

Latent Dirichlet Allocation

Sargur Srihari

srihari@cedar.buffalo.edu

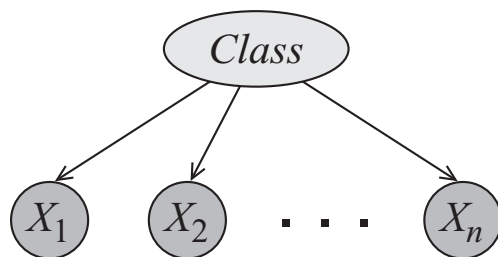
Topics

1. Bag-of-words for Text Classification
 1. Bernoulli Naïve Bayes
 2. Multinoulli Naïve Bayes
2. Latent Dirichlet Allocation

Bag-of-words Text Classification

- Document doc to be put into a category
 - Assume doc belongs to a single category
 - E.g., sports, economics
- Bag-of-words model
- Distribution of bag in different categories
- Naïve Bayes model
 - There still are design choices affecting performance

Naiive Bayes Model



$$P(C, X_1, \dots, X_n) = P(C) \prod_{i=1}^n P(X_i | C)$$

$$P(C | X_1, \dots, X_n) = \frac{P(C, X_1, \dots, X_n)}{\sum_c P(C, X_1, \dots, X_n)}$$

Encoded using a very small number of parameters

Linear in the number of variables

Random Variable Selection

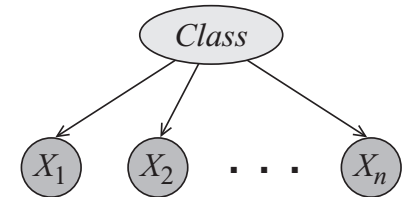
- Remove extraneous characters
 - Such as punctuation marks
- Remove stop words (which are content free)
 - the, and, ..
- Map words to canonical words
 - In pre-defined dictionary \mathcal{D}
 - apples \rightarrow apple
 - used \rightarrow use
 - running \rightarrow run

Two approaches to define features

1. Bernoulli Naïve Bayes model

- Binary attribute (Feature) X_i indicates whether $w_i \in \mathcal{D}$ appears in doc
 - Not how many times it appears in doc

$n = \text{No. of words in } \mathcal{D}; \text{Val}(X_i) \in \{0,1\}$



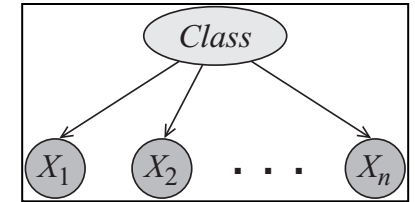
2. Multinomial Naïve Bayes model

- Attributes describe specific sequence of words
 - Attribute X_i indicates which \mathcal{D} -word appears in i^{th} pos
 - Thus each X_i takes one of many values, one for each possible word

$n = \text{No. of positions in doc}, \text{Val}(X_i) \in \mathcal{D}$

Parameters for Two approaches

- Bernoulli Naïve Bayes: X_i is binary
 - Learn frequency over a document of each dictionary word over each category C
 - E.g., probability that ball appear/not appear in sports
 - We learn a parameter for each (dictionary word, category)
- Multinomial Naïve Bayes: X_i is multi-valued
 - Simplifying assumption: word in position i does not depend on i i.e., $P(X_i=w)$ given the topic is same as $P(X_j=w)$
 - We use parameter sharing between $P(X_i|C)$ and $p(X_j|C)$
 - We need probability of ball appearing in sports
 - No. of parameters is again one for each (dictionary word, category)



Differences between two models

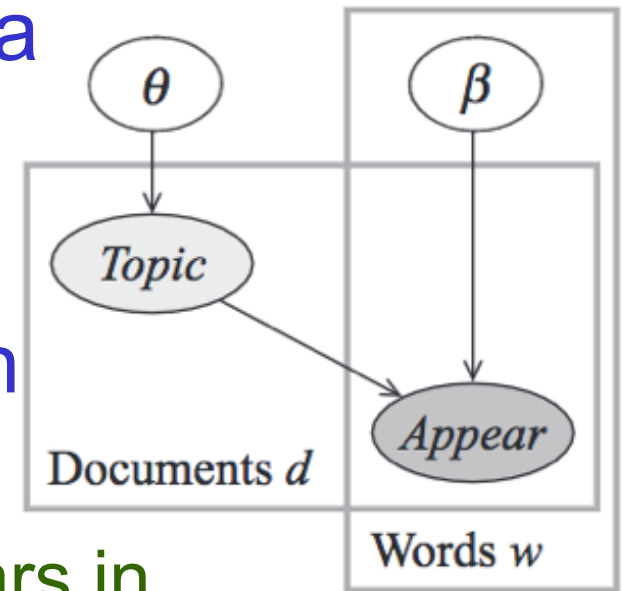
- Distributions are different
 - Two models give rise to different distributions
 - If word w appears in several positions in doc
 - Bernoulli ignores no. of occurrences
 - Multinomial multiplies $P(w|C)$ several times
 - If $P(w|C)$ is small in one category, probability of the document given the category will decrease
- Role of document length
 - Bernoulli: each document has same no of variables
 - Multinomial: documents of different lengths have different no of random variables
- Plate model makes subtle differences explicit₈

