# 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 $X$

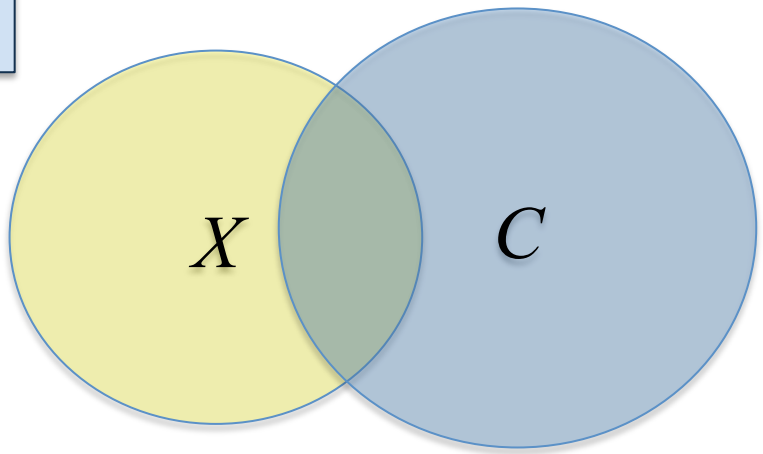
Likelihood

Class prior probability

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

Posterior probability

Predictor prior probability or evidence



# Probability Model

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

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

The values of the features are given

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

$$P(C) \times P(X_1, \ldots, X_n \mid C)$$

# Probability Model

- The conditional probability is equivalent to the joint probability

$$P(C) \times P(X_1, ..., X_n \mid C) = P(C, X_1, ..., X_n)$$

- Applying the chain rule for join probability

$$P(A, B) = P(A) \times P(B \mid A)$$

$$\begin{aligned}
& P(C) \times P(X_1, ..., X_n \mid C) \\
&= P(C) \times P(X_1 \mid C) \times P(X_2, ..., X_n \mid C, X_1) \\
&= P(C) \times P(X_1 \mid C) \times P(X_2 \mid C, X_1) P(X_3, ..., X_n \mid C, X_1, X_2) \\
&\quad \dots \\
&\quad \dots \\
&= P(C) \times P(X_1 \mid C) \times P(X_2 \mid C, X_1) ..... P(X_n \mid C, X_1, ..., X_{n-1})
\end{aligned}$$

# 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 $i \neq j$.

$$P(X_i \mid C, X_j) = P(X_i \mid C)$$

- Similarly,

$$P(X_i \mid C, X_j, ..., X_k) = P(X_i \mid C)$$

- Hence:

$$P(C) \times P(X_1 \mid C) \times P(X_2 \mid C, X_1) ..... P(X_n \mid C, X_1, ..., X_{n-1})$$

$$= P(C) \times P(X_1 \mid C) \times P(X_2 \mid C) ..... P(X_n \mid C)$$

# 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

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

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

  $$\langle d,c \rangle \in D \times C$$

- *Example:* a document, and a class label

  <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 $P(t_k|c)$ = probability of a term $t_k$ occurring in a document of class $c$

Intuitively:

- $P(t_k|c) \sim$ how much evidence $t_k$ contributes to the class $c$
- $P(c) \sim$ 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

$$c_{map} = \underset{c \in C}{\operatorname{argmax}} \left[ \log P(c) + \sum_{k=1}^{n_d} \log P(t_k \mid c) \right]$$

Term weight of $t_k$ in class $c$

Todo: estimating these probabilities

# Maximum Likelihood Estimate

- Based on relative frequencies
- Class prior probability

$$P(c) = \frac{N_c}{N}$$

#of documents labeled as class *c* in training data

total #of documents in training data

- Estimate $P(t|c)$ as the relative frequency of $t$ occurring in documents labeled as class $c$

