Project Plan

LDA with Deep Neural Networks

Aaditya M Nair (201302161)
Parth Kolekar (201301143)
Atul Agarwal(201330188)

What is LDA?

Latent Dirichlet allocation (**LDA**) is a generative statistical model that allows sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. It represents documents as a mixture of topics that spit out words with certain probabilities. It is a bag of words model.

Learning

We have set of documents and we want to learn the topic representation of each document and the words associated with each topic. We will apply collapsed Gibbs sampling:

1. We choose some K topics to be discovered.
2. We go to each document and randomly assign each word to those K topics. It gives us topic representation and words associated with each topic(not accurate).
3. For each document d we go to each word w we compute two things, (a) for each topic t we compute p(t|d), i.e. , the proportion of the words in document d that are currently assigned to topic t;(b) for that topic we compute p(w|t) , i.e. , proportion of assignment to topic t that comes from the word w.
4. Then we choose the new topic according to the p(t|d) * p(w|t). In other words, we assume that the rest of the assignment is correct and we assign the new topic accordingly.
5. After repeating these four steps large number of times we will converge to a steady state and then we can estimate the topic mixture of the document by calculating proportion of words assign to each topic within the document and the words associated with each topic by counting the words assign to each topic.

We will get a LDA model for the chosen topics for the set of documents.

Use of Deep Learning

As we can see above, LDA is highly resource intensive and inference is very slow and often intractable. We are interested in training a DNN to perform an LDA. Specifically, we will be using an LDA as a teacher model to guide the training of a DNN, so that the DNN can approximate the behaviour and the performance of an LDA.

Since a deep neural network is much faster than an LDA, this model will be much more useful for online classification tasks.

Plan of Action.

1. **Get a dataset.** We plan to use some easily available text dataset like ones from UCI ML database (e.g. *reuters21578*) for the project. We will pre-process this data to remove unnecessary words.
2. **Topic Modelling.** We then run the LDA on the dataset to determine the topic for the each of the data.
3. **Deep Learning.** We now use this topic classified data to train the neural network using term frequency as the feature.
4. **Test and Compare.** Now we compare the output from the trained DNN and the LDA on a separate test data.

Scope of the project

1. In this project we are going to compare the performance of LDA vs DNN for different number of topics.
2. For the deep neural network, we will be comparing a two-layer DNN with a three-layer DNN.
3. If time permits, we will analyse the effect of different loss functions on the performance of the neural network.

Technologies Used

1. Sklearn - For a model of Latent Dirichlet Allocation using online variational Bayes algorithm.
2. Tensorflow - For a DNN model