Topic Modeling Using Latent Dirichlet Allocation

CSE250B, Winter 2014, Project 3

Prabhav Agrawal
UC San Diego
prabhav@ucsd.edu

Soham Shah
UC San Diego
srshah@ucsd.edu

March 8, 2014

---

1 Abstract

Text documents are typically made up of multiple topics in varying proportions. Although it very likely that one topic dominates a document, there are still hints of other topics in very small proportions. In other instances, documents are fundamentally span over multiple topics, and in such cases predicting a single output topic is just not acceptable. We want a more general model, which can capture the distribution of topics. LDA, an inherently probabilistic generative model, serves as a perfect candidate to learn the hidden topics in the corpus and give the distribution of these latent topics for each document. We use LDA to perform topic modeling on two datasets, Classic400 and the BBC dataset and use inspection, VI distance and clustering accuracy to analyze these clusters. On inspection we find meaning topics and our clustering accuracy and VI distance along with the 3D mapping agree very well with our inferences from inspection.

2 Introduction

Our task is to train an unsupervised learning model to discover hidden topics in a given corpus of documents. We use a latent Dirichlet allocation (LDA) model to learn the hidden topics for each word instance in the given set of documents. Using word instances allows us to assign different topics for each occurrence of word in the document, enabling us to achieve the notion of representing multiple topics per document.
We evaluate the effectiveness of LDA on two datasets, Classic400 and the BBC dataset. We perform LDA using different values for the hyper parameters number of topics K, α and β. We run collapsed Gibbs sampling for 500 epochs and manually examine the top 10 words for each topic. We see meaningful topics such as science, medicine and dynamics for the Classic400 dataset and topics such as football, cricket, rugby, tennis and athletics for the BBC dataset. Given the set of true labels for each document we test the quality of clustering of the LDA model on the datasets, using VI distance and clustering accuracy. We also discuss issues associated with Gibbs sampling such as goodness of fit and overfitting.Varying the number of topics shows us interesting results such as splitting of topics for higher values of K, such as splitting the topic dynamics in Class400 into aerodynamics and fluid dynamics. Similarly, for lower values of K, we see topics with high counts split into two. On manual inspection we see most meaningful clustering when K is equal to number of true labels, $alpha = \frac{2}{K}$ and $beta = 1$. Our plot of the dataset in LDA topic space and VI distance measures also gives best results for this combination. Section 3 talks about LDA and the collapsed Gibbs sampling technique used to train such a model. Section 4 then explains the design of our experiments followed by Section 5 which depicts our results, observations and lessons learnt through the course of the project.

3 Model analysis

3.1 Latent Dirichlet allocation

Latent Dirichlet allocation (LDA) is an unsupervised learning technique to discover latent semantic topics in a collection of documents . LDA is based on the intuition that each document consists of words from different topics, where the proportion of each topic varies but the topics themselves are the same for all documents. LDA achieves this notion of multiple topics per document by assigning each word instance to its appropriate

topic. Learning in this model is unsupervised because the input data is incomplete: the corpus provides only bag-of-words representation of documents; there are no training labels provided.

3.2 Representation of Data

Ideally we would like to represent data in a form that retains the ordering of words, however this representation would lead to each document having varying lengths. Instead it is sufficient to use a simple bag-of-words representation that only retains the count of the words in the document. We need to identify the set of all words that occur in the corpus, we define this set of words as vocabulary V. Further optimizations can be made, such as including only words that occur at least twice and removing "stop" words - words that are common in most documents. After the vocabulary is fixed, each document is represented as a vector with integer entries of length m where m is the size of the vocabulary. x_j represents the number of occurrence of word j in the document. The length of the document is $n=\sum_{j=1}^{m} x_j$. The collection of documents is represented as a two-dimensional matrix where each row describes a document and each column corresponds to a word.

3.3 Generation Process

Unsupervised learning is often done by assuming that the data were generated by some probabilistic process, and then tries to learn the parameters of this process. The generative process assumed by LDA is as follows.

