CS 3190: Foundations of Data Analysis **Due: Friday 12.07 at 11:59pm**

Homework 5: Clustering and Classification

Instructions: Your answers are due at 11:59pm on the due date. You must turn in a pdf through canvas. I recommend using latex (`http://www.cs.utah.edu/~jeffp/teaching/latex/`) for producing the assignment answers. If the answers are too hard to read you will lose points, entire questions may be given a 0 (e.g. **sloppy pictures with your phone's camera are not ok, but very careful ones are**)

Please make sure your name appears at the top of the page.

You may discuss the concepts with your classmates, but write up the answers entirely on your own. **Be sure to show all the work involved in deriving your answers! If you just give a final answer without explanation, you may not receive credit for that question.**

We will use two datasets, here: `https://www.cs.utah.edu/~zhe/data/P.csv` and here: `https://www.cs.utah.edu/~zhe/data/Q.csv`
There are many ways to import data in python (see Canvas for a discussion). The `pandas` package seems to be the most general one.

1. [**40 points**] Download data sets P and Q. Both have 120 data points, each in 6 dimensions, can be thought of as data matrices in $\mathbb{R}^{120\times 6}$. For each, run some algorithm to construct the k-means clustering of them. You can use your own or third-party implementation of k-means. Diagnose how many clusters you think each data set should have by finding the solution for k equal to 1, 2, 3, $\ldots$, 10.

2. [**20 points**] Construct a data set X with 5 points in $\mathbb{R}^2$ and a set S of $k = 3$ sites so that k-means algorithm will have converged, but there is another set S', of size $k = 3$, so that $\textsf{cost}(X, S') < \textsf{cost}(X, S)$. Explain why S' is better than S, but that k-means algorithm will not move from S.

3. [**10 points**] Consider a "loss" function, called an *double-hinged loss function*

$$\ell_i(z) = \begin{cases} 0 & \text{if } z > 1 \\ 1 - z & \text{if } 0 \leq z \leq 1 \\ 1 & \text{if } z \leq 0. \end{cases}$$

   where the overall cost for a dataset (X, y), given a linear function $g(x) = \langle (1, x), \alpha \rangle$ is defined $\mathcal{L}(g, (X, y)) = \sum_{i=1}^{n} \ell_i(y_i \cdot g(x_i))$.

   (a) What problems might a gradient descent algorithm have when attempting to minimize $\mathcal{L}$ by choosing the best α?

   (b) Explain if the problem would be better or worse using stochastic gradient descent?

4. [**30 points**] Consider a data set (X, y), where each data point $(x_{1,i}, x_{2,i}, y_i)$ is in $\mathbb{R}^2 \times \{-1, +1\}$. Provide the psuedo-code for the Perceptron Algorithm using a polynomial kernel of degree 2. You can have a generic stopping condition, where the algorithm simply runs for T steps for

some parameter T. *(There are several correct ways to do this, but be sure to explain how to use a polynomial kernel clearly.)*