CS4786/5786: Machine Learning for Data Science, Spring 2016
09/01/2016: Assignment 1: PCA, CCA and Random Projections

Instructions Due at 11:59pm September 9th on CMS. Submit what you have at least once by an hour before that deadline, even if you haven't quite added all the finishing touches — CMS allows resubmissions up to, but not after, the deadline. If there is an emergency such that you need an extension, contact the professor. You have a slip day of at most one day for the assignment. The assignment is to be done individually. You will submit both a writeup and some datafiles you create. The writeup can be handwritten or typeset, but please make sure it is easily readable either way. Keep an eye on the course webpage for any announcements or updates.

Academic integrity policy We distinguish between "merely" violating the rules for a given assignment and violating *academic integrity*. To violate the latter is to commit *fraud* by claiming credit for someone else's work. For this assignment, an example of the former would be getting an answer from person X but stating in your homework that X was the source of that particular answer. You would cross the line into fraud if you did not mention X. The worst-case outcome for the former is a grade penalty; the worst-case scenario in the latter is academic-integrity hearing procedures.

The way to avoid violating academic integrity is to always document any portions of work you submit that are due to or influenced by other sources, even if those sources weren't permitted by the rules.[1]

---

*Q1 (***PCA).**
A trick often used as preprocessing in machine learning approaches is to take a multi-dimensional dataset and normalize it such that every coordinate has variance of 1. So, if $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is the original set of n points in d dimensions each (and here we assume that the points are centered, meaning that $\frac{1}{n}\sum_{t=1}^{n} \mathbf{x}_t = 0$). Then, we can obtain the normalized dataset $\mathbf{x}'_1, \ldots, \mathbf{x}'_n$ by setting for each $t \in \{1, \ldots, n\}$ and for each $i \in \{1, \ldots, d\}$,

$$\mathbf{x}'_t[i] = \frac{1}{\sqrt{\frac{1}{n}\sum_{s=1}^{n} \mathbf{x}_s[i]^2}} \mathbf{x}_t[i]$$

In this question we explore the effect of such pre-processing on PCA algorithm.

1. Generate a 20-dimensional dataset say X consisting of 100 points that are centered and are such that the direction of projection from PCA (with $K = 1$) **remains unchanged** when we renormalize the dataset so that each coordinate is made to have variance 1.

[1] We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials. To minimize documentation effort, we also do not expect you to credit the course staff for ideas you get from them, although it's nice to do so anyway.

2. Generate a 20-dimensional dataset say XX consisting of 100 points that are centered and are such that the direction of projection from PCA (with $K = 1$) **changes** when we renormalize the dataset so that each coordinate is made to have variance 1.

First, **explain you answer conceptually. Describe mathematically or in words (but precisely) why in the first case the projections remain unchanged and why and how the projections change in the second case.** You can also use scatter plots of the one dimensional projections before and after renormalization in both the cases above to visually demonstrate the success of your solution. To make the visual clear you might want to color points red or blue and show how the scatter plots in the first case are same and in the second differ. The scatter plots are not necessary as we will on our end check first your explanation and second the changes in projections for the points you provide ourselves.

Submit the generated points to cms as two csv (comma separated values) files named `X.csv` and `XX.csv`. The files should each consist of 100 lines, one for each data-point. Each line in turn should consist of 20 numbers separated by ',' symbols.

*Q2 (***CCA).** The goal of this question is to get a better understanding of CCA, specifically the scale free nature of CCA.

Generate a 100-dimensional dataset consisting of 1000 points. We shall consider the first 50 coordinates of this dataset as view 1 and next 50 to be view 2. The data points should be such that:

1. When we perform CCA with $K = 1$ and plot the points on a line (in each of the views), the first 500 points and last 500 points are well-separated.

2. If instead, on each of the views, we first perform PCA to reduce dimensionality to 49 and then subsequently perform CCA with $K = 1$ using these two 49 dimensional (reduced dimensional) views, the first 500 points and last 500 points should not be separable.

When we use the term separable, we mean that if we scatter plot the 1000 points, and color the first 500 red and the next 500 blue, the red points should form one separate blob and the blue points another separate blob.

First, **explain your answer mathematically or precisely in words. Tell us why your solution should work.** Also **submit you** 1000**,** 100 **dimensional points in file `XCCA.csv` as comma separated value files.** That is as 1000 lines each consisting of 100 numbers separated by commas. If you like, you may include scatter plots you created in your writeup.