Given: Dirichlet distribution with parameter vector α of length K
Given: Direchlet distribution with parameter vector β of length V
for topic number 1 to topic number K
  draw a multinomial with parameter vector ϕ_k according to β
for document number 1 to document number M
  draw a topic distribution, i.e. a multinomial θ according to α
  for each word in the document
    draw a topic z according to θ
    draw a word w according to ϕ_z

3.4 Training the model

We assume the prior distributions α and β are fixed and known, along with the number of topics K, number of documents M, length of each document N_m and the size of vocabulary V. Our goal is to infer (i) the document-topic distribution θ for each document m and (ii) topic-word distribution ϕ for each topic k.
We use collapsed Gibbs sampling to infer the hidden value z for each word occurrence in each document. We use the term "word occurrence" for multiple appearances of the same word, in the same or different document, may be assigned to different topics; thus every word has its own z value.
Let $\bar{w}$ be the sequence of words making up the entire corpus, and let $\bar{z}$ be a corresponding sequence of z values. Pick a random word occurrence w_i. Suppose we know the value of z for every word occurrence in the corpus except for word w_i. The idea of Gibbs sampling is to draw a z_i for i randomly according to its current distribution, then assign this topic z_i for the word i and then repeat this process for next word occurrence w_i. It can be proved that eventually this process converges to a correct distribution of z values for all the words in the corpus. We perform Gibbs sampling using the following equation:

$$p(z_i = j|\bar{z}', \bar{w}) \propto \frac{q'_{jw_i} + \beta_{w_i}}{\Sigma_t q'_{jt} + \beta_t} \frac{n'_{mj} + \alpha_j}{\Sigma_k n'_{mk} + \alpha_k} \tag{1}$$

where n_{mk} is the number of times $z_i = k$ within document m and q_{jt} is the number of times that word t occurs with topic j in the while corpus. n'_{mk} and q'_{jt} are n_{mk} and q_{jt} respectively with one subtracted for the topic z_i and word occurrence w_i respectively.

4 Design of experiments

4.1 Datasets

We conduct experiments on two datasets to evaluate performance of LDA in topic modeling. First one is the Classic400 dataset which has 400 documents with a 6205 word vocabulary. The data has been provided as a sparse matrix where each row corresponds to the count of non-zero words in a document.

Second dataset is the BBC dataset which consists of documents from the BBC Sport website corresponding to sports news articles in five topical areas: athletics, cricket, football, rugby, tennis. This dataset has 737 documents over a vocabulary of 4613 words. The preprocessing steps included removal of stop-words, applying Porter's stemmer on the wordlist and then converting document-word counts into a sparse matrix format

4.2 Goodness-of-fit measurement

Evaluating goodness-of-fit of topic models is a tricky task. If topics ϕ_1 to ϕ_k are given, then goodness-of-fit can be modeled as $p(\mathbf{w}|\phi, \alpha)$ which is the probability of data or likelihood of the model. It is defined as:

$$p(\mathbf{w}|\phi, \alpha) = \int_\theta p(\mathbf{w}|\theta, \phi)p(\theta|\alpha) = \int_\theta \int_z p(\mathbf{w}|\mathbf{z}, \phi)p(\mathbf{z}|\theta) \tag{2}$$

If the topic ϕ are not given, the goodness-of-fit can be modeled as $p(w|\alpha, \beta)$. This is defined as:

$$p(\mathbf{w}|\alpha, \beta) = \int_\phi p(\mathbf{w}|\phi, \alpha)p(\phi|\beta) \tag{3}$$

This is difficult to calculate and several approximations exists for this.[2]

4.3 Analysis of clustering

The document-topic distributions produced by LDA are a soft clustering. (i.e. we have the probabilities for topics in a document).
One naive approach can be to assign each document the topic with highest probability. When number of topics K is equal to the natural classes given, then a clustering accuracy can be computed. But, we do not know the correspondence between LDA topics and natural classes, we did not use this approach. Instead, we trained a KNN (K nearest neighbour) classifier by giving feature vector for a document as the probabilities of topics associated with it. The prediction accuracy obtained by the classifier is used for measuring clustering quality in our experiments.
Another approach is to treat the given categorization as a deterministic topic distribution for each document, and then compute distance between this distribution and LDA topic distribution. We use the Variation of Information distance (VI-distance)[...ref..] as the distance measure.

