**Machine Learning 2 Finals**

**Vivek D. Kulkarni 16th December 2017**

**1.** Supposedly, if you understand your discipline, you should be able to explain anything you do to your grandmother and she would understand. Explain what Machine Learning is, and how it differs from traditional statistics, in less than 100 words, not for your grandmother, but for a reasonably educated business professional who may be thinking of hiring you.

**Answer 🡪**

ML is backed by artificial intelligence and CS, on the other hand statistics has mathematical functions that formulizes relationship between two things. In ML we learn from data and try to predict / classify as per business problem. These algorithms with the help of programming and traditional statistics provides better solution without any human bias in decision making. ML banks on computing power and statistics on assumptions related to data and how distributed the data is. ML has open approach to learn and solve, however statistics has old-fashioned approach.

**2. Work with the diabetes data from women, PD.csv.**

npreg: the number of pregnancies

glucose: the plasma glucose concentration in an oral glucose tolerance test

bp: the diastolic blood pressure (mm Hg)

skinth: triceps skin-fold thickness measured in mm

bmi: the body mass index (weight in kg/(height in m)^2)

pedprob: the diabetes pedigree function, based on family history. Higher values indicate more relatives have diabetes.

age: age in years

Diabetes: Person has diabetes 5 years later, Yes or No

We want to come up with the best model to predict diabetes in a new sample.

a. Run all applicable ML methods using all variables. Be sure to run at a minimum Linear Regression, Linear Discriminant Analysis, Logistic regression, kNN, SVM, Naïve Bayes, Random forest, CART (Boosted/Bagged Decision Trees), Neural Network, Xgboost. Compile your results in a table. Display confusion matrix for each algorithm

b. Which ML method, of every applicable method, predicts the occurrence of diabetes best?

c. Which predictors perform best?

d. Based on your results, what recommendations would you make to the women to help them avoid diabetes?

**3. Work with the NIJ\_BMOP.csv data.**

We want to come up with the best model to predict "out" (the outcome) in a new sample.

a. Describe the predictors (p1 through p91). Screen them for too many missing values, which will limit the sample sizes. Ideally, you want as many predictors without sacrificing sample size or distribution of your outcome.

b. Run all applicable ML methods. Be sure to run at a minimum Linear Regression, Linear Discriminant Analysis, kNN, SVM, Random Forest, CART (Boosted/Bagged Decision Trees), Neural Network, Xgboost, Partial Least Square Regression, Principal Component Regression, and Generalized Additive Model. Remove redundant or noisy predictors as necessary. Compile your results in a table. Display confusion matrix for each algorithm.

c. Which ML method, of every applicable method, predicts the outcome best?

d. Which predictors help predict the best?

e. Which ML method was *easiest* to use on these data?

P.S. Refer lecture notes for theory behind algorithms Due 14th midnight 11:59pm.