CSCI-B555 MACHINE LEARNING

PROGRAMMING PROJECT 4

Experiments with LDA

In this programming project we implement Latent Dirichlet Allocation (LDA) and inspect its performance both in an unsupervised manner and when used as a preprocessing step for supervised learning. Your goals in this assignment are to (i) implement the collapsed Gibbs sampler for LDA inference, and (ii) compare the LDA topic representation to a "bag-of-words" representation with respect to how well they support document classification.

Task 1: Gibbs Sampling

In this portion, your task is to implement the collapsed Gibbs sampler for LDA. In the case of LDA, the output represents a sample of the (hidden) topic variables for each word. Recall that in LDA we sample the hidden topic variables associated with words in the text. This sample of topic variables can be used to calculate topic representations per document.

Results:

5 most frequent words from each topic. (*topicwords.csv*)

eliot	engines	diesels	cars	writes
edu	henry	writes	toronto	article
science	space	information	internet	nasa
space	nasa	long	work	gov
car	ford	bad	dealer	probe
station	launch	option	redesign	capability
don	even	make	want	find
sky	earth	temperature	satellite	good
edu	writes	article	apr	people
book	cost	saturn	buying	blue
two	time	high	large	used
hst	mission	pat	access	mass
edu	insurance	writes	geico	uiuc
shuttle	system	orbit	mars	mission
oil	come	service	used	time
power	speed	air	mph	turbo
engine	cars	feel	small	manual
car	clutch	shifter	sho	mustang
world	idea	once	don	city
bill	edu	gif	uci	ics

Conclusion:

Most of the words in each row are related to a particular topic. For example, in (Shuttle, system, orbit, mars, mission) all the words related to the similar topic. Similarly, in (power, speed, air, mph, turbo) also have all the words related to one topic. We also have many rows which have words not related to the main topic. This can be attributed to the random sampling. Even after that, out 5 there are at least 3 to 4 words which are related to the topic. For example, in (sky, earth, temperature, satellite, good) the first 4 words that are related to one topic. So yes, words in topics do make sense.

Task 2: Classification

In this portion we will evaluate the dimensionality reduction accomplished by LDA in its ability to support document classification and compare it to the bag of words representation.

Results:

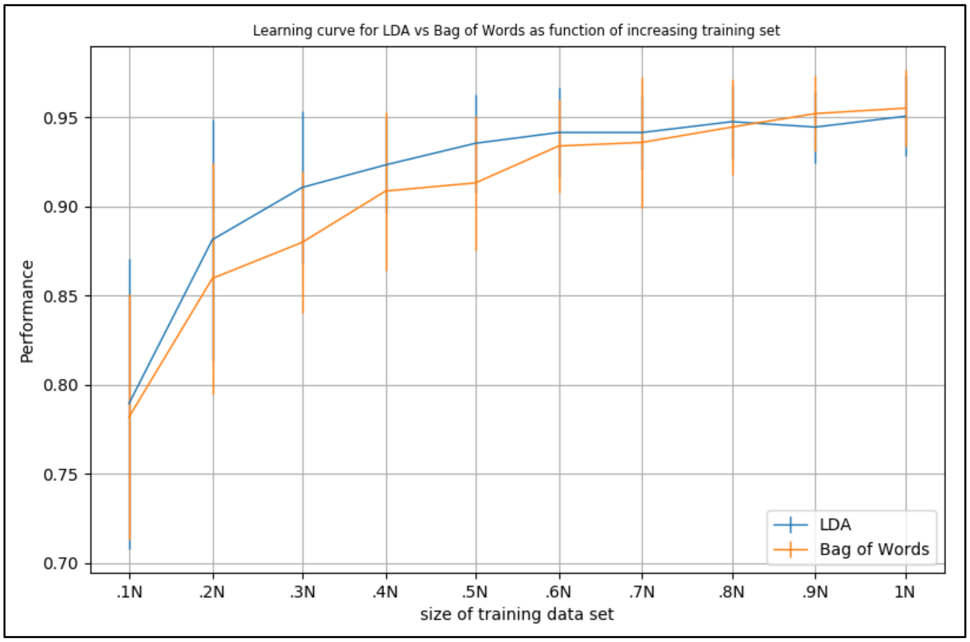

Figure1: Plots for LDA vs Bags of Words

Conclusion:

We can see from the above graph that, as long as the train size is small LDA performs better than the bag of words. But as the train size increases Bag of words starts producing better results. Reason for this behavior can be attributed to dimensionality reduction in LDA. LDA performs well when the train size is small and dimensionality reduction does not affect it much preserving the data integrity. However, as the train size increases Bag of words performs better and the effects of dimensionality reductions leads to misclassification of words to topic for LDA.