Exemple de méthodes de classification/catégorisation de textes

- Le classifieur Bayésien naïf (Naive Bayes Classifier)
  - principe général du classifieur Bayésien= à classer, choisir la classe c qui maximise P(c | o)
    - étant donné une observation o
      - par exemple ici o = le document

$$\hat{c} = \arg\max_{c} P(c \mid o)$$

    - Loi de Bayes, et le fait que P(o) est constant pour toute classe, on obtient :

$$\hat{c} = \arg\max_{c} P(c \mid o) = \arg\max_{c} \frac{P(o \mid c)P(c)}{P(o)} = \arg\max_{c} P(o \mid c)P(c)$$

Exemple de méthodes de classification/catégorisation de textes

- $$\hat{c} = \arg\max_c P(c \mid o) = \arg\max_c \frac{P(o \mid c)P(c)}{P(o)} = \arg\max_c P(o \mid c)P(c)$$
- Le classifieur Bayésien naïf (suite)
  - Naïf : hypothèse d'indépendance forte entre les caractéristiques de l'observation
    - o = doc et (m1, …., mN) les mots du document o
    - P(o/C) = P(m1,…,mN/C) $= \prod_{i=1}^{N} P(mi/C)$ -> passer en log

$$\hat{c} = \arg\max_{c \in \mathbb{R}} [log(P(c)) + \sum_{i=1}^{N} log(P(m_i/c))]$$

  - Apprentissage sur un ensemble de documents
    - Estimation de p(c) et de p(mi/c)
      - P(c) = nombre de docs dans la classe C/nombre total de docs
      - P(mi/c) = fréquence du mot i dans la classe C

TP Sentiment analysis avec NB

TRAINMULTINOMIALNB($\mathbb{C}, \mathbb{D}$)
1 $V \leftarrow$ EXTRACTVOCABULARY($\mathbb{D}$)
2 $N \leftarrow$ COUNTDOCS($\mathbb{D}$)
3 **for each** $c \in \mathbb{C}$
4 **do** $N_c \leftarrow$ COUNTDOCSINCLASS($\mathbb{D}, c$)
5 $prior[c] \leftarrow N_c/N$
6 $text_c \leftarrow$ CONCATENATETEXTOFALLDOCSINCLASS($\mathbb{D}, c$)
7 **for each** $t \in V$
8 **do** $T_{ct} \leftarrow$ COUNTTOKENSOFTERM($text_c, t$)
9 **for each** $t \in V$
10 **do** $condprob[t][c] \leftarrow \frac{T_{ct}+1}{\sum_{t'}(T_{ct'}+1)}$
11 **return** $V, prior, condprob$

APPLYMULTINOMIALNB($\mathbb{C}, V, prior, condprob, d$)
1 $W \leftarrow$ EXTRACTTOKENSFROMDOC(V, d)
2 **for each** $c \in \mathbb{C}$
3 **do** $score[c] \leftarrow \log prior[c]$
4 **for each** $t \in W$
5 **do** $score[c] \mathrel{+}= \log condprob[t][c]$
6 **return** $\arg\max_{c \in \mathbb{C}} score[c]$

▶ **Figure 13.2** Naive Bayes algorithm (multinomial model): Training and testing.