EE5239 Optimization Homework 3 Notes

In this note I provide a few pointers about the homework. Here I just want to help set up the problem. Please see the following:

1. Hope that you already got access to the data and can import and plot them as well. If not, please consult the code 'read_plot.m' that I have updated online.

2. Now let's look at the models. Your task is to do a linear regression on this problem, how do we set up the model?

   Features: We have two features, 'intensity' and 'symmetry', and about 7291 training samples. So the data matrix $\tilde{\mathbf{A}}$ is of size 7291×2. Our decision variable will be $\tilde{\mathbf{x}} = [x_1, x_2]$. Is this enough? Not really. Recall that in Lecture 1, we mentioned that the linear prediction model should be something of the following form (the notation here is a bit different from Lecture 1, but you got the idea...)

   $$f(\mathbf{x}) = \mathbf{x}^T \mathbf{a} + x_0$$

   That is, for any given data point $\mathbf{a}$, the model produces the prediction of the above form. So here x_0 is the variable that characterizes the "intercept" of the line. Therefore in our problem, we need the following classifier

   $$x_1 a_1 + x_2 a_2 + x_0 = 0$$

   which precisely defines a line on a 2-dimensional plane. Therefore, we need another decision variable, so the final variable vector should be $\mathbf{x} = [x_1, x_2, x_0]^T \in \mathbb{R}^3$, and the final data matrix should be

   $$\mathbf{A} = [\tilde{\mathbf{A}}, \mathbf{1}] \in \mathbb{R}^{7291 \times 3}, \text{ with } \mathbf{1} \text{ being the all 1 colume vector of size 7291}$$

   Labels: What's the label $\mathbf{b}$? In order to do a linear regression for classification, we need to *setup our own labels*. Suppose I am going to classify digit '2' and the rest. Then for all instances i correspond to the letter 2, I will label $b_i = +1$, and for the rest of the instances, I will label their b_j's as -1. In the training stage, the problem will be

   $$\min_{\mathbf{x}} \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$$

   where $\mathbf{A}$ and $\mathbf{b}$ are defined above. The question is how to optimize the above problem to find the global optimal solution? Can you run and compare different types of gradient based algorithm and compare how fast these algorithms do in terms of reducing the objective? You will see that for gradient descent with *constant stepsize*, if you don't pick the stepsize carefully, the algorithm will diverge.