CS 3190: Foundations of Data Analysis **Due: Wednesday 10.28 at 1:10pm**

Homework 3: Regression and Gradient Descent

Instructions: Your answers are due at 1:10, 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/`, see also `http://overleaf.com`) 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 found here:
`http://www.cs.utah.edu/~jeffp/teaching/FoDA/x.csv`
`http://www.cs.utah.edu/~jeffp/teaching/FoDA/y.csv`
There are many ways to import data in python, the `genfromtext` command in numpy provides an easy solution.

1. **[50 points]** Let $\mathtt{x} \in \mathbb{R}^n$ hold the data for an explanatory variable, and $\mathtt{y} \in \mathbb{R}^n$ be the data for the dependent variable. Here $n = 100$.

   (a) [10 points] Run simple linear regression to predict `y` from `x`. Report the linear model you find. Predict the value of y for the new x values of 4 and 8.5.

   (b) [10 points] Split the data into a training set (the first 80 values) and the test set (the last 20 values). Run simple linear regression on the training set, and report the linear model. Again predict the y value at x value of 4 and of 8.5.

   (c) [15 points] Using the testing data, report the residual vector (it should be 20-dimensional) for the model built on the full data, and another one using the model built just from the training data. Report the 2 norm of each vector.

   Also compute the 2-norm of the residual vector for the training data (a 80-dimensional vector) for the model build on the full data, and also for the model built on the training data.

   (d) [15 points] Expand data set `x` into a $n \times (p + 1)$ matrix $\tilde{X}_p$ using standard polynomial expansion for $p = 3$. Report the first 3 rows of this matrix.

   Build and report the degree-3 polynomial model using this matrix on the training data.

   Report the 2 norm of the residual vector built for the testing data (from a 20-dimensional vector) and for the training data (from a 80-dimensional vector).

2. **[25 points]** Reginald has input data (X, y) where $X \in \mathbb{R}^{n\times n}$ and $y \in \mathbb{R}^n$ and the columns of X are linearly independent.

   (a) [5 points] What is the span of the columns of X?

(b) [5 points] A matrix is invertible if and only if it is square and has full rank. Is X invertible?

(c) [10 points] Reginald fits an ordinary least squares model to his entire data set. In particular, he correctly finds the $\alpha \in \mathbb{R}^n$ minimizing $\|X\alpha - y\|_2^2$ and calls it $\hat{\alpha}$. To assess the goodness of fit of his model, he then computes $\|X\hat{\alpha} - y\|_2^2$ (which is the sum of squared errors (SSE)). What is $\|X\hat{\alpha} - y\|_2^2$?

(d) [5 points] Reginald concludes based on the value of $\|X\hat{\alpha} - y\|_2^2$ that the independent variables in his X matrix are very useful in predicting his dependent variable y. Is he necessarily correct? Why or why not?

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

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

(a) Run gradient descent on f_1 with starting point $(x, y) = (2, 3)$, $T = 20$ steps and $\gamma = .05$. Report the function value at the end of each step

(b) Run gradient descent on f_2 with starting point $(x, y) = (0, 2)$, $T = 100$ steps and $\gamma = .0015$. Report the function value at the end of each step

(c) Run any variant of gradient descent you want for f_1. Try to get the smallest function value after $T = 20$ steps

(d) Run any variant of gradient descent you want for f_2. Try to get the smallest function value after $T = 100$ 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. To obtain extra points, a detailed description of how the gradient descent is performed is required.*