CS328: Homework 2

Due date 17/2/2019 @11:59 pm

Submitting a Jupyter notebook is preferable for each of the coding questions. Copying code is not allowed, from others or any sources. Discussion is fine, but please write down names of people you discussed with.

---

1. (5 points) Suppose you are given a matrix A. You have calculated SVD of A to be $A = U\Sigma V^t$. You claim that the matrix UU^t is the *projection matrix* for the subspace spanned by the columns of A. However, your roommate was working all night and claims that s/he has a new projection matrix that serves the same purpose– $(A^tA)^{-1}A^t$. How can this be?

2. (5 points) Suppose you are trying to define an "inverse" for a $n \times n$ matrix A that is not necessarily invertible. In a desperate attempt, you resort to taking the SVD $A = U\Sigma V^t$. And then define something like an inverse $A^+ = \sum_{i \leq rank(A)} \frac{1}{\sigma_i} v_i u_i^t$. Note the flips of u and v. We will check whether this makes sense. What are i) AA^+ ii) A^+A iii) AA^+A.

3. (10 points) The MNIST is a set of images of handwritten set of digits. We will provide you sample images of every digit (unlabelled) as training as well as a sample of it as test. You task is the following:

   (a) Using the sklearn library function, apply k-means on the digits, with $k = 10$, say. Do not use labels even if available. Use the `init='random'` initializer. You should get 10 clusters. Calculate accuracy of your clustering to be as follows: using the labels of the training set, take a majority vote and name each cluster to be one of the digits. Note that in this process, depending on the clustering, multiple clusters could be named the same digit, and some digit may not be assigned as a cluster name. Now calculate the accuracy as follows: for each image from the test set, assign it to the nearest cluster center and give it the corresponding label of that cluster. Find out the accuracy, i.e. the fraction of test set that is misclassified.

   (b) Run the same exercise, but now first take the low-rank approximation of the training data. Try out $k = 2, 4, 16, 32$. Report the accuracies for each choice of k.

4. (5 points) If P represents the projection matrix onto a subspace of rank-k, lying in R^n, what are the singular values of A? What is the physical interpretation?

5. (5 points) Suppose matrix A is a database of restaurant ratings: each row is a person, each column is a restaurant, and an entry (i, j) represents how much person i likes restaurant j. What might v_1 represent? What about u_1 ? How about the gap $\sigma_1 - \sigma_2$ (speculation is fine here)?

6. (10 points) The directed webgraph of Stanford is available here: https://snap.stanford.edu/data/web-Stanford.html. Download it, and take the largest weakly connected component (feel free to use library functions in SNAP or networkX, or any other package with citation). Using only the largest component run the following pseudocode, i.e. the HITS algorithm for 10 iterations. Report the top 5 hubs and authorities along with their scores (2 digits of precision is enough).

   (a) Initialize h and a to be all ones.

   (b) Run HITs algorithm for 10 iterations.