Final Exam
Statistical Machine Learning
Spring 2018

Welcome to the Machine Learning final exam. Your deliverable should include ***a writeup*** of your work that explains any relevant decisions, and discusses the output of you work. **No analysis is sufficient without some discussion of the results.** You need not include all the relevant code, but feel free to do so if you prefer. You should provide your writeup and code in a **single PDF**, with the write-up of the entire exam first, followed by any relevant code you are including (labeled with the relevant question number for reference). ***Failure to follow this format will result in the automatic subtraction of 5 points from your final score.***

Question 1: Simulation study

We talked in class about methods for dimension reduction, including clustering, PCA, etc. One method that I mentioned works surprisingly well is a method known as random projections. In this question you will be running a simulation study to assess, in this very particular case, the effectiveness of random projections. I have provided code "generate_data.R" that generates a high dimensional dataset (inspired by one of the data generating models in "Random-projection ensemble classification" by Timothy I. Cannings, Richard J. Samworth). Our data generating model is:

$$X \sim \left(\frac{1}{2}N(\mu,\Sigma)+\frac{1}{2}N(-\mu,\Sigma)\right)I(Y=0)+F*I(Y=1)$$

where $N(\mu,\Sigma)$ is a multivariate normal, μ is a d-dimensional vector with the first $floor(d*.10)$ dimensions are equal to 1, and the remaining are 0, and F is a distribution such that the first $floor(d*.10)$ are standard Cauchy and the rest are standard Normals. For now, let's think of d between 50 and 250.

The goal of random projections is to preserve the distance relationship between observations. Formally, a random projection is calculated by multiplying each dimension of each observation by a random normal, and then adding up across dimensions. So if $X_i=[x_{i1},x_{i2},\dots,x_{id}]$ is observation i, then we can represent it in terms of p random projection $Y_i=[y_{i1},y_{i2},\dots,y_{ip}]$ where $Y_{ik}=a_{k1}x_{i1}+a_{k2}x_{i2}+\cdots a_{kd}x_{kd}$ where $a_{kj}\sim\frac{1}{\sqrt{p}}N(0,1)$. You can use different distributions for the a_{kj} and different choices have different properties. In matrix notation, we have $Y=WX$ where again $Y\in R^p, X\in R^d, W\in R^{p\times d}$

Letting $||\cdot||$ represent distance, then we want $\frac{||Wx_1-Wx_2||}{||x_1-x_2||}\approx 1$. In our case $x's$ are elements in a vector space, so this reduces to $\frac{||Wx||}{||x||}\approx 1$.

The proof that this works is actually quite simple, relying on concentration inequalities and the chi square distribution.

But I digress. To create this matrix W, use the R code *W <- 1/sqrt(p) * matrix(rnorm(p*d), nrow=p)* where p is the number of random projections you want to take (the dimensionality of the reduced observation) and d is the dimensionality of the original observation. Matrix multiplication in R is accomplished with %*%, and if you have your observations in a matrix of size n x d, where n is the

number of observations, you can compute a reduced dimensional observation matrix as follows: *Y=X%*%W*. Convince yourself that this is correct.

a. As you can imagine, the number of projections used has an effect on the accuracy of the distance calculated in the projected space. Design and run a simulation that studies this effect. Set *d=100*, say. Pick a reasonable number of observations *n*. In this simulation study, you want to run enough trials to characterize and compare the effect (so, lots… number in the 10s or 100s are not appropriate. If you are not in the 1000s you should have a really good reason why)! Also, recall that *W* is random, so it needs to be resampled many times per dimension p.
b. Design and run a simulation to understand the effect of the dimensionality *d* of the original observations on the accuracy.
c. When might you want to use random projections?

Question 2: Building unsupervised profiles

Using the 'acting.Rdata' dataset and the 'genre.Rdata' dataset, use unsupervised learning to build a profile for actors and actresses. While this is similar to the user profiles you did for your homework, there are a number of differences that must be *carefully thought through* and *justified* in your writeup. The deliverable here is a description of the method with analysis of the results (including examples where appropriate).

Question 3: Your final prediction task of ML, Spring 2018

Using the 'acting.Rdata' dataset, the 'genre.Rdata' dataset, and the 'directors.Rdata' dataset, design an algorithm to predict the genre of a movie. You may use the two other datasets as well (I have provided the script for how I extracted acting, genre, and directors for your reference), but it is not required, though they may be quite helpful. Please remember to work in the machine learning paradigm. Your deliverable should include, as always, a robust writeup of your methodology and choices.