Given two class distributions for a document: $p(c = j|d_m)$ and $p(z = k|d_m)$ with true labels $j = 1...J$ and $m = 1....M$. We evaluate class probabilities $p(c = j = \frac{1}{M}\Sigma_m p(c = j|d_m))$ and $p(z = k) = \frac{1}{M}\Sigma_m p(z = k|d_m)$ as well the probabilities of co-occurrence of pairs of classes $p(c = j, z = k) = \frac{1}{M}\sum_{m=1}^{M} p(c = j|d_m)p(z = k|d_m)$
VI-distance between two distributions is defined as:

$$D_{VI}(C, Z) = H(C) + H(Z) - 2I(C, Z) \tag{4}$$

where $H(Z) = -\sum_{k=1}^{K} p(z = k)log_2p(z = k)$ and $H(C) = -\sum_{j=1}^{J} p(c = j)log_2p(c = j)$ $H(Z)$ and $H(C)$ are the entropies of corresponding distributions Mutual information $I(C, Z)$ is defined as:

$$I(C, Z) = \sum_{j=1}^{J}\sum_{k=1}^{K} p(c = j, z = k)[log_2p(c = j, z = k) - log_2p(c = j)p(z = k)] \tag{5}$$

We also report results of VI-distance between distribution of true labels and LDA topic model distribution.

4.4 Convergence and overfitting

We are checking for convergence by monitoring the values of VI-distance and classifier accuracy on the training set. We can check for overfitting by computing the test set classifier accuracy and compare it with that on training set. For no overfitting, difference between two should be less.

4.5 Effect of hyperparameters

Hyperparameter α represents the strength of topics in a document and β represents the strength of words in a topic. Since LDA model treats all the topics and the words in the same way, we choose same α for all topics and same β for all words. If the values of α and β are high, that means our prior beliefs are strong and more iterations of Gibbs sampling are needed to revise the beliefs. α and β also have a smoothing effect on document-topic and topic-word distribution as well. Lower the value, less will be the smoothing and distributions will be well separated. In our experiments, we have used $\alpha\epsilon\{1/K, 5/K\}$ and $\beta\epsilon\{0.1, 1\}$.

4.6 Number of Topics

An obvious choice for K or number of topics is to take them equal to number of natural classes given. If K is smaller than the number of natural classes, then it is possible that some of the topics may merge. Similarly, for K larger than number of given categories, we might discover sub-topics from the corpus or we may discover the noise as a topic. In our experiments, we choose $K = 3, 4$ for Classic400 dataset with 3 natural categories and $K = 4, 5, 6$ for BBC dataset with 5 natural categories.

4.7 Implementation Details

We initialize a random z_i for every word instance w_i in the corpus. Using these z_i values we compute document-topic distribution count $\bar{n}_m$ for every document m and topic-word distribution count $\bar{q}_k$ for every topic k. Once initialized we use these document-topic distribution and topic-word distribution counts along with α and β priors to sample a new topic for a randomly chosen word instance w_i accordingly to equation 1. More specifically, we draw a random number uniformly between 0 and 1, and index into the unit interval which is divided into subintervals of length p(z_i=j—$\bar{z}'$, $\bar{w}$). Each time, we assign a new topic to a word instance w_i we update both these distributions in constant time.

The LDA generative model treats each word in each document individually. However, the specific order in which words are generated does not influence the probability of a document according to the generative model. Similarly, the Gibbs sampling algorithm works with a specific ordering of the words in the corpus, however again any ordering will do. Thus we do not need to know the ordering of words in the original document corpus, we just need to take an arbitrary ordering and work with it. Thus a LDA model can be learnt on a standard bag-of-words model.

To compute equation 1 in constant time, we initially compute both the distribution counts as well as the sum of these distributions along their rows. As a result each term in equation one is obtained in constant time. As a word instance w_i is assigned a new topic we update both the distributions and the sum of the distributions, since at most one entry can change in each of these entities, all this can be done in constant time. Let there be K possible topics and N be the total number of words in the all the documents, then the time to do one epoch according to our implementation is O(NK) time.

