Analisi dei dati con R

Prova d'esame di Data Mining del 25.6.2020

Problema di classificazione

Si tratta di un dataset reale, ma le variabili sono state opportunamente anonimizzate.

- $Y \in \{Yes, No\}$ è la variabile risposta binaria
- $X = (X_1, \ldots, X_7)'$ sono $p = 7$ predittori
- Training set: (y_i, x_i) per $i = 1, \ldots, n$ con $n = 200$
- Test set: (y_i^*, x_i^*) per $i = 1, \ldots, m$ con $m = 332$

L'obiettivo è di prevedere la variabile risposta $y_1^*, \ldots, y_m^*$ per il test set.

La metrica di valutazione è l'accuratezza:

$$\text{Acc}_{\text{Te}} = \frac{1}{m} \sum_{i=1}^{m} 1(y_i^* = \hat{y}_i^*)^2$$

La percentuale di punti verrà calcolata come $\min\left(\frac{x - 0.7}{0.8 - 0.7}, 100\%\right)$ dove x è il punteggio di Acc_{Te} ottenuto.

Il punteggio con il quale confrontarsi $\text{Acc}_{\text{Te}} = 70.48\%$ è stato ottenuto dal seguente modello

```
library(kknn)
yhat<-kknn(y~., tr, te, k=1)$fitted.values
head(yhat)
# name the .txt file with your badge number, e.g. 2575.txt
write.table(file="2575.txt", yhat, row.names = F, col.names = F)
```

Regole

Riceverete training e test set (file `trte.RData`), insieme ad un template (file `2575.Rmd`) da completare con il codice R utilizzato.

Entro **70 MINUTI** bisogna:

1. Inviare il file **[BADGE].txt** contenente le proprie previsioni
2. Inviare il file **[BADGE].html** (generato da R Markdown) contenente il codice R utilizzato per le previsioni

Altri formati non verranno accettati.