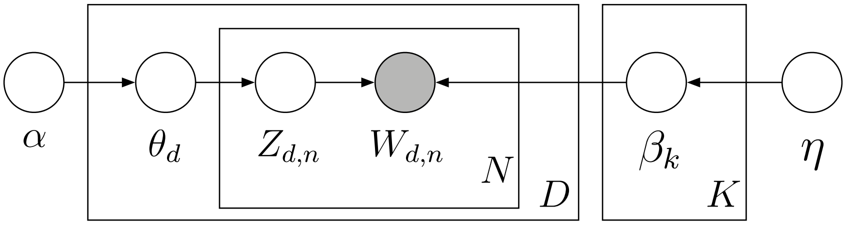
COS 424: Interacting with Data

Lecturer: David
Lecture: Final Lecture!
Scribe: Ben Sonday

1 Topic models

Today we discuss topic models, which are essentially an extension of mixture models. The difference is that with mixture models, we assumed that each document's composition of words was determined by a distribution over a set of words in a vocabulary, and that this distribution is determined by the class of the document. Topic models allow for a document to be a "mixture" of different classes (classes are called "topics" in this setting). For instance, a scientific paper in *Science* about hereditary diseases in fish populations might be a linear combination of the topics *Genetics*, *Evolution*, and *Disease*. More advanced topic models ultimately allow us to do things like construct networks showing the interrelatedness of topics, determine the evolution of the "lingo" associated with a particular topic, or even automatically caption pictures.

1.1 The Graphical Model for Topics



This is often referred to as a mixed membership model since each word in a document is now assigned to a particular topic ($Z_{d,n} \to W_{d,n}$) instead of the document as a whole, and hence, each document has "mixed" membership with mixing proportions given by θ_d. Before, for document classification, we used a mixture model, where, as said already, each word in a particular document comes from the *same* class (topic), and each document belongs to only onc class.

1.2 The Posterior

Unfortunately, computing the posterior for the above problem is intractable:

$$p(\theta, z_{1:N}|w_{1:N}, \alpha, \beta_{1:K}) = \frac{p(\theta|\alpha)\Pi_{n=1}^{N} p(z_n|\theta)p(w_n|z_n, \beta_{1:K})}{\int_\theta p(\theta|\alpha)\Pi_{n=1}^{N} \sum_{z=1}^{K} p(z_n|\theta)p(w_n|z_n, \beta_{1:K})} \tag{1}$$

In fact, this equation is a nasty special hypergeometric function. The approach to use is variational inference. The basic idea of variational inference is to formulate the computation of a marginal or conditional probability in terms of an optimization problem. This (generally intractable) problem is then "relaxed," yielding a simplified optimization problem that depends on a number of free parameters, known as variational parameters. Solving for the variational parameters gives an approximation to the marginal or conditional probabilities of interest (from paper of Blei, 2004). On the above problem, a variational EM scheme is used, where on the E step, we estimate the posterior using variational inference, and then do the M step as usual.

2 Class Summary

In this class, we have discussed essentially four main classes of algorithms.

2.1 Classification

Naive Bayes, Support vector machines, Logistic regression, Boosting, Decision trees, and Nearest-neighbor/K-nearest-neighbor

Naive Bayes can be mathematically coerced into logistic regression form, and both boosting and SVM are margin-based algorithms.

2.2 Clustering

Agglomerative clustering, K-means, Mixture modeling

K means is closely related to mixture modeling as we showed in homework #3. Naive Bayes is also related to mixture modeling, since in naive Bayes we can observe the class of a "document," and in mixture modeling it is hidden.

2.3 Prediction

Linear regression, Polynomial regression (note that logistic regression is a classification algorithm and doesn't really belong on this list)

Linear regression is closely related to logistic regression since both are of the class of generalized linear models.

2.4 Dimension Reduction

Principal component analysis, Factor analysis

PCA and factor analysis are very closely related with the major difference being that factor analysis allows for noise that has different variances in different directions. Both are related to linear regression when finding the optimal projections.

2.5 An analogy

Classification is to clustering as prediction is to dimension reduction. Classification and prediction are both models whose optimal parameters can be determined by maximum likelihood, and clustering and dimension reduction are latent variable models whose optimal parameters can be determined using the EM algorithm. Classification and clustering give categorical labels to data, and prediction and dimension reduction work in the real numbers.