COS 424: Interacting with Data

Lecturer: Dave Blei — Lecture #9
Scribes: James An and Tsheko Mutungu — March 4, 2008

1 Mixture modeling

Recall naive Bayes classifier.

Joint distribution:

$$p(c, w_{1:N}) = p(c) \prod_{n=1}^{N} p(w_n|c) \tag{1}$$

Posterior:

$$p(c|w_{1:N}) = \frac{p(c) \prod_{n=1}^{N} p(w_n|c)}{\sum_c p(c) \prod_{n=1}^{N} p(w_n|c)} \tag{2}$$

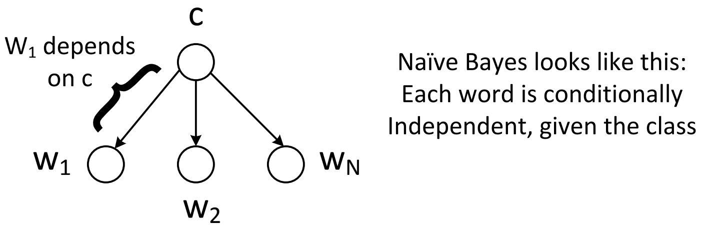
2 Graphical Models ("Bayesian Networks")

GMs capture joint distributions over an ensemble of random variables and encode the conditional probabilities within the ensemble.

- Directed Acyclic Graphs
- Nodes = random variables
- Edges = dependence relationships
- GMs represent a factorization of the joint:

$$p(c, w_{1:N}) = p(c) \prod_{n=1}^{N} p(w_n|c) \tag{3}$$

Note that this is equivalent to equation (1).


c
W1 depends on c
W1
W2
WN
Naïve Bayes looks like this:
Each word is conditionally
Independent, given the class

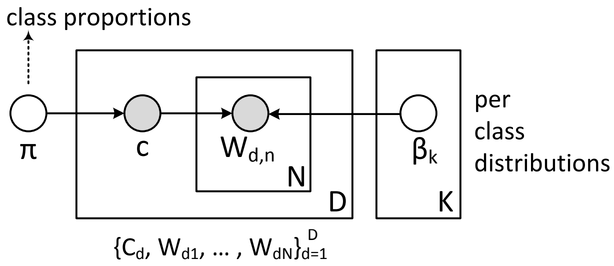
The nodes are ordered such that no node appears before its parents and the child nodes are conditioned on their parents

In general,

$$p(X_1, ..., X_N) = \prod_{n=1}^{N} p(X_n|P_a(X_n)) \tag{4}$$

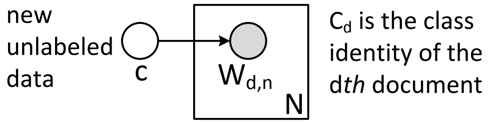
$P_a(X_n)$ = parents of X_n

“Plate” notation

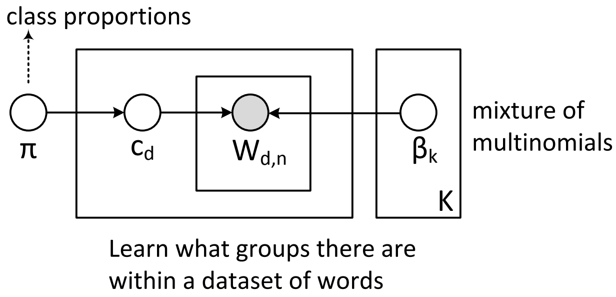


Observed parameters are shaded; computed parameters are not

At testing time, we only observe N words



What if we’ve never seen c_d?



Each document is associated with some “latent” group

Goal: to sort data into groups à la clustering.

3 Mixture modeling

Data are divided into K groups. We never observe the groups.

3.1 Estimating parameters

Find π and $\beta_{1:k}$ from data.

π are the mixture proportions.

$\beta_{1:k}$ are the mixture components.

Data is “what is observed”. The data are documents: $\{w_{d,1:N}\}_{d=1}^{D}$

Use maximum likelihood estimation to estimate π and $\beta_{1:k}$

3.2 MLE

$$\log p(D|\pi,\beta,k) = \sum_{d=1}^{D} \log p(w_{d,1:N}|\pi,\beta_{1:k}) \tag{5}$$

$$\log p(w_{d,1:N}) = \log \sum_{c=1}^{k} p(c, w_{d,1:N}) = \log \sum_{c=1}^{k} p(c) \prod_{n=1}^{N} p(w_{d,1:N}|c) \tag{6}$$

Log likelihood of data:

$$\log p(D|\beta_{1:k},\pi) = \sum_{d=1}^{D} \log(\sum_{c=1}^{k} p(c|\pi) \prod_{n=1}^{N} p(w_n|\beta_c)) \tag{7}$$

3.3 Parameters

How can we optimize with respect to π and $\beta_{1:k}$?

One option is to use numerical methods (i.e. some optimization technique).

Another option is the Expectation-Maximization (EM) algorithm. This algorithm was secretly invented at Princeton by the Defense Department. The algorithm is used for finding the MLE (a good local optimum) in the face of latent variables.