necessitam formar uma hipótese explícita da função alvo sobre o espaço das instâncias.
- Eles formam uma aproximação local da função alvo para cada nova instância a “classificar”.

Resumo

- O k -NN é um algoritmo baseado em instâncias para aproximar funções alvo de valor real ou de valor discreto, assumindo que as instâncias correspondem a pontos em um espaço d -dimensional.
- O valor da função alvo para um novo ponto é estimada a partir dos valores conhecidos dos k exemplos de treinamento mais próximos.

Resumo

- Vantagens:
 - habilidade para modelar funções alvo complexas por uma coleção de aproximações locais menos complexas.
 - A informação presente nos exemplos de treinamento nunca é perdida.
- Dificuldades:
 - Tempo ?
 - Determinação de uma métrica.