$$P(t \mid c) = \frac{T_{ct}}{\sum_{t' \in V} T_{ct'}}$$

total #of occurrences of t in documents $d \in c$

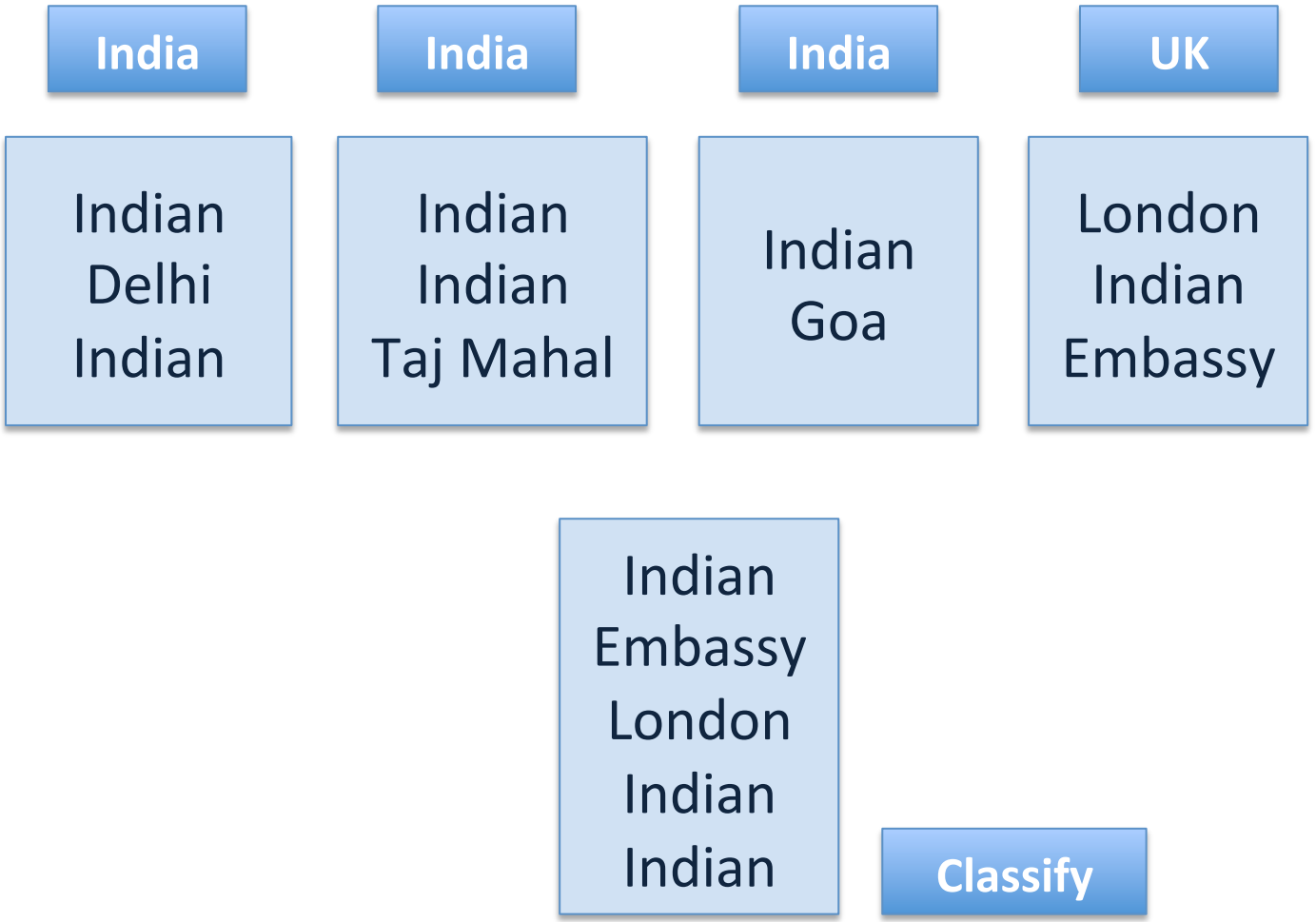
total #of occurrences of all terms in documents $d \in c$

# 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*

$$P(t \mid c) = \frac{T_{ct} + 1}{\sum_{t' \in V} (T_{ct'} + 1)} = \frac{T_{ct} + 1}{\left( \sum_{t' \in V} T_{ct'} \right) + |V|}$$

# Example


India
India
India
UK
Indian
Delhi
Indian
Indian
Indian
Taj Mahal
Indian
Goa
London
Indian
Embassy
Indian
Embassy
London
Indian
Indian
Classify

