STA 221: Homework 1

- Homework due in Canvas: 05/01/2020 at 11:59PM. Please follow the instructions provided in Canvas about homeworks, carefully.

1. **Basics (6 points)**. Answer true or false for each of the question below and give justification.

   (a) A rectangular matrix of size $n \times m$ is a *linear* transformation.

   (b) Only square matrices have Eigenvalue decompositions.

   (c) `Power Method` can be used to find only eigenvectors (and not singular vectors).

   (d) Singular vectors are orthogonal to each other.

   (e) Kernel PCA is a linear dimension reduction technique.

   (f) Spectral Clustering is a non-linear dimension reduction technique.

2. **Python practice via statistical concepts (10 points)**. This questions helps you brush-up your numerical computing skills in python by implementing some basic statistics concepts. Let X be a random variable that takes values $+1, -1$ with equal probability. That is:

$$P(X = +1) = P(X = -1) = 1/2.$$

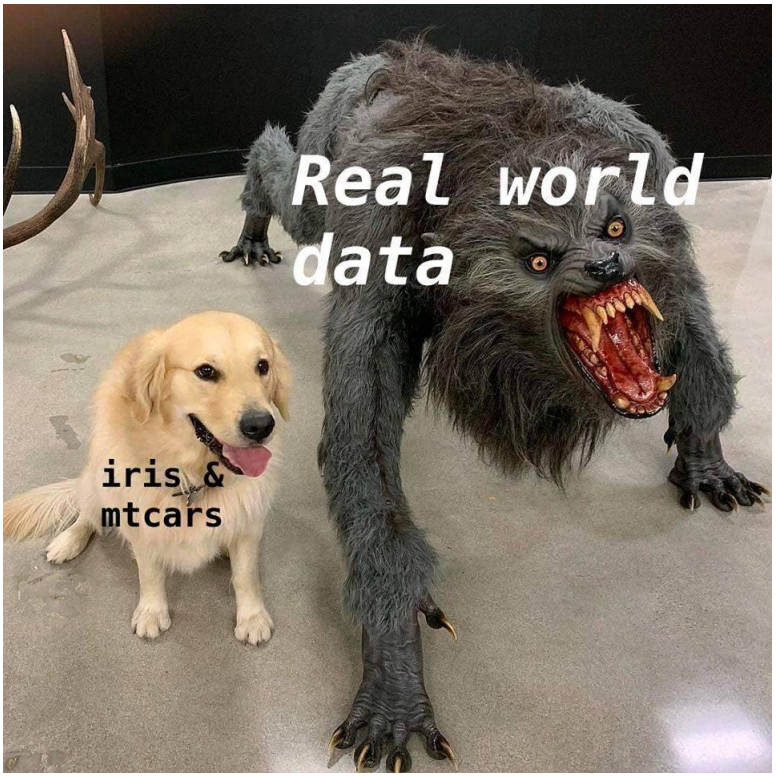
   Generate $N = 10,000$ datasets, each of which has n data points. For this simulation, we consider $n = \{10, 100, 1000, 10000\}$. (Hint: Write a function that samples from the uniform distribution between 0 and 1. If the result is less than 0.5, set it to -1. Otherwise, set it to 1). Let $\overline{X}_n^{(i)}$ be the sample average of i^{th} dataset, $\mu = E(X) = 0$ and $\sigma^2 = \text{Var}(X) = 1$.
   (Hint: Once you compute the sample averages, you will not need the individual data points from each dataset. Therefore, to save memory, you need only store the $\overline{X}_n^{(i)}$ rather than all the data points. It is highly recommended that you do this to avoid freezing or crashing your computer). Plot and intepret the following:

   (a) $\log_{10}(n)$ v.s. $\overline{X}_n^{(1)} - \mu$;
   (Hint: This plot illustrates how the deviation $\overline{X}_n^{(1)} - \mu$ converges to 0 as n increases).