Objects for Plate Models

- Both models have two different kinds of objects: documents and individual words in documents
- Document objects d are associated with attribute T representing the document topic
- However the notion of “word objects” is different in the different models

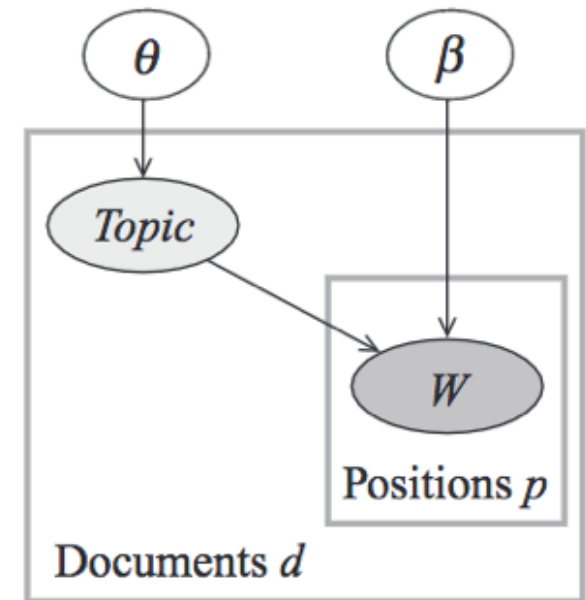
Bernoulli Naïve Bayes Model

- Words correspond to words in a dictionary
 - E,g., cat, computer, etc
- Binary attribute $A(d, w)$ for each document d , dictionary word w
 - Takes value true if word w appears in document d
- Can model this using pair of intersecting plates
 - One for documents and the other for dictionary words



Multinomial Naïve Bayes Model

- Word objects correspond not to dictionary words but to word positions P within the document
- Thus we have an attribute W of records representing pairs (D, P) where D is a document and P is a position within it
 - Attribute takes values in space of dictionary word p in document d
 - However all generated from same multinomial which depends on topic

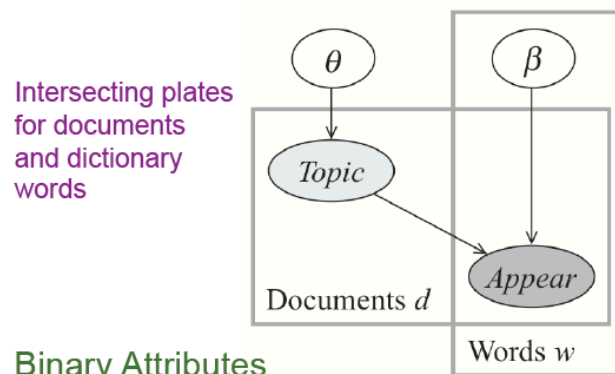


Two different Plate Models for Text

Both associate document d with Topic T

Bernoulli Naïve Bayes

No of features =
No of words in \mathbf{D}

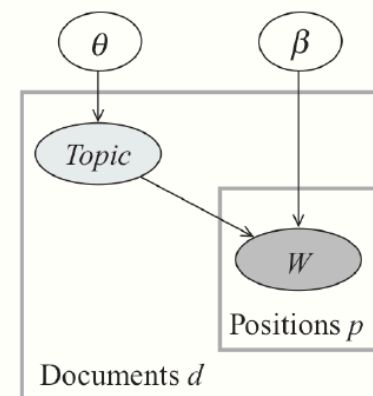


$Appear(d, w)$ for document d and word w takes value 1 if w appears in d

Bernoulli parameter $\beta_w[w]$ is different for different words

Multinomial Naïve Bayes

No of features =
No of positions in doc



Multinomial is generally more successful than Bernoulli

$W(d, p)$ for document d and position p takes value of dictionary word

Parameter β_w is the same for all positions

Parameter Estimation for Text

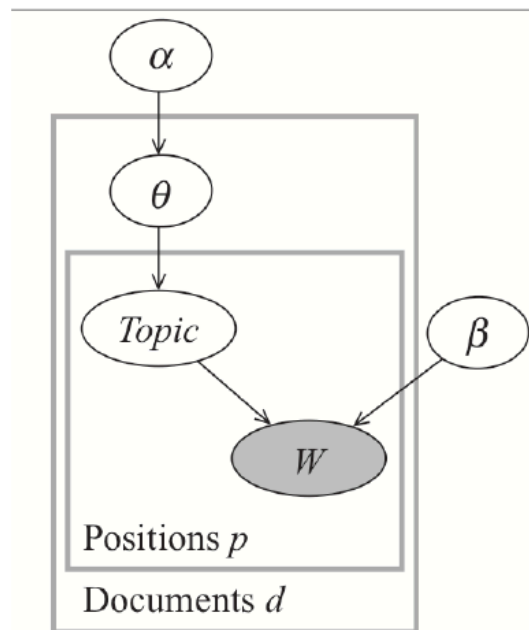
- In both models
 - Parameter estimated from data
 - Model used for classifying new documents
- Parameters measure prob. of word given topic
 - E.g., “bank” given “economics”
 - Bayesian parameter estimation avoids over-fitting
 - Especially ascribing zero to words not in training set
- With Bayesian estimation
 - can learn naïve Bayes using small corpus
 - Principal advantage of Naïve Bayes
 - More realistic language models are harder