# 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:

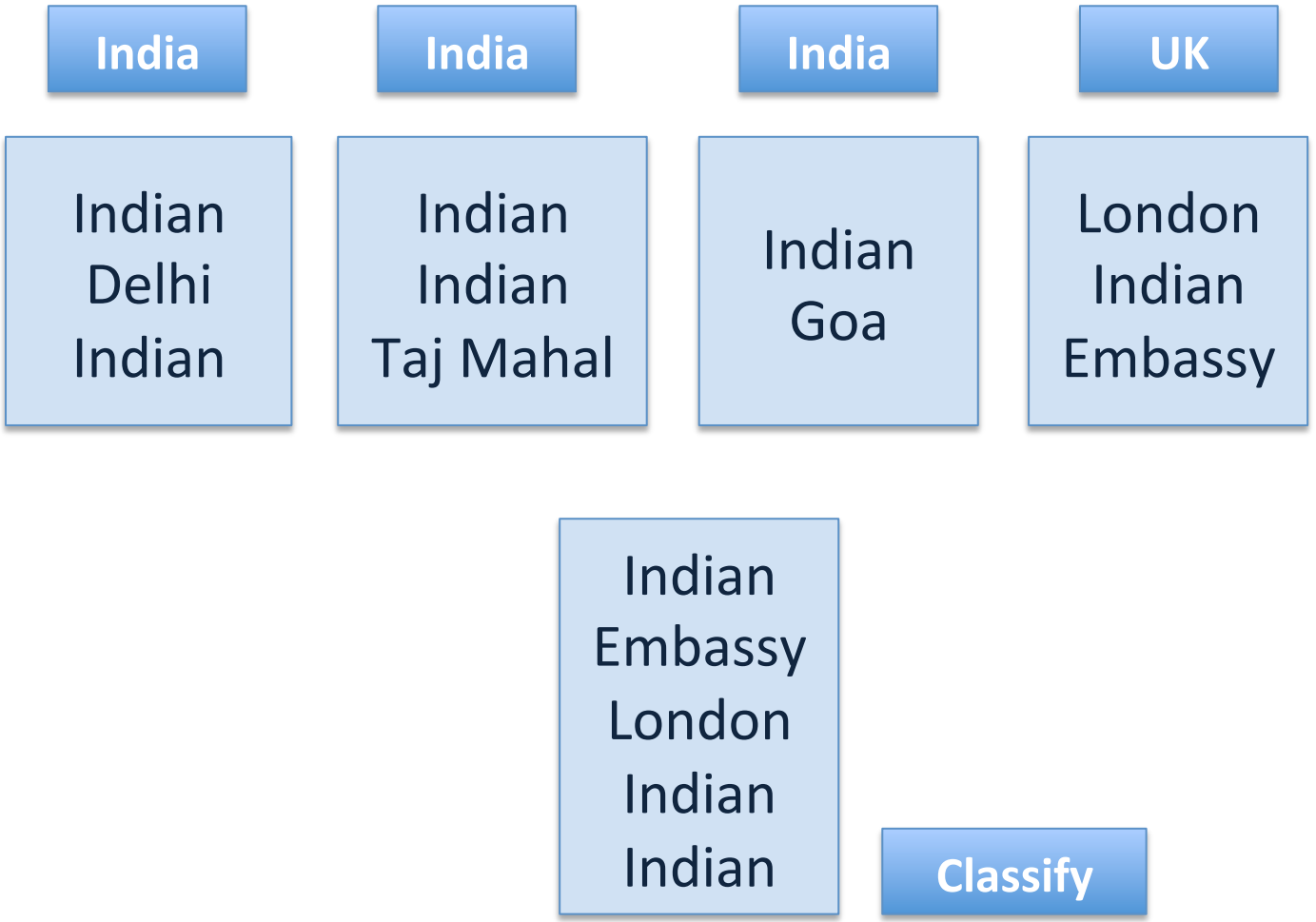
$$P(C \mid X_1,\ldots,X_n) = \prod_{i=1}^{n}\left[X_i P(t_i \mid C) + (1 - X_i)(1 - P(t_i \mid C))\right]$$