We are able to run one epoch in under one second on the Classic400 dataset and under 6 seconds on the BBC dataset. We consider this performance high enough thus we do not try to achieve further optimization of the inner loop for Gibbs Sampling.

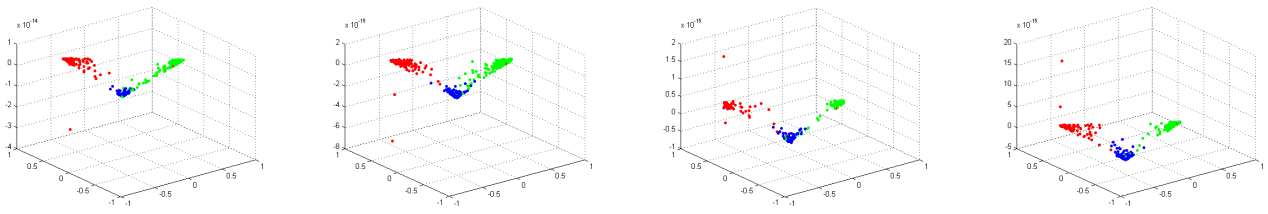
We train the model for 500 epochs, varying the values of $\alpha \in \{\frac{1}{K}, \frac{5}{K}\}$ and $\beta \in \{0.1, 1\}$. For Classic400 we vary $K \in \{3, 4\}$ and for BBC dataset we vary $K \in \{4, 5, 6\}$.

We test for both accuracy and VI distance at the end of each epoch. For testing accuracy, we train a classifier using $X = \theta$ and $Y = true_labels$ and use it to predict the true labels, where accuracy is measured by the percentage of matches between the the predicted value of the classifier and the actual true labels. We also used equation 4 and 5 to compute the VI distance at the end of every epoch and plot its value against the number of iterations.

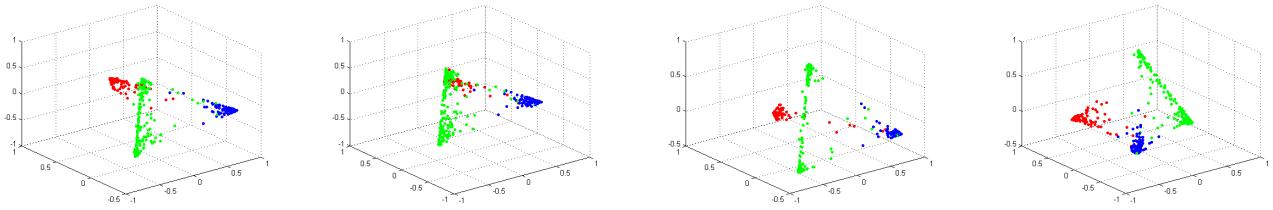
5 Results and discussion

5.1 Classic400 dataset

We show the 3D scatter plot of documents in topic space in Figure 1. Each document is represented as a point with probabilities of each topic in the document is taken as its coordinates. The color of the point represents the truelabel associated with it. For small α, documents are well separated and located towards the corners but for larger α, they are clustered close to each other and topic mixing is more as compared to small α. For a higher value of β i.e. for $\beta = 1$, we are obtaining good clustering as compared to $\beta = 0.1$ For $K = 3$, the documents reside on a 3-simplex. For $K = 4$, we first used principal component analysis (PCA) to project the data on a 3D space and then plotted them.



(a) $\alpha = 0.66, \beta = 0.1, K = 3$ (b) $\alpha = 1.66, \beta = 0.1, K = 3$ (c) $\alpha = 0.66, \beta = 1, K = 3$ (d) $\alpha = 1.66, \beta = 1, K = 3$



(e) $\alpha = 0.5, \beta = 0.1, K = 4$ (f) $\alpha = 1.25, \beta = 0.1, K = 4$ (g) $\alpha = 0.5, \beta = 1, K = 4$ (h) $\alpha = 1.25, \beta = 1, K = 4$

