Data Mining e Scoperta di Conoscenza

Progetto 4

Classificazione Bayesiana e spam detection

1. Si individui un dataset di messaggi di posta elettronica che contenga messaggi spam (ad esempio, si possono utilizzare i datasets riferiti da http://www.kdnet.org/kdnet/control/dataset_details?item_id=40. http://www.kdnet.org/kdnet/control/dataset_details?item_id=39 e http://listserv.linguitlist.org/archives).
2. Si estenda Weka con la classe **`weka.core.MessageInstance.java`**, che implementa la lettura di un insieme di messaggi di posta elettronica, ne estrae le features, e le rappresenta in formato tabellare. In particolare, la classe deve implementare:
    a. lemmatizzazione
    b. riconoscimento e rimozione di stopwords
    c. estrazione di ulteriori features (ad esempio dominio del mittente, finestra temporale di ricezione, ecc.).
    d. rappresentazione del messaggio come istanza in Weka, utilizzando due modelli differenti:
        i. rappresentazione binaria
        ii. rappresentazione per frequenza (TF/IDF).
3. Si apprenda un classificatore bayesiano che implementa il riconoscitore dello spam. In pratica, il riconoscitore prende in input un messaggio e, dopo averlo trasformato nel formato desiderato, lo classifica come soggetto spam/non spam. (OPZIONALE: il classificatore può leggere direttamente da una casella reale utilizzando pop/imap. Si consulti in proposito la documentazione SUN JavaMail).
4. Si studi la performance della classificazione con particolare riferimento alle curve di ROC:
    a. Come variano le performance del classificatore al variare dell'insieme di addestramento?
    b. Come varia l'accuratezza
    c. Si studi l'effetto dell'analisi delle componenti principali sull'errore
    d. Si determinino delle trasformazioni dei dati che permettano un miglioramento generale delle performances del classificatore.
5. Si determini la corretta topologia di una rete bayesiana che permetta di migliorare le performances della classificazione rispetto al normale classificazione Naive Bayes.

NOTE PER L'ESECUZIONE DEL PROGETTO

1. Scrivi un rapporto di circa 10 pagine in cui
    a. Descrivi analiticamente l'algoritmo che hai utilizzato e la tecnica che hai usato per la valutazione.
    b. Commenti le parti essenziali del codice Java che hai scritto, e metti in un'appendice l'intero codice
    c. commenti e illustri graficamente e quantitativamente gli esperimenti effettuati.
2. Prepara delle slides Powerpoint (non più di 10 slides) in cui riassumi gli esiti del progetto