Advanced Predictive Modelling
Quiz 2, Fall 2017 Total Marks = 15; *Time: 15 min.*

Declaration: By submitting this quiz for grading, I affirm that I have neither given nor received help from another examinee and acknowledge that **this is a closed-book, closed-notes test.**

Name: ________________________________

Signature: ________________________________

Q1. (2+2 pts) What is a scree plot? For a given dataset, how can a scree plot be used to determine if PCA could potentially be a useful dimensionality reduction method or to indicate what the dimension of the reduced space should be?

A scree plot shows the cumulative eigenvalues against an accumulation of the principal components. ie (sum of top k e.v.s)/(sum of all e.v.s), which reflects the percentage of the total variance retained by the top k eigenvalues. If a few factors can provide large gains in variance, then PCA would be good approach to dimensionality reduction. The plot can be used to identify a cut off for the number of factors to include.

Q2. (3 pts). PCA is a linear and unsupervised method, so it does not use class labels. Sketch a simple example of a classification problem where using PCA will hurt rather than help a classifier learnt on the reduced (i.e. projected) dataset.

see page 2

Q3 (2+2 pts). What do you understand by an MLP with a single hidden layer being a Universal Approximator? Why may it still be beneficial to use several hidden layers (as in deep learning) even though a single hidden layer has the property stated above?

MLPs are universal approximators because through their hidden units and activation functions they are able to model any nonlinear relationship between inputs and responses. Adding more layers could be beneficial for number of epochs needed to converge and the ability to handle additional complexity, preventing overfitting, etc.

Q4 (2+2 pts). When SGD is applied to a regression problem, suppose your model at some time t is realizing a certain function $h_t(x)$

(i) Of what function is the gradient taken, and what variable(s) is the gradient "with respect to"? Function: Loss function of h_t(x) versus the true value at time t.
The gradient will be taken with respect to the weights.

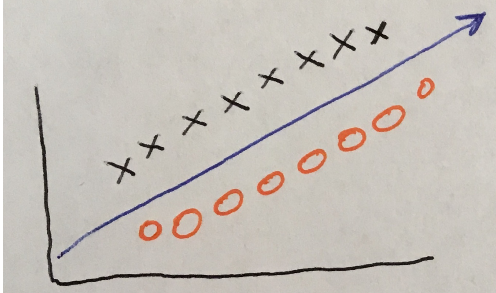
(ii) What is one advantage and one disadvantage of using "mini-batches" rather than pure SGD?

Advantage:

SGD w/MB is more stable (ie, less likely to converge to a suboptimal answer than SGD)

Disadvantage:

but at the cost of taking longer since you are evaluating however many points in your batch size at each step, though this can be made more efficient with vector operations.



In this case both the X's and O's will be projected onto the 1st component vector in a way that does **not** allow us to distinguish between the classes