

Corso di Riconoscimento e Recupero  
dell'Informazione per Bioinformatica - A.A.  
2011/2012  
Laboratorio - Esercitazione 6

Dr. Marco Cristani

7 maggio 2012

## 1 K-Nearest Neighbor

Si consideri l'insieme "irisSet", contenente dati riguardanti il seguente problema:

- problema di classificazione della specie degli iris
- 150 elementi (ogni riga del file é un elemento);
- 4 features, corrispondenti alle caratteristiche del fiore (le prime quattro colonne);
- 3 classi, corrispondenti alle tre specie (l'ultima colonna identifica la classe);

Dopo aver caricato l'insieme,

1. Dividerlo in due insiemi casuali di uguale dimensione: uno verrà usato come training set e uno come testing set (mescolare l'insieme prima di dividerlo)
2. Costruire il classificatore K-NN, che dato un elemento del testing set ne calcoli la classe sulla base del training set (utilizzando la metrica euclidea)
3. Calcolare l'errore che il K-NN commette sul testing set per K=1, K=3 e K=5

Ricordo che si ha un errore quando la classe determinata dal KNN è diversa dalla classe effettiva dell'oggetto. L'errore totale è la somma di tutti gli errori. Ricordo anche che la distanza euclidea tra due vettori  $\mathbf{x} = [x_1 \dots x_n]^T$  e  $\mathbf{y} = [y_1 \dots y_n]^T$  è definita come

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

## Primitive di Matlab

Permutazioni:	<i>randperm, permute</i>
Gaussiane:	<i>mean, var, std, cov</i>
Visualizzazione:	<i>plot, scatter, plot3, scatter3, imagesc</i>
Gestione file:	<i>load, save</i>