CS 4964: Foundations of Data Analysis **Due: Wednesday 11.01 at 2:45pm**

Homework 3: Linear Regression and Gradient Descent

Instructions: Your answers are due at 2:45, before the beginning of class 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 a dataset found here: `http://www.cs.utah.edu/~jeffp/teaching/FoDA/D3.csv`
There are many ways to import data in python. The `pandas` package seems to be the best one.

1. **[50 points]** Let the first column of the data set be the explanatory variable `x`, and let the fourth column be the dependent variable `y`. [That is: ignore columns 2 and 3 for now]

   (a) [10 points] Run simple linear regression to predict `y` from `x`. Report the linear model you found. Predict the value of `y` for new `x` values 0.3, for 0.5, and for 0.8.

   (b) [10 points] Use cross-validation to predict generalization error, with error of a single data point (`x`, `y`) from a model M as $(M(\mathtt{x}) - \mathtt{y})^2$. Describe how you did this, and which data was used for what.

   (c) [20 points] On the same data, run polynomial regression for $p = 2, 3, 4, 5$. Report polynomial models for each. With each of these models, predict the value of `y` for a new `x` values of 0.3, for 0.5, and for 0.8.

   (d) [10 points] Cross-validate to choose the best model. Describe how you did this, and which data was used for what.

2. **[25 points]** Now let the first three columns of the data set be separate explanatory variables `x1`, `x2`, `x3`. Again let the fourth column be the dependent variable `y`.

   - Run linear regression simultaneously using all three explanatory variables. Report the linear model you found. Predict the value of `y` for new (`x1`,`x2`,`x3`) values $(0.3, 0.4, 0.1)$, for $(0.5, 0.2, 0.4)$, and for $(0.8, 0.2, 0.7)$.
   - Use cross-validation to predict generalization error, with error of a single data point (`x1`, `x2`, `x3`, `y`) from a model M as $(M(\mathtt{x1}, \mathtt{x2}, \mathtt{x3}) - \mathtt{y})^2$. Describe how you did this, and which data was used for what.

3. **[25 points]** Consider two functions

$$f_1(x, y) = (x - 5)^2 + (y + 2)^2 \qquad f_2(x, y) = (1 - (y - 4))^2 + 35((x + 6) - (y - 4)^2)^2$$

Starting with $(x, y) = (0, 0)$ run the gradient descent algorithm for each function. Run for T iterations, and report the function value at the end of each step.

(a) First, run with a fixed learning rate of $\gamma = 0.5$.

(b) Second, run with any variant of gradient descent you want. Try to get the smallest function value after T steps.

For f_1 you are allowed only $T = 10$ steps. For f_2 you are allowed $T = 100$ steps.

[**+5 points**] *If any students do significantly better than the rest of the class on f_2 in part (b), we will award up to 5 extra credit points.*