Machine Learning	Homework 5
CSCI 5622 Fall 2017	Due Time Dec 8, 2017

Submission Guidelines

- In order to download files required for the homework, clone `https://github.com/BoulderDS/csci_5622_hws`.
- For programming questions, submit python source files in a zip file.
- For other questions, submit a PDF file of no more than 16 pages.

All homework submissions are done through Moodle.

1 KMeans (35 points)

In this homework, you will use SVD and KMeans in sklearn to find topics in the *20newsgroup* dataset. For more information about the 20newsgroup dataset, check here `http://qwone.com/~jason/20Newsgroups/`. You are going to use *Vectorizer* in sklearn to vectorize text data, SVD to do dimensionality reduction, and KMeans to do clustering.

We are going to perform clustering on both words and documents:

- Words. We represent a word based on the documents in which it occurs (word-document matrix), reduce the dimensionality ($|V| \times K$), and then cluster words into topics (T clusters).
- Documents. We represent a document based on the words that it contains (document-word matrix), reduce the dimensionality ($D \times K$), and then cluster documents into clusters.

1.1 Programming questions (15 points)

Finish `KMeans.py`.

1. Finish the *DimensionalityReduction* class to reduce the dimension of data from vocabulary size to D (default n_components=100).
2. Finish the *KM* class to do clustering on the dimension-reduced data.

1.2 Analysis (20 points)

Try different reduced dimensions K (50, 100, 200) and number of clusters T (20, 50, 100) for words.

- What are the words in the same topic with "university" for different K and T? (Listing 10 words for each setting is sufficient.) (5 points)
- What are the words in the same topic with "jesus" for different K and T? (Listing 10 words for each setting is sufficient.) (5 points)

Try different reduced dimensions K (50, 100, 200) and cluster all documents into 20 clusters.

- Does the document clusters map to existing labels in 20newsgroup? What are the **three** most aligned existing label? (10 points) Hint: compute the number of news articles in cluster j for each existing label i.

2 LDA (35 points)

In this question, you are going to use LDA with variational inference. The functions and classes are completed for you. Your task is to run LDA on the 20newsgroup dataset. A toy dataset is provided so that you can learn how LDA works. You need to write your testing code on the main function and analyze the topics of 20newsgroup dataset.

2.1 Programming questions (10 points)

In `lda.py`:

1. Replace toy dataset with 20newsgroup dataset in main function. (you may need to use tokenizer.)

2.2 Analysis (25 points)

Run LDA with 20 topics.

1. Explain how functions in `lda.py` implement the variational inference algorithm and write down the update equations for different variables (10 points).
2. How do existing labels align with topics in 20newsgroup? (10 points) Hint: compute a matrix that is 20 by 20 and each entry holds the sum of topic probability of news articles with label i for a particular topic j.
3. Generate top 10 words in each topic (5 points).

3 Residual Error of PCA (30 pts)

Derive the residual error for PCA.

1. Prove that

$$\|\boldsymbol{x}_i - \sum_{j=1}^{K} z_{ij}\boldsymbol{v}_j\|^2 = \boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{j=1}^{K} \boldsymbol{v}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{v}_j$$

   Use the fact that $\boldsymbol{v}_j^T\boldsymbol{v}_j = 1$ and $\boldsymbol{v}_j^T\boldsymbol{v}_k = 0$ for $k \neq j$. Also, recall $z_{ij} = \boldsymbol{x}_i^T\boldsymbol{v}_j$

2. Now show that

$$J_K \triangleq \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{j=1}^{K} \boldsymbol{v}_j^T\boldsymbol{x}_i\boldsymbol{x}_i^T\boldsymbol{v}_j) = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i^T\boldsymbol{x}_i - \sum_{j=1}^{K}\lambda_j$$

   Hint: recall $\boldsymbol{v}_j^T\boldsymbol{C}\boldsymbol{v}_j = \lambda_j\boldsymbol{v}_j^T\boldsymbol{v}_j = \lambda_j$

3. If $K = d$ there is no truncation of dimension, so $J_d = 0$. Use this to show that the error from only using $K < d$ dimensions is given by

$$J_K = \sum_{j=K+1}^{d}\lambda_j$$

   Hint: partition the sum $\sum_{j=1}^{d}\lambda_j$ into $\sum_{j=1}^{K}\lambda_j$ and $\sum_{j=K+1}^{d}\lambda_j$