# Lecture 16: Naïve Bayes Text Classification

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
August–December 2020

# Text classification

- Classify text documents using topics
- Useful for automatic segregation of newsfeeds, other internet content
- Training data has a unique topic label per document — e.g., Sports, Politics, Entertainment
- Want to use a naïve Bayes classifier
- Need to define a generative model
- How do we represent documents?

# Set of words model

- Each document is a set of words over a vocabulary $V = \{w_1, w_2, \dots, w_m\}$
- Topics come from a set $C = \{c_1, c_2, \dots, c_k\}$
- Each topic $c$ has probability $Pr(c)$
- Each word $w_i \in V$ has conditional probability $Pr(w_i \mid c_j)$ with respect to each $c_j \in C$
- Generating a random document $d$
  - Choose a topic $c$ with probability $Pr(c)$
  - For each $w \in V$, toss a coin, include $w$ in $d$ with probability $Pr(w \mid c)$
- $Pr(d \mid c) = \prod_{w_i \in D} Pr(w_i \mid c) \prod_{w_i \notin D} (1 - Pr(w_i \mid c))$
- $Pr(d) = \sum_{c \in C} Pr(d \mid c)$

# Naïve Bayes classifier

- Training set $D = \{d_1, d_2, \ldots, d_n\}$
  - Each $d_i \subseteq V$ is assigned a unique label from $C$
- $Pr(c_j)$ is fraction of $D$ labelled $c_j$
- $Pr(w_i \mid c_j)$ is fraction of documents labelled $c_j$ in which $w_i$ appears
- Given a new document $d \subseteq V$, we want to compute $\arg\max_c Pr(c \mid d)$
- By Bayes' rule, $Pr(c \mid d) = \dfrac{Pr(d \mid c)Pr(c)}{Pr(d)}$
  - As usual, discard the common denominator and compute $\arg\max_c Pr(d \mid c)Pr(c)$
- Recall $Pr(d \mid c) = \prod_{w_i \in D} Pr(w_i \mid c) \prod_{w_i \notin D} (1 - Pr(w_i \mid c))$

# Bag of words model

- Each document is a multiset or bag of words over a vocabulary $V = \{w_1, w_2, \ldots, w_m\}$
  - Count multiplicities of each word
- As before
  - Each topic $c$ has probability $Pr(c)$
  - Each word $w_i \in V$ has conditional probability $Pr(w_i \mid c_j)$ with respect to each $c_j \in C$
  - Note that $\displaystyle\sum_{i=1}^{m} Pr(w_i \mid c_j) = 1$
  - Assume document length is independent of the class

# Bag of words model

- Generating a random document $d$
  - Choose a document length $\ell$ with $Pr(\ell)$
  - Choose a topic $c$ with probability $Pr(c)$
  - Recall $|V| = m$.
    - To generate a single word, throw an $m$-sided die that displays $w$ with probability $Pr(w \mid c)$
    - Repeat $\ell$ times
- Let $n_j$ be the number of occurrences of $w_j$ in $d$
- $Pr(d \mid c) = Pr(\ell)\ \ell! \displaystyle\prod_{j=1}^{m} \frac{Pr(w_j \mid c)^{n_j}}{n_j!}$

# Parameter estimation

- Training set $D = \{d_1, d_2, \ldots, d_n\}$
  - Each $d_i$ is a multiset over $V$ of size $\ell_i$
- As before, $Pr(c_j)$ is fraction of $D$ labelled $c_j$
- $Pr(w_i \mid c_j)$ — fraction of occurrences of $w_i$ over documents $D_j \subseteq D$ labelled $c_j$
  - $n_{id}$ — occurrences of $w_i$ in $d$
  - $Pr(w_i \mid c_j) = \dfrac{\displaystyle\sum_{d \in D_j} n_{id}}{\displaystyle\sum_{t=1}^{m} \sum_{d \in D_j} n_{td}} = \dfrac{\displaystyle\sum_{d \in D} n_{id}\ Pr(c_j \mid d)}{\displaystyle\sum_{t=1}^{m} \sum_{d \in D} n_{td}\ Pr(c_j \mid d)},$

    since $Pr(c_j \mid d) = \begin{cases} 1 & \text{if } d \in D_j, \\ 0 & \text{otherwise} \end{cases}$

# Classification

- $Pr(c \mid d) = \dfrac{Pr(d \mid c)\ Pr(c)}{Pr(d)}$
- Want $\arg\max_c\ Pr(c \mid d)$
- As before, discard the denominator $Pr(d)$
- Recall, $Pr(d \mid c) = Pr(\ell)\ \ell!\ \displaystyle\prod_{j=1}^{m} \frac{Pr(w_j \mid c)^{n_j}}{n_j!}$, where $|d| = \ell$
- Discard $Pr(\ell), \ell!$ since they do not depend on $c$
- Compute $\arg\max_c\ Pr(c)\ \displaystyle\prod_{j=1}^{m} \frac{Pr(w_j \mid c)^{n_j}}{n_j!}$

# Summary

- We can use naïve Bayes classifiers to assign topics to documents
- Need to define a suitable probabilistic model for generating random documents
- Set of words — each document $d$ is a subset of the vocabulary $V$
- Bag of words — each document $d$ is a multiset of the vocabulary $V$
- In the bag of words model, we assume that document length is independent of topic