**CS 3190: Foundations of Data Analysis** **Due: Tuesday 10.26 at 11:50pm**

# Homework 3: Regression and Gradient Descent

---

**Instructions:** Your answers are due at 11:50, 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 8 and 12.

   (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 8 and of 12.

   (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 $p = 4$. Report the first 3 rows of this matrix.

   Build and report the degree-4 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]** It is the summer of 2021 and the school of computing is prepping to teach CS 3190. They have a roster of the 1000 students in the course and want to predict how well the students will do. Oksana is in charge.

For the 1000 students that took CS 3190 in 2020, she collects their final grade for the course, and their final grade in CS 1410, Math 2270, CS 2100 (all grades are collected on a scale of 0 - 100). Imagine also every University of Utah student is required to generate a uniformly distributed number between -50 and 50 everyday, and enter their name and that number into a security application in order to gain access to their email.

Oksana also collects the random number that each student had over the last 96 days. This results in a dataset $X \in \mathbb{R}^{1000\times 100}$ **where each row is for a student that took CS 3190 in 2020**, the first column is all 1s for the intercept, the second, third, and fourth columns are the CS 1410, Math 2270, CS 2100 grades respectively, and the remainder of the columns give the security number for each student over the last 96 days (where each column is one of these days). In addition, Oksana collects $y \in \mathbb{R}^{1000}$ giving the grades of the students in CS 3190 in fall 2020.

(a) [8 points] Oksana fits a linear model using only the first 4 columns, and then another using all 100 columns on the data from the class of 2020 students. Which model will have smaller SSE (sum of squared error)?

(b) [8 points] Now Oksana collects $X^{new} \in \mathbb{R}^{1000\times 100}$ for the roster of students planning to take CS 3190 in the fall of 2021. Note that the none of the incoming students know any of the students that have taken the course in the past, and nothing at all has changed about the format of the course between 2020 and 2021. To predict their performance in the course, she needs to choose either the 4 column model or the 100 column model. Which model would you tell her to use and why?

(c) [9 points] Oksana is skeptical. In 2022 the same prediction task needs to be done for CS 3190. How could you leverage the now complete 2021 data (which includes $X^{new} \in \mathbb{R}^{1000x100}$ and $y^{new} \in \mathbb{R}^{100}$ which are the scores of the 2021 students) to convince Oksana to again use the model form you chose in part b?

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

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

For (a) - (d) below, you also need to report the function value, gradient, and 2-norm of the gradient at the end of each iteration – one set of values per line. This is a common way to see if your results make sense, and is not hard to quickly read through.

(a) [5 points] Run gradient descent on $f_1$ with starting point $(x,y) = (0,4)$, $T = 25$ steps and $\gamma = .01$

(b) [5 points] Run gradient descent on $f_2$ with starting point $(x,y) = (-10,6)$, $T = 100$ steps and $\gamma = .001$

(c) [7 points] Run any variant of gradient descent you want for $f_1$ with starting point $(x,y) = (0,4)$. Try to get the smallest function value after $T = 25$ steps. Also explain in a couple of sentences the procedure you used.

(d) [8 points] Run any variant of gradient descent you want for $f_2$ with starting point $(x,y) = (-10,6)$. Try to get the smallest function value after $T = 100$ steps. Also explain in a couple of sentences the procedure you used.

**[+5 points]** *If any students do significantly better than the rest of the class on $f_2$ in part (d), 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.*