Figure 1: Plots of Classic400 dataset in LDA topic space

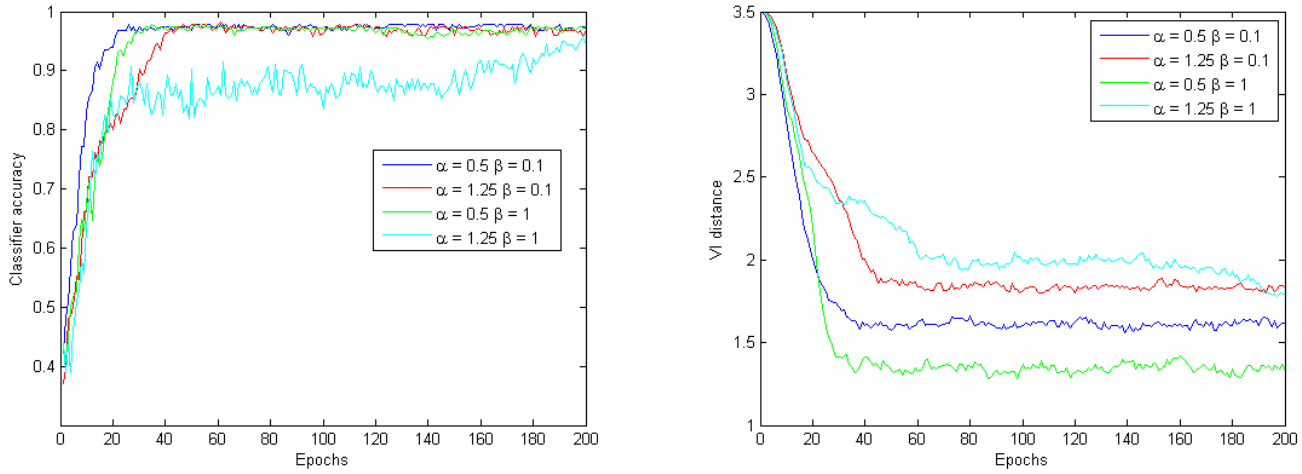
We display the top ten words for each LDA topic in Table1. We have identified three categories from Classic400 dataset - Dynamics, Medicine and Science according to documents given. When $K = 3$, top ten words in each LDA topic are suggestive of the categories identified. For $K = 4$, we observe that the topic Dynamics splits into two sub-topics - one related to Aerodynamics and other to Fluid-dynamics. It's important to note here that count of words assigned to Dynamics topic for $K = 3$ is almost two times that of other two topics.

Aerodynamics	Science	Fluid-dynamics	Medicine
wing	system	boundary	patients
mach	scientific	solution	ventricular
supersonic	research	plate	left
wings	retrieval	laminar	fatty
ratio	science	layer	cases
shock	language	transfer	acids
lift	methods	temperature	aortic
aerodynamic	systems	field	blood
layer	problems	fluid	normal
boundary	journals	problem	glucose

Dynamics	Medicine	Science
boundary	patients	system
layer	ventricular	scientific
wing	fatty	research
mach	left	retrieval
supersonic	nickel	science
ratio	cases	language
wings	acids	subject
velocity	aortic	methods
shock	blood	systems
effects	normal	journals

Table 1: Top 10 words for each LDA topic for K=3 and K=4 in Classic400 dataset

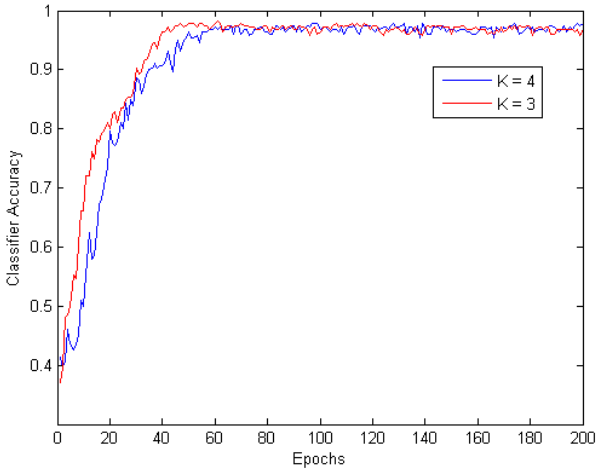
Higher classification accuracy suggests that our model fits the true labels well. Smaller VI-distance suggests that learned document-topic distribution has higher similarity to distribution of true labels. In Figure 2, results also support our observation in Figure1 that smaller α leads to better clustering. Similarly, a larger value of β is preferred but effect of α seems to dominate effect of β. A stability in the value of both classification accuracy and VI-distance suggests that our topic models have achieved convergence. For all combination of hyperparameters, we achieve more than 85% accuracy.



