Progetto 2006-2007

AntiSpam

Descrizione:

Raccogliete un congruo numero di e-mail spam ed e-mail non spam (almeno 150 e 150).
Suddividete il campione in k blocchi di uguale dimensione (k>=3) per effetuare una k-fold cross evaluation.

Un primo compito consiste nel decidere come rappresentare i messaggi: quali "features" scegliere, come rappresentarle. Non vi dò suggerimenti, perchè questa è la parte più "creativa" dell'esercitazione.
Quidni, selezionate due algoritmi di apprendimento, fra quelli studiati a lezione, e disponibili sul sito WEKA.
Addestrate l'algoritmo con l'insieme dei dati del LS, e testatelo sul TS, per k volte, modificando LS e TS secondo il procedimento della k-fold cross evaluation.

Il sito WEKA vi fornisce la possibilità di diagrammare i risultati.

Producete un rapporto di una decina di pagine, che consegnerete qualche giorno prima della verbalizzazione.

Il progetto può essere svolto da due studenti (ma il vincolo è che alla verbalizzazione ci si presenti in due).