Esercizio: costruire un Naive Bayes classifier per classificare documenti in argomenti.
Come features si considerano la presenza o meno di parole di un vocabolario nel documento da classificare.

Vocabolario: (baseball calcio Moratti riforma sport studio)

La funzione di classificazione f(x) assume due valori: sport, scuola.
Le x_i (i documenti) sono rappresentati mediante vettori a 6 valori binari. Il valore "1" della feature x_{ij} indica che il termine j-esimo (j=1, 2, ..., 6) del vocabolario è presente nel documento i-esimo.

Insieme di addestramento D:
x_1 = ((0,0,1,0,1,0), sport)
x_2 = ((0,0,1,0,0,1), scuola)
x_3 = ((1,1,1,0,0,0), sport)
x_4 = ((1,0,0,0,1,0), sport)
x_5 = ((0,0,0,1,0,1), scuola)
x_6 = ((0,0,0,1,0,1), scuola)

Si vuole assegnare una classe all'istanza x = (1,0,1,0,0,0) sulla base di un Naive Bayes classifier:

$$f(x) = \arg\max_{c_i \in C} P(c_i) \prod P(val_j \mid c_i)$$

f(x): assume i valori in C = {sport, scuola}
a_j (valore del j-esimo attributo) = 0,1 (tutti gli attributi sono binari)

Nell'esempio x da classificare ho i seguenti valori per i 6 attributi:
baseball=1 calcio=0 Moratti=1 riforma=0 sport=0 studio=0

Le (stime delle) probabilità a priori delle due classi sono uguali, cioè P(sport)=P(scuola)=3/6=1/2
Devo stimare sull'insieme di addestramento D le seguenti probabilità condizionate:

P(baseball=1|sport) P(calcio=0|sport) P(Moratti=1|sport)
P(riforma=0|sport) P(sport=0|sport) P(studio=0|sport)

P(baseball=1|scuola) P(calcio=0|scuola) P(Moratti=1|scuola) P(riforma=0|scuola)
P(sport=0|scuola) P(studio=0|scuola)

avrò le seguenti stime per le probabilità condizionate (senza m-stime):

P(baseball=1|sport)=2/3 P(calcio=0|sport)=2/3 P(Moratti=1|sport)=2/3
P(riforma=0 |sport)=3/3 P(sport=0|sport)=1/3 P(studio=0|sport)=3/3

P(baseball=1|scuola)=1/3 P(calcio=0|scuola)=1/3 P(Moratti=1|scuola)=1/3
P(riforma=0|scuola)=0 P(sport=0|scuola)=2/3 P(studio=0|scuola)=0

Dunque,

$P(sport)\prod P(a_j|sport)=1/2*2/3*2/3*2/3*1*1/3*1=4/81 = 0,05$
$P(scuola)\prod P(a_j|scuola)=1/2*1/3*1/3*1/3*0*2/3*0=0$

quindi l'argmax fornisce c_{NB}=sport