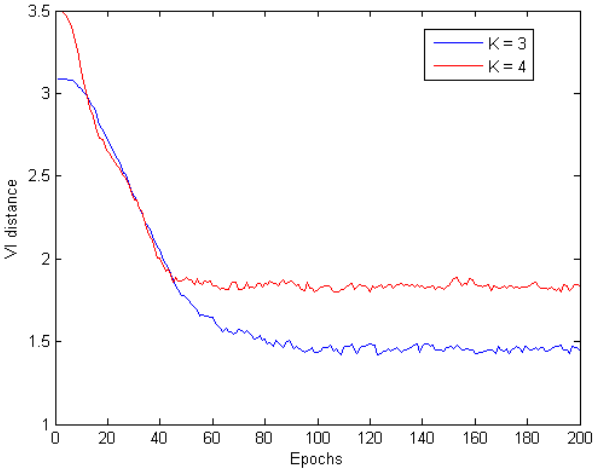
(a) Classification accuracy for $K = 4$ (b) VI distance for $K = 4$

Figure 2: Plots of clustering quality measures for Classic400 dataset

For same value of α and β, we can see in Figure 3 that when K is equal to number of given categories, then a better quality model is obtained. This difference is more prominent for VI-distance as shown in Figure 3(b).



(a) Classification accuracy for $\alpha = 1.25, \beta = 0.1$

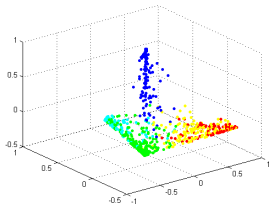


(b) VI distance for $\alpha = 1.25, \beta = 0.1$

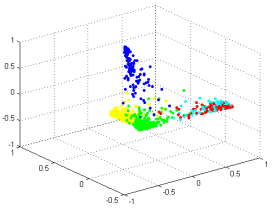
Figure 3: Plots of clustering quality measures for Classic400 dataset

5.2 BBC dataset

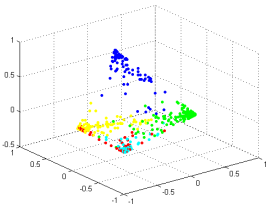
Figure 4 shows the plots of documents in topic space for BBC dataset. For $K >= 4$, we use PCA to reduce dimensionality of data to 3 dimensions. Clustering follows the same pattern as for Classic400 dataset i.e. for smaller α and comparatively larger β, clusters are more separated from each other.



