CAP 6617 ADVANCED MACHINE LEARNING

NAME: SANKET KUMAR UF ID: 4513-3352

LATENT DIRICHLET ALLOCATION

AIM: Latent Dirichlet Allocation (LDA), a generative probabilistic model for collections of discrete data is described.

- Latent Dirichlet allocation (LDA) is a generative probabilistic model of a corpus. The basic idea is that documents are represented as random mixtures over latent topics, where each topic is characterized by a distribution over words. LDA is a Bayesian version of Probabilistic Latent Semantic Indexing (PLSI). The generative process for each document w in corpus D is: 1) Choose a sequence of words 2) Choose a parameter through Dirichlet prior 3) For each of the words: Choose a topic and choose a word from a multinomial probability conditioned on the topic.
- The term frequency-inverse document frequency approach provided small amount of reduction in description length and revealed little of the statistical structure of the document. This lead to the discovery of Latent Semantic Indexing (LSI), which uses a singular value decomposition of the tf-idf matrix to identify a linear subspace. The next big thing was Probabilistic LSI which models each word in a document as a sample from a mixture model, where the mixture components are multinomial random variables that can be viewed as representations of "topics". Each document is represented as a list of mixing proportions for these mixture components and thereby reduced to a probability distribution on a fixed set of topics. But there were two problems: 1) Number of parameters grow linearly with size of corpus. 2) It is not clear how to assign probability to a document outside of the training set.
- The assumption in LDA is that the topics are infinitely exchangeable within a document. Each topic is a latent multinomial variable corresponding to a distribution over words. It is a 3-level model; the topic node is sampled repeatedly within the document. Under this model, documents can be associated with multiple topics.
- A quick comparison of latent variable models is: 1) Unigram chooses one point in the word simplex. 2) Mixture of unigrams chooses k points in the word simplex and each document corresponds to a point. 3) PLSI chooses k points in the word simplex and chooses one point in the topic simplex for each document in the corpus. 4) LDA chooses k points in the word simplex and then learns a distribution over the topic simplex.
- Parameters can be estimated using Maximum Likelihood estimator. There are fewer parameters (only 2: α and β) in LDA in comparison to PLSI. But now the key problem is computing the posterior distribution of the hidden variables for a document, which is generally computationally intractable. Though intractable, a wide variety of approximation inference algorithms can be considered for LDA.
- The basic idea of convexity-based variational inference is to make use of Jensen's inequality to obtain an adjustable lower bound on the log likelihood. By dropping the coupling edges between topic distribution for documents, topic distribution for i-th word in a document and the document nodes, and endowing the resulting simplified graphical model with free variational parameters, we obtain a family of distributions on the latent variables. These variational parameters call for optimization that determines its final values. These are found by minimizing KL divergence between the variational distribution and true posterior via an iterative fixed point method.
- The variational distribution is a conditional distribution because the optimization problem is conducted for a fixed document, and thus yields optimizing parameters that are a function of the document. We can write the resulting variational distribution where we can make the dependence on a document explicit and thus, this distribution can be viewed as an approximation to the posterior distribution.
- We wish to find parameters α and β that maximal the marginal log likelihood of the data, where w_d's are documents in a corpus:

$$\ell(\alpha, \beta) = \sum_{d=1}^{M} \log p(\mathbf{w}_d \mid \alpha, \beta).$$

We can use variational EM to approximate empirical Bayes estimates for the LDA model since *p(w|* α*,* β*)* is computationally intractable. The iterative algorithm is: 1) E-step: For each document, find the optimizing values of the variational parameters. 2) M-step: Maximize the resulting lower bound on the log likelihood with respect to the model parameters α and β. This corresponds to finding maximum likelihood estimates with expected sufficient statistics for each document under the approximate posterior which is computed in the E-step. There two steps are rerepeated until the lower bound on the log likelihood converges. A new document is very likely to contain words that did not appear in any of the documents in a training corpus. By smoothing multinomial parameters, we assign positive probabilities to all vocabulary items whether or not they are observed in the training set.