CS4786/5786: Machine Learning for Data Science, Spring 2016
3/14/2016: Diagnostic Assignment P1

Instructions Due at 11:59pm Thursday, March 17th 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. No collaborations on this assignment, do it individually. Submit your writeup on CMS.

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.

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]

[1] We make an exception for sources that can be taken for granted in the instructional setting, namely, the course materials.

Q1 For each of PCA, CCA and random projection, suggest a heuristic for choosing K, the number of dimensions to project data down to. The heuristic should be such that you don't loose too much information but on the other hand, K should not be too large. Explain your answer for how you choose the K and why you think the heuristic is a good one. We are looking for a heuristic so you don't need to provide precise mathematical formulation for your heuristic Hint: the heuristic could depend on the data and for non-degenerate cases lead to non-trivial choice of K (ie. not lead to $K = 1$ or $K = d$).

Q2 Recall that the K-means objective is to minimize:

$$M_1(C_1, \ldots, C_K) = \sum_{j=1}^{K} \sum_{\mathbf{x}_t \in C_j} \|\mathbf{x}_t - \mathbf{r}_j\|_2^2$$

where $\mathbf{r}_j$ is the centroid of C_j defined as $\mathbf{r}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_t \in C_j} \mathbf{x}_t$. Show that for any cluster assignment, this objective of minimizing M_1 is equivalent to the minimizing the objective,

$$M_2(C_1, \ldots, C_K) = \sum_{j=1}^{K} \frac{1}{|C_j|} \sum_{\mathbf{x}_s, \mathbf{x}_t \in C_j} \|\mathbf{x}_s - \mathbf{x}_t\|_2^2$$

Specifically, show that for any cluster assignment $C_1, \ldots, C_K$,

$$M_1(C_1, \ldots, C_K) = \frac{1}{2} M_2(C_1, \ldots, C_K)$$

Q3 Let us consider the problem of clustering nodes of a graph into K groups. Say the graph given to us is $G = (V, E)$ where V is the set of vertices and E is the set of edges. Show that the objective of finding a K-clustering of the nodes of a graph that

"minimizes the number of edges between clusters"

is equivalent to finding a K-clustering that

"maximizes number of edges within the clusters"

Can you suggest an algorithm that produces a clustering of nodes that (perhaps approximately) optimizes the above, equivalent objectives?