Insegnamento di "Data mining"
Prova d'esame del 5 aprile 2013

1. Spiegare che cosa si intende per "problema della maledizione della dimensionalità", e come questo si ripercuote sulle modalità di analisi dei dati.

2. Spiegare quale sia l'obiettivo dell'algoritmo delle K-medie, e descrivere in via informale il suo funzionamento.

3. Presentare sinteticamente gli elementi essenziali della formulazione dei modelli additivi.

4. Si spieghi che cosa è un intervallo di confidenza di livello 95% e si presenti come un tale intervallo potrebbe essere scelto.

5. Una parte della matrice di correlazione tra tre variabili (diciamo X_1, X_2, X_3) è la seguente (i puntini indicano elementi non forniti)

$$\begin{bmatrix} \cdot & \cdot & 0.86 \\ -0.34 & \cdot & 0.1 \\ \cdot & \cdot & \cdot \end{bmatrix}$$

   (a) completare la matrice

   (b) sapendo che gli scarti quadratici medi di X_1, X_2, X_3 sono rispettivamente 1, 2 e 3, scrivere la matrice delle varianze e covarianze.

   (c) tra i due modelli di regressione lineare semplice

$$A : X_1 = \alpha + \beta X_2 + \varepsilon \qquad \text{e} \qquad B : X_1 = \alpha + \beta X_3 + \varepsilon$$

   quale usereste per calcolare delle previsioni per X_1?

   (d) Disegnare degli ipotetici diagrammi di dispersione per tutte e tre le coppie di variabili (ovvero X_1 rispetto a X_2, X_1 rispetto a X_3, e X_2 rispetto a X_3) compatibili con tutte le informazioni date.

6. (facoltativo)

   Si considerino due variabili X e Y osservate su n unità e il modello lineare che le lega: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon$ con $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$, per $i =, 1, \ldots,$. Siano $r_1, r_2, \ldots, r_n$ i residui del modello lineare. Si dimostri che la varianza dei residui è sempre non più grande della varianza delle variabile risposta Y.