The University of Texas at Austin
Department of Electrical and Computer Engineering

EE379K: Data Science Lab — Spring 2018

Lab Seven

Caramanis/Dimakis Due: Friday, April 6, 3:00pm 2018.

Problem 1. Consider the dataset `https://www.kaggle.com/c/GiveMeSomeCredit/data`. This data set asks you to predict who will have a serious delinquency on their loan. If someone is a high-risk individual (i.e., the model predicts $y = 1$, they would be denied a loan).

0. Using some your new-found Kaggle competition skills, fit a good model to these data.

1. Model interpretability: What is the effect of `MonthlyIncome` to the prediction? Quantify as much as you can how 1000, 2000 or 3000 extra per month affect the probability of delinquency. Do this by fitting a simple model on the dataset and using your best model.

2. What is the most important variable in predicting delinquency? What is the most important pair of variables? Make a data science argument supported by data.

3. The Age Discrimination in Employment Act (ADEA) forbids age discrimination against people who are age 40 or older. Does your good model (from part 0 above) discriminate against older people? Make the best argument you can.

4. Your manager asks if the number of dependents in the family (spouse, no of children) has an effect on loan delinquency. What do the data say? Calculate a p-value to express how confident you are.

Problem 2. a) Create two random variables that are uncorrelated but dependent.
b) Create two continuous random variables X, Y so that X and Y are strongly dependent but the best linear regression fit $y = \beta_1 x + \beta_0$ has the optimal $\beta_1 = 0$. Show a scatter plot of x, y pairs.

Problem 3. (Starting with MNIST) Install Tensorflow and Keras 2.0. Use the amazon instances and complete this tutorial: `https://www.tensorflow.org/get_started/mnist/pros`