

## 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

$$\hat{c} = \operatorname{argmax}_c P(c | o)$$

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

$$\hat{c} = \operatorname{argmax}_c P(c | o) = \operatorname{argmax}_c \frac{P(o | c)P(c)}{P(o)} = \operatorname{argmax}_c P(o | c)P(c)$$

## Exemple de méthodes de classification/catégorisation de textes

- $$\hat{c} = \arg \max_c P(c | o) = \arg \max_c \frac{P(o | c)P(c)}{P(o)} = \arg \max_c P(o | c)P(c)$$

- Le classifieur Bayésien naïf (suite)

- Naïf : hypothèse d'indépendance forte entre les caractéristiques de l'observation

- ✦  $o = \text{doc}$  et  $(m_1, \dots, m_N)$  les mots du document  $o$
    - ✦  $P(o/C) = P(m_1, \dots, m_N/C) = \prod_{i=1}^N P(m_i/C) \rightarrow$  passer en log

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

- Apprentissage sur un ensemble de documents

- ✦ Estimation de  $p(c)$  et de  $p(m_i/c)$ 
      - $P(c)$  = nombre de docs dans la classe  $C$  / nombre total de docs
      - $P(m_i/c)$  = fréquence du mot  $i$  dans la classe  $C$

# TP Sentiment analysis avec NB

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

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

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