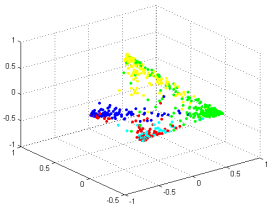
(a) $\alpha = 0.5, \beta = 0.1, K = 4$



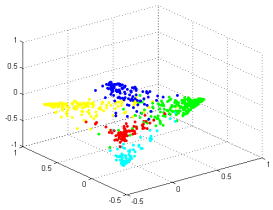
(b) $\alpha = 1.25, \beta = 0.1, K = 4$



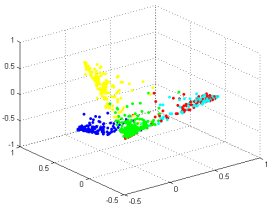
(c) $\alpha = 0.5, \beta = 1, K = 4$



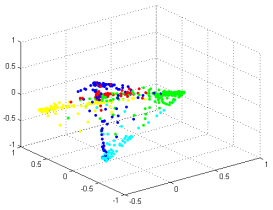
(d) $\alpha = 1.25, \beta = 1, K = 4$



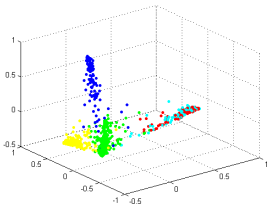
(e) $\alpha = 0.4, \beta = 0.1, K = 5$



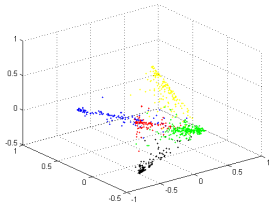
(f) $\alpha = 1, \beta = 0.1, K = 5$



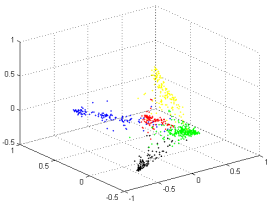
(g) $\alpha = 0.4, \beta = 1, K = 5$



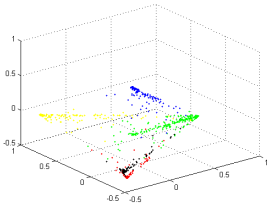
(h) $\alpha = 1, \beta = 1, K = 5$



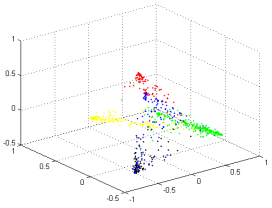
(i) $\alpha = 0.33, \beta = 0.1, K = 6$



(j) $\alpha = 0.83, \beta = 0.1, K = 6$



(k) $\alpha = 0.33, \beta = 1, K = 6$



(l) $\alpha = 0.83, \beta = 1, K = 6$

Figure 4: Plots of BBC dataset in LDA topic space

Table 2 shows the top frequency words for $K = 4, 5, 6$. For $K = 4$, we observe that the topics Tennis and Athletics have merged together. This can be due to the reason that both of them are individual sports and are

clustered together. For $K = 5$, we see top words in LDA topics resemble the natural classes given. For $K = 6$, we see that words from topic football split to two sub-topics. For K=5, count of words assigned topic as Football is almost double than that the count for other topics. This agrees with similar observation on Classic400 dataset as well that topics from bigger cluster divide into two subtopics.

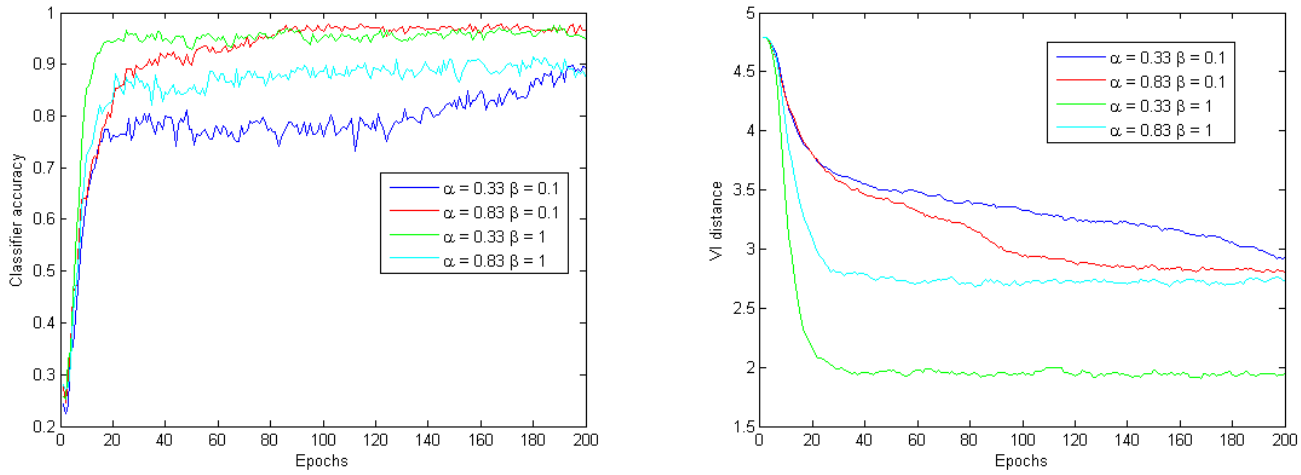
Rugby	Athletics	Cricket	Tennis	Football
england	second	test	plai	player
ireland	world	cricket	win	game
wale	minut	australia	first	plai
game	race	seri	open	club
against	olymp	pakistan	match	go
rugbi	win	plai	set	leagu
nation	european	first	england	chelsea
six	goal	india	two	want
plai	indoor	match	south	footbal
franc	athlet	tour	year	team