Richer representations

- Can capture finer-grained structure in distribution
 - LDA extends multinomial naïve Bayes model
- As with multinomial naïve Bayes we have a set of topics associated with a set of multinomial distributions θ_w over words
 - We do not assume that the entire document is about a single topic
 - Rather a continuous mixture of topics defined using $\theta(d)$

Latent Dirichlet Allocation

- Extends the multinomial naïve Bayes model
- Set of topics associated with set of multinomial distributions θ_W over words



Document d is associated with a continuous mixture of topics defined using parameters $\theta(d)$

Parameters selected independently from each document d , from a Dirichlet distribution parameterized for a set of hyper-parameters α

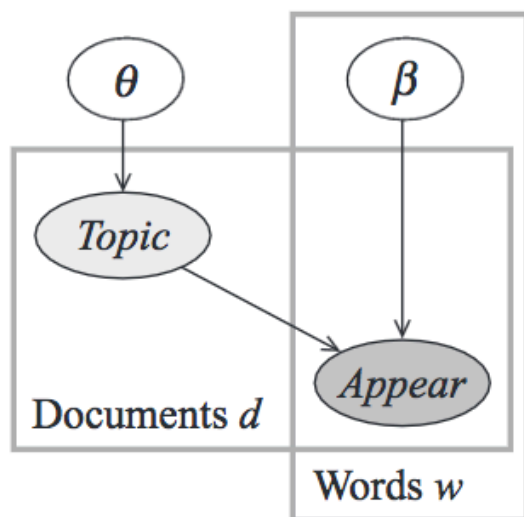
Word in position p of document d is selected by first selecting topic $Topic(d,p)=t$ from mixture $\theta(d)$ and then selecting specific dictionary word from Multinomial β , associated with topic t

Avoids over-fitting problems with the two naïve Bayes models

Summary of plate models for text

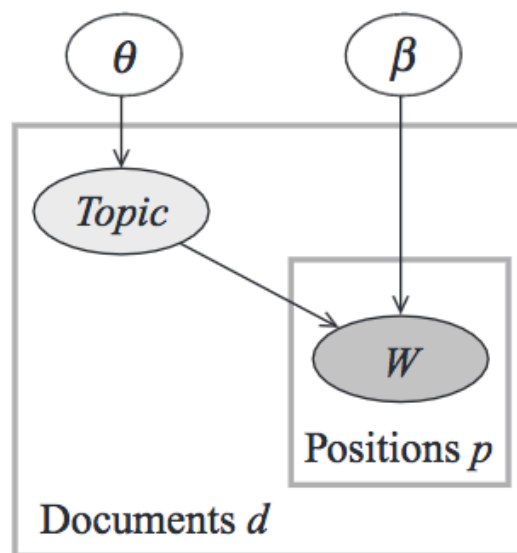
Latent Dirichlet Allocation

Bernoulli Naïve Bayes

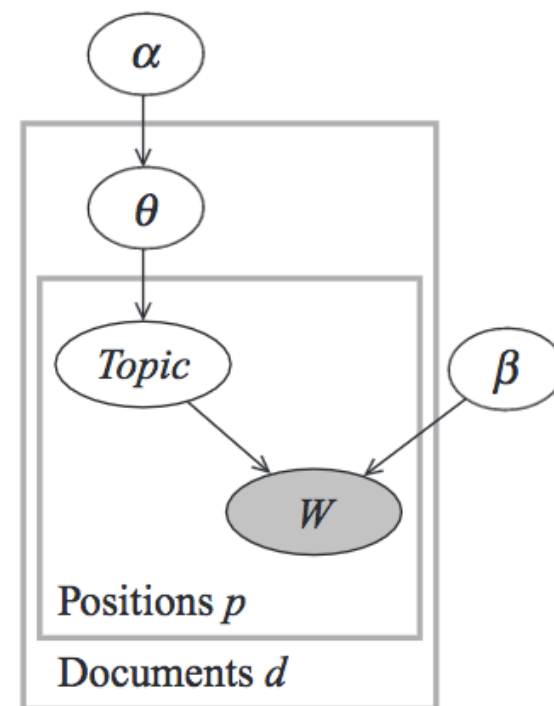


(a)

Multinomial Naïve Bayes



(b)



(c)

Summary of BN Parameter Estimation

- Examined parameter estimation for Bayesian networks
 - When data are complete
- Discussed two approaches
 - MLE and Bayesian