CS4786/5786: Machine Learning for Data Science, Spring 2016
09/13/2016: Assignment 2: Random Projections and Kernel Method

---

**Instructions** Due at 11:59pm September 20th 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.

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-cartesianse 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 (***Random Projections).**

Generate a data set consisting of $100$ $1000$-dimensional points. For each point $\mathbf{x}_t$ in the data set, ensure that their norm (distance to $\mathbf{0}$) is exactly $1$. We shall perform PCA and random projections on this data set. To evaluate our projections we shall use the following metric on how well each projection preserves distances:

$$\mathrm{Err}(\mathbf{y}_1, \ldots, \mathbf{y}_n, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{2}{n(n-1)} \left| \sum_{t=1}^{n} \sum_{s=t+1}^{n} \left( \|\mathbf{y}_t - \mathbf{y}_s\|_2 - \|\mathbf{x}_t - \mathbf{x}_s\|_2 \right) \right|$$

You shall pick $K = 1$ and perform PCA and random projections on the data set. Your task in this problem is to create the data sets such that **the** $\mathrm{Err}$ **of Random Projection is much smaller compared to that of PCA.** (the $\mathrm{Err}$ of PCA should be at least $0.9$ more than the $\mathrm{Err}$ of random projection).

[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.

Submit your dataset as csv files `RpBeatsPCA.csv` (so, the file should consist of 100 lines, each with thousand numbers separated by commas). Also, in your assignment writeup, explain how you generated the data set and the rationale behind this choice. Specifically, give the mathematical intuition with some formalism as to why the error in PCA will be much higher than that of random projections. Your rationale should explain how you used the properties of what RP and PCA produce to guide your thinking.

*Q2 (***Kernel Method).**
We shall consider here the case (for simplicity) when dimensionality of the original data set is $d = 1$. Show that the function

$$k(x, y) = \exp\left(\frac{x \cdot y}{\sigma}\right)$$

is a valid kernel function for $\sigma \in [0, 1]$. That is, there exists a feature space mapping $\Phi$ (mapping to possibly infinite dimensional space) such that, $k(x, y) = \Phi(x)^\top \Phi(y)$. Hint: use the power expansion of the exponential function