Football - A	Rugby	Football - B	Tennis	Athletics	Cricket
goal	england	player	plai	olymp	test
minut	ireland	game	open	world	cricket
unit	wale	club	win	athlet	england
arsen	game	plai	first	race	first
score	rugbi	go	match	test	plai
ball	against	want	set	indoor	seri
chelsea	nation	think	final	year	south
second	six	team	roddick	champion	australia
chanc	franc	footbal	year	athen	ball
refere	win	leagu	world	european	run

Table 2: Top 10 words for each LDA topic for K=4, K=5 and K=6 in BBC dataset

Athletics+Tennis	Cricket	Rugby	Football
win	test	england	game
world	cricket	ireland	player
year	england	wale	plai
set	plai	game	club
final	first	against	chelsea
open	seri	nation	arsen
olymp	south	rugbi	leagu
plai	australia	six	goal
second	ball	coach	unit
champion	run	franc	footbal

Figure 5 shows model quality measures for BBC dataset. All choices of hyperparamters obtain more than 85% accuracy. Small α and large β performs better both in classification accuracy and VI-distance.

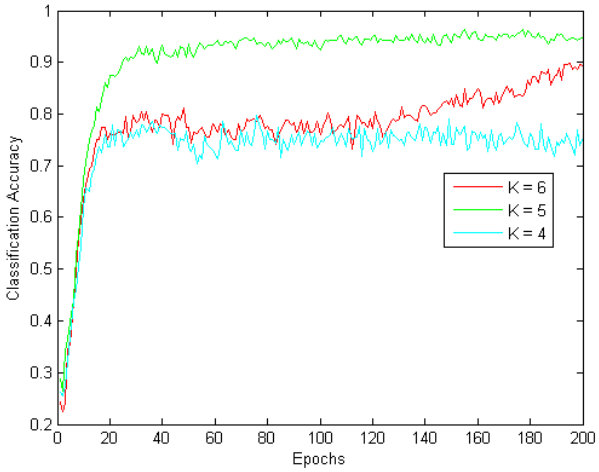


(a) Classification accuracy for $K = 6$

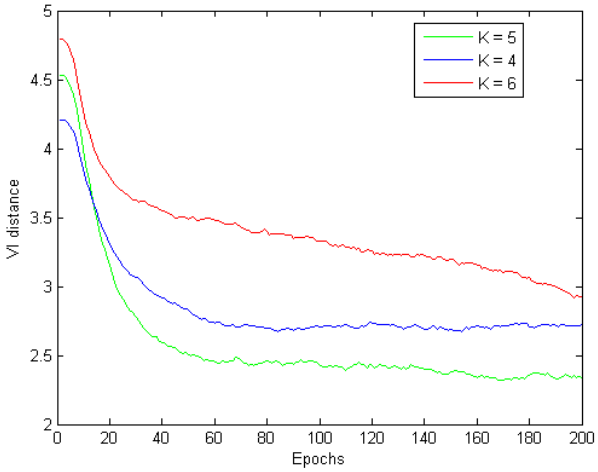
(b) VI distance for $K = 6$

Figure 5: Plots of clustering quality measures for BBC dataset

In Figure 6, better clustering performance is achieved when $K = 5$ i.e. number of LDA topics match the number of natural classes.



(a) Classification accuracy for $\alpha = 0.66, \beta = 0.1$



(b) VI distance for $K = 6$ $\alpha = 0.66, \beta = 0.1$

Figure 6: Plots of clustering quality measures for BBC dataset

6 Conclusion

References

[1] G. Doyle and C. Elkan: Accounting for Word Burstiness in Topic Models - In Proceedings of the 26th International Conference on Machine Learning (ICML), July 2009

[2] Gregor Heinrich. Parameter estimation for text analysis. http://www. arbylon.net/publications/text-est.pdf, 2005.

[3] Machine Learning Group Datasets, University College Dublin. http://mlg.ucd.ie/datasets/bbc.html.