Naïve Bayes Classification

Debapriyo Majumdar
Data Mining – Fall 2014
Indian Statistical Institute Kolkata

August 14, 2014
Bayes’ Theorem

- Thomas Bayes (1701-1761)
- Simple form of Bayes’ Theorem, for two random variables $C$ and $X$

$$P(C \mid X) = \frac{P(X \mid C) \times P(C)}{P(X)}$$

**Likelihood**

**Posterior probability**

**Class prior probability**

**Predictor prior probability or evidence**

$X$ and $C$
Probability Model

- Probability model: for a *target class variable C* which is dependent over features $X_1, \ldots, X_n$

$$P(C \mid X_1, \ldots, X_n) = \frac{P(C) \times P(X_1, \ldots, X_n \mid C)}{P(X_1, \ldots, X_n)}$$

- So the denominator is effectively constant
- Goal: calculating probabilities for the possible values of $C$
- We are interested in the numerator:

$$P(C) \times P(X_1, \ldots, X_n \mid C)$$
The conditional probability is equivalent to the joint probability

\[ P(C) \times P(X_1, \ldots, X_n \mid C) = P(C, X_1, \ldots, X_n) \]

Applying the chain rule for joint probability

\[
P(C) \times P(X_1, \ldots, X_n \mid C) \\
= P(C) \times P(X_1 \mid C) \times P(X_2, \ldots, X_n \mid C, X_1) \\
= P(C) \times P(X_1 \mid C) \times P(X_2 \mid C, X_1) P(X_3, \ldots, X_n \mid C, X_1, X_2) \\
\ldots \\
\ldots \\
= P(C) \times P(X_1 \mid C) \times P(X_2 \mid C, X_1) \ldots P(X_n \mid C, X_1, \ldots, X_{n-1})
\]

\[ P(A, B) = P(A) \times P(B \mid A) \]
Strong Independence Assumption (Naïve)

- Assume the features $X_1, \ldots, X_n$ are conditionally independent given $C$
  - Given $C$, occurrence of $X_i$ does not influence the occurrence of $X_j$, for $i \neq j$.

$$P(X_i | C, X_j) = P(X_i | C)$$

- Similarly,

$$P(X_i | C, X_j, \ldots, X_k) = P(X_i | C)$$

- Hence:

$$P(C) \times P(X_1 | C) \times P(X_2 | C, X_1) \cdots \times P(X_n | C, X_1, \ldots, X_{n-1})$$

$$= P(C) \times P(X_1 | C) \times P(X_2 | C) \cdots \times P(X_n | C)$$
Naïve Bayes Probability Model

\[
P(C \mid X_1, \ldots, X_n) = \frac{P(C) \times P(X_1 \mid C) \times P(X_2 \mid C) \times \cdots \times P(X_n \mid C)}{P(X_1, \ldots, X_n)}
\]

- Class posterior probability
- Known values: constant
Classifier based on Naïve Bayes

- Decision rule: pick the hypothesis (value $c$ of $C$) that has highest probability
  - Maximum A-Posteriori (MAP) decision rule

$$\arg\max_{c} \left\{ P(C = c) \prod_{i=1}^{n} P(X_i = x_i \mid C = c) \right\}$$

Approximated from frequency in the training set

Approximated from relative frequencies in the training set

The values of features are known for the new observation
Example of Naïve Bayes
Reference: The IR Book by Raghavan et al, Chapter 6

Text Classification with Naïve Bayes
The Text Classification Problem

- Set of class labels / tags / categories: $C$
- Training set: set $D$ of documents with labels $<d, c> \in D \times C$
- Example: a document, and a class label $<\text{Beijing joins the World Trade Organization, China}>$
- Set of all terms: $V$
- Given $d \in D'$, a set of new documents, the goal is to find a class of the document $c(d)$
Multinomial Naïve Bayes Model

- Probability of a document \( d \) being in class \( c \)

\[
P(c \mid d) \propto P(c) \times \prod_{k=1}^{n_d} P(t_k \mid c)
\]

where \( P(t_k \mid c) \) = probability of a term \( t_k \) occurring in a document of class \( c \)

Intuitively:
- \( P(t_k \mid c) \) ~ how much evidence \( t_k \) contributes to the class \( c \)
- \( P(c) \) ~ prior probability of a document being labeled by class \( c \)
Multinomial Naïve Bayes Model

- The expression

\[ P(c) \times \prod_{k=1}^{n_d} P(t_k \mid c) \]

has many probabilities.

