Corso di Teorie e Tecniche Del Riconoscimento - A.A. 2011/2012

Laboratorio - Esercitazione di ripasso

Dr. Marco Cristani

6 giugno 2012

1 Simulazione di classificazione complessa

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 , contenute nella matrice `x`;
- 3 classi, corrispondenti alle tre specie, il cui vettore di etichette è registrato in `l`;

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. Costruire la matrice di confusione della classificazione
4. Valutare *precision* e *recall* di ogni classe
5. Calcolare l'errore che il K-NN commette sul testing set per K=1, K=3 e K=5

Ripetete i punti qui sopra 100 volte, in modo da ottere una statistica sulle performance più informativa. Ricordo che si ha un errore quando la classe determinata dal KNNè 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 $\mathbf{x} = [x_1...x_n]^T$ e $\mathbf{y} = [y_1...y_n]^T$ è definita come

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

Ripetere l'esercizio pre-processando i dati con PCA (parametrizzate voi l'estrazione delle features). Ripetete la classificazione di cui sopra, e confrontate le performance. Provate a cambiare la proporzione di dati di training e testing, valutando cosa si ottiene con e senza estrazione delle feature.