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

EE379K: Data Science Lab — Fall 2017

Lab Seven

Caramanis/Dimakis Due: Monday Nov 6th, 3:00pm 2017.

Problem 1. The dataset you played on the Kaggle competition was derived from this dataset `https://www.kaggle.com/c/GiveMeSomeCredit/data`. As you can see, you were predicting who will have a serious Deliquency on their loan. If someone is a high-risk individual (i.e. the model predicts $y = 1$ they would be denied a loan).

1. The Kaggle data file `Data Dictionary.xls` explains the real features. The other features were artifically generated. Identify and name the real features in the in-class Kaggle training dataset.

2. 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 probablity of deliquency. Do this by fitting a simple model on the dataset and using your best model.

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

4. The Age Discrimination in Employment Act (ADEA) forbids age discrimination against people who are age 40 or older. Look at the best models you used in your Kaggle competion. Were they discriminating against older people ? Make the best argument you can.

5. Your manager asks if the number of dependents in the family (spouse, no of children) has an effect on loan deliquency. What does 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`