Homework 7: Dimensionality Reduction **CSCE 633**
Due: 11:59pm on April 26, 2024

Instructions for homework submission
a) Please write a brief pdf report including your experiment details, results and discussion; **submit the pdf file on Canvas**. Also **submit a jupyter notebook** including your code.
b) Make sure TAs can run your code on Google Colab. For each question, please explain your thought process, results, and observations in a markdown cell after the code cells. Please do not just include your code without justification.
c) **You can use any available libraries for this homework.**
d) Please start early :)

Part A: Feature Selection

(a) Load Data Load the Iris dataset from sklearn.

(b) Data Exploration What are variable names of the features and target?

(c) Feature Comparison

Option 1: Make two scatter plots of data points in the Iris dataset, using two feature combinations as axes respectively: (i) sepal width and length; (ii) petal width and length. Colors of points in scatter plots indicate their flower type. Are all classes (flower types) easily identifiable based on one of these feature combinations?

Option 2: For convenience, if you are not willing to plot and analyse using sklearn data directly as above, feel free to make a pandas dataframe (columns are features and target) using the data instead. If so, plot figures using *seaborn.violinplot* to find out differences between flower species based on each of the different features.

Finally tell me, are the different flower types distinguished more easily by their sepal or petal features? Which features do you prefer to select?

Part B: Feature Extraction

(a) Example We are still using Iris Dataset as an example here.

(1) Apply PCA to the dataset, taking original feature variables and finding three principal components.

(2) Now make a 3D scatter plot, visualizing data points across the first three PCA components. Then check the explained variances and explained variance ratios for each of the three components. What can you infer from these results? Do you still think we need three principal components?

(3) If you decide to change the number of components, please apply PCA to original Iris dataset again, and show that there is still an efficient separation between different target classes (flower types).

(b) Load Data Let's apply PCA to word embeddings in NLP. Download GloVe embeddings from link. In this assignment, we use *glove.6B.200d.txt*.

(1) Write a function *loadGlove* to load the GloVe data, with txt file as input, so that you can access the related vector for a word. You may consider writing and returning a dictionary, with words and related embedding vectors as key-value pairs.

(2) Now take a look at two words - "man" and "woman", in order to test if your function works and see how words are represented.

(c) Pick Data Print all words in GloVe in a list. Then select about 12 - 15 of them of interest, and access their corresponding vectors. Note that some of the words you pick must share something in common. Examples are as below:

- 'man', 'woman', 'boy', 'girl', 'father', 'father-in-law', 'mother', 'mother-in-law', 'uncle', 'aunt', 'lord', 'lady', 'king', 'queen', 'prince', 'princess', 'waiter', 'waitress'
- 'people', 'team', 'group', 'public', 'police', 'military', 'department', 'government', 'president', 'united', 'states', 'u.s.', 'city', 'state', 'capital', 'country'
- 'monday', 'tuesday', 'wednesday', 'thursday', 'friday', 'week', 'month', 'season', 'year', 'home', 'work', 'house', 'school', 'university', 'college', 'company', 'office'

(d) Principal Component Analysis

(1) Apply PCA (number of components = 10) to vectors of your chosen words, then check the explained variance ratios.

(2) If you think you can adjust the number of components or change to another set of words, you may give it a try.

(3) Eventually, plot the results in a 3D figure, with each of the first three components represented by one of three axes.

(e) Result Analysis What can you see from the results above? Does PCA actually work? Are the relationships between the chosen words successfully visualized?