- May result a floating point underflow.
  - Add logarithms of the probabilities instead of multiplying the original probability values

\[
\begin{align*}
    c_{map} &= \arg\max_{c \in C} \left[ \log P(c) + \sum_{k=1}^{n_d} \log P(t_k \mid c) \right] \\
    \text{Term weight of } t_k \text{ in class } c
\end{align*}
\]

Todo: estimating these probabilities
Maximum Likelihood Estimate

- Based on relative frequencies
- Class prior probability
  \[ P(c) = \frac{N_c}{N} \]

- Estimate \( P(t|c) \) as the relative frequency of \( t \) occurring in documents labeled as class \( c \)
  \[ P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}} \]
Handling Rare Events

- What if: a term $t$ did not occur in documents belonging to class $c$ in the training data?
  - Quite common. Terms are sparse in documents.
- Problem: $P(t|c)$ becomes zero, the whole expression becomes zero
- Use add-one or Laplace smoothing

$$P(t|c) = \frac{T_{ct} + 1}{\sum_{t'|\in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\left(\sum_{t'|\in V} T_{ct'}\right) + |V|}$$
Example
Bernoulli Naïve Bayes Model

- Binary indicator of occurrence instead of frequency
- Estimate $P(t|c)$ as the fraction of documents in $c$ containing the term $t$
- Models absence of terms explicitly:

$$P(C \mid X_1,\ldots,X_n) = \prod_{i=1}^{n} [X_i P(t_i \mid C) + (1 - X_i)(1 - P(t_i \mid C))]$$

$X_i = 1$ if $t_i$ is present, 0 otherwise.

Absence of terms

Difference between Multinomial with frequencies truncated to 1, and Bernoulli Naïve Bayes?
Example

- India
  - Indian
  - Delhi
  - Indian

- India
  - Indian
  - Taj Mahal

- India
  - Indian
  - Goa

- UK
  - London
  - Indian
  - Embassy

- Indian
  - Embassy
  - London
  - Indian
  - Indian

Classify
Naïve Bayes as a Generative Model

- The probability model:

\[ P(c \mid d) = \frac{P(c) \times P(d \mid c)}{P(d)} \]

Multinomial model

\[ P(d \mid c) = P(\langle t_1, \ldots, t_{n_d} \rangle \mid c) \]

Terms as they occur in \( d \), exclude other terms

where \( X_i \) is the random variable for position \( i \) in the document
- Takes values as terms of the vocabulary

- Positional independence assumption \( \rightarrow \) Bag of words model:

\[ P(X_{k_1} = t \mid c) = P(X_{k_2} = t \mid c) \]
Naïve Bayes as a Generative Model

- The probability model:

\[ P(c \mid d) = \frac{P(c) \times P(d \mid c)}{P(d)} \]

Bernoulli model

\[ P(d \mid c) = P(\langle e_1, \ldots, e_{|V|} \rangle \mid c) \]

All terms in the vocabulary

\[ P(U_i = 1 \mid c) \] is the probability that term \( t_i \) will occur in any position in a document of class \( c \)
## Multinomial vs Bernoulli

<table>
<thead>
<tr>
<th></th>
<th>Multinomial</th>
<th>Bernoulli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event Model</td>
<td>Generation of token</td>
<td>Generation of document</td>
</tr>
<tr>
<td>Multiple occurrences</td>
<td>Matters</td>
<td>Does not matter</td>
</tr>
<tr>
<td>Length of documents</td>
<td>Better for larger documents</td>
<td>Better for shorter documents</td>
</tr>
<tr>
<td>#Features</td>
<td>Can handle more</td>
<td>Works best with fewer</td>
</tr>
</tbody>
</table>
On Naïve Bayes

- **Text classification**
  - Spam filtering (email classification) [Sahami et al. 1998]
  - Adapted by many commercial spam filters
  - SpamAssassin, SpamBayes, CRM114, …

- **Simple: the conditional independence assumption is very strong (naïve)**
  - Naïve Bayes may not estimate right in many cases, but ends up classifying correctly quite often
  - It is difficult to understand the dependencies between features in real problems