   (b) Draw $\log_{10}(n)$ v.s. $\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{|\overline{X}_n^{(i)} - \mu| > \epsilon\}$ for $\epsilon = 0.5, \epsilon = 0.1, \epsilon = 0.05$;
   (Hint 1: This plot illustrates the convergence of empirical averages to true expectation.)
   (Hint 2: For some statement S, the indicator function $\mathbb{1}\{S\}$ is defined as $\mathbb{1}\{S\} = 1$ if S is true and $\mathbb{1}\{S\} = 0$ otherwise.)

   (c) Draw histograms of $\sqrt{n}(\overline{X}_n^{(i)} - \mu)/\sigma$ for N datsets for $n = 10, n = 1,000, n = 10,000$. You may choose your histogram bins or you may let Python choose automatically–any meaningful plot will do.
   (Hint: This plot illustrates the Central Limit Theorem.)

3. **Amazon Review Analysis (14 points)** In class we used IRIS data to visualize several dimensionality reduction techniques. But the difference between such datasets and real-data could be summarized nicely as below:



In this question, we will take the raw reviews from Amazon and go through several steps to extract Document-Term matrix and TF-IDF matrix representation of the documents (each review is defined as a document). This process will result in a matrix of size `number of documents` x `number of words in our dictionary considered`. After this, we will try out different dimension reduction techniques on this dataset.

(a) The dataset `Amazon.RData` consists of real reviews of different products in Amazon. Unfortunately, it is provided to you in `Rdata` format (the preferred data format for `R programming language`). This scenario is quite common in practice. To process this data, you need to load the data in python. In order to proceed, install `pyreadr` package in Python. Note: to install with pip, use `pip install pyreadr`. You are welcome to explore any other ways of importing this data into Python. After loading the data, you will use only the `review` field in this question.

(b) The next pre-processing step we consider is called as stemming. This process fixes the words in our dictionary. For example, some common words (like 'the', 'and') are ignored, numbers are ignored, the 'root' of the word is used (i.e., running, ran are all treated as a single word run). To perform this step, execute the following commands:

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.feature_extraction.text import CountVectorizer
from nltk.stem.snowball import FrenchStemmer
stemmer =FrenchStemmer()
analyzer= CountVectorizer().build_analyzer()
def stemmed_words(doc):  return (stemmer.stem(w) for w in analyzer(doc))
```

Note that above we used FrenchStemmer. Feel free to explore other options depending on the actual text you see.

(c) Now we will extract the Document-Term matrix of this dataset.This will help build a matrix where each row represents one document and each column represents a different word. To do so, use the command `CountVectorizer`. For the analyzer, use `stemmed_words` option. This will give you a matrix which corresponds term counts in each document.

(d) Next, we will extract the TF-IDF matrix of this dataset. To do so, use the command `TfidfVectorizer` with the option `token_pattern='[a-z]{3,15}'`. This will give you an alternate representation of the same dataset.

(e) How many rating values are present in the dataset ? How many reviews of each rating value are there in the entire dataset? You can think of the dataset as having roughly as many clusters as the number of rating values.

(f) Perform, (i) PCA and (ii) kernel PCA on both Document-Term representation and TF-IDF representation, all with number of components being set to 2. Note that, you might have to set other parameters as well for some of the above methods – you are welcome to explore different options. Produce the best figure (for each of the above method) that identifies the cluster structure (if it can) after dimension reduction (to 2 dimensions).

(g) Now, perform (i) k-means clustering and (ii) spectral clustering on both Document-Term representation and TF-IDF representation of the data. Is the cluster assignment you obtained, consistent with the true rating of each document ?

Pledge:

Please include the following pledge and sign it before you submit your assignment in canvas. If you can not honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

- We pledge that we are honest students with academic integrity and we have not cheated on this homework.
- These answers are our own work.
- We did not give any other students assistance on this homework.
- We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
- We understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of "F" for the course.

Team Member 1 Team Member 2