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

Absence of terms

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

# Example

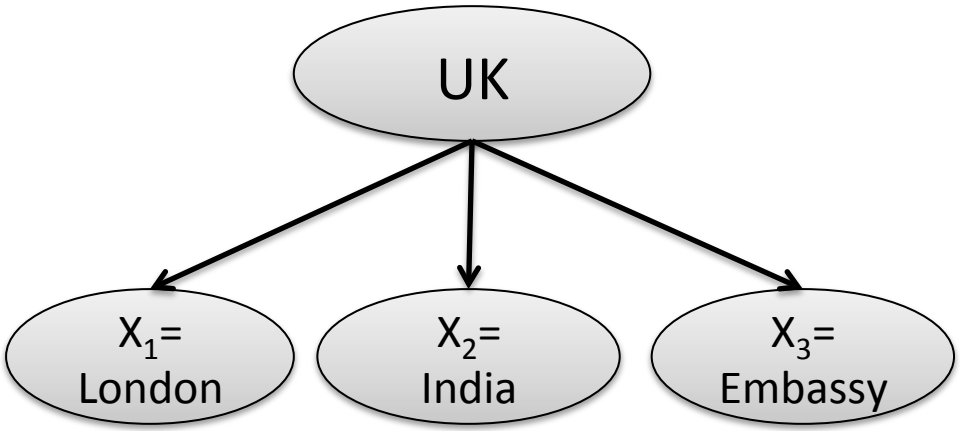

India
India
India
UK
Indian
Delhi
Indian
Indian
Indian
Taj Mahal
Indian
Goa
London
Indian
Embassy
Indian
Embassy
London
Indian
Indian
Classify

# 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




$$P(d \mid c) = P\left(\left\langle t_1, \ldots, t_{n_d} \right\rangle \middle| c\right)$$

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 → Bag of words model:

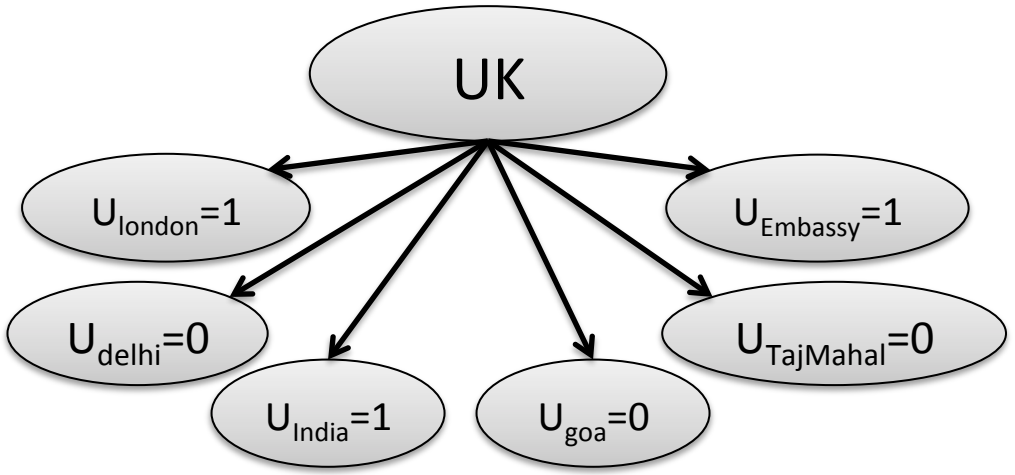
$$P(X_{k_1} = t \mid c) = P\left(X_{k_2} = t \middle| c\right)$$

# 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




$$P(d \mid c) = P\left(\left\langle e_1, \ldots, e_{|V|} \right\rangle \middle| c\right)$$

All terms in the vocabulary

$P(U_i{=}1|c)$ is the probability that term $t_i$ will occur in any position in a document of class $c$

# Multinomial vs Bernoulli

| | Multinomial | Bernoulli |
|---|---|---|
| Event Model | Generation of token | Generation of document |
| Multiple occurrences | Matters | Does not matter |
| Length of documents | Better for larger documents | Better for shorter documents |
| #Features | Can handle more | Works best with fewer |

# 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