FINAL EXAM

IS 605 FUNDAMENTALS OF COMPUTATIONAL MATHEMATICS - SPRING 2016

The final exam will consist of 3 types of questions. The first set of 10 questions are designed to evaluate your understanding of essential concepts that you learnt this semester. For each of these questions, a simple 1 line answer will suffice. There is no need to submit any code. These will be followed by 3 questions that might require you to write small programs or functions that use one or more topics. Finally, there will be a mini-project question that will require you to solve a machine learning task. You'll be required to analyze the data, apply some transformations and write code to construct and evaluate the model. The first 10 questions are worth 1.5 point each. The next 3 questions are worth 5 points each. The mini-project will have a maximum of 20 points for a total of 50 points for the final exam.

Please use R and submit your final exam response as an R markdown document. Good luck!

1. Review of Essential Concepts - 15 points

(1) What is the rank of the following matrix?

$$\begin{bmatrix} 1 & -1 & 3 & -5 \\ 2 & 1 & 5 & -9 \\ 6 & -1 & -2 & 4 \end{bmatrix} \tag{1}$$

(2) What is the reduced row-echelon form of the above matrix?

(3) Define orthonormal basis vectors. Please write down at least one orthonormal basis for the 4-dimensional vector space R^4.

(4) Given the following matrix, what is its characteristic polynomial?

$$A = \begin{bmatrix} 6 & 1 & 1 \\ 0 & 7 & -1 \\ -1 & 3 & 0 \end{bmatrix} \tag{2}$$

(5) What are its eigenvectors and eigenvalues? Please note that it is possible to get complex vectors as eigenvectors.

(6) Given a column stochastic matrix of links between URLs, what can you say about the PageRank of this set of URLs? How is it related to its eigendecomposition?

(7) Assuming that we are repeatedly sampling sets of numbers (each set is of size n) from an unknown probability density function. What can we say about the mean value of each set?

(8) What is the derivative of $e^x \sin^2(x)$?

(9) What is the derivative of $\log(x^3 + \sin(x))$?
(10) What is $\int e^x \cos(x) + \sin(x)dx$? Don't forget the constant of integration.

2. Mini-coding assignments - 15 points

2.1. Sampling from function. Assume that you have a function that generates integers between 0 and 50 with the following probability distribution: $P(x == k) = \binom{50}{k} p^k q^{50-k}$ where $p = 0.2$ and $q = 1 - p = 0.8$ and $x \in [0, 50]$. This is also known as a Binomial Distribution. Write a function to sample from this distribution. After that, generate 1000 samples from this distribution and plot the histogram of the sample. Please note that the Binomial distribution is a discrete distribution and takes values only at integer values of x between $x \in [0, 50]$. Sampling from a discrete distribution with finite values is very simple but it is not the same as sampling from a continuous distribution.

2.2. Principal Components Analysis. For the auto data set attached with the final exam, please perform a Principal Components Analysis by performing an SVD on the 4 independent variables (with mpg as the dependent variable) and select the top 2 directions. Please scatter plot the data set after it has been projected to these two dimensions. Your code should print out the two orthogonal vectors and also perform the scatter plot of the data after it has been projected to these two dimensions.

2.3. Sampling in Bootstrapping. As we discussed in class, in bootstrapping we start with n data points and repeatedly sample many times with replacement. Each time, we generate a candidate data set of size n from the original data set. All parameter estimations are performed on these candidate data sets. It can be easily shown that any particular data set generated by sampling n points from an original set of size n covers roughly 63.2% of the original data set. Using probability theory and limits, please prove that this is true. After that, write a program to perform this sampling and show that the empirical observation also agrees this.

3. Mini-project - 20 points

In this mini project, you'll perform a Multivariate Linear Regression analysis using Stochastic Gradient Descent. The data set consists of two predictor variables and one response variable. The predictor variables are living area in square feet and number of bedrooms. The response variable is the price of the house. You have 47 data points in total.

Since both the number of rooms and the living area are in different units, it makes it hard to compare them in relative terms. One way to compensate for this is to *standardize* the variables. In order to standardize, you estimate the mean and standard deviation of each variable and then compute new versions of these variables. For instance, if you have a variable x, then then standardized version of x is $x_{std} = (x - \mu)/\sigma$ where μ and σ are the mean and standard deviation of x, respectively.

As we saw in the gradient descent equations, we introduce a dummy variable $x_0 = 1$ in order to calculate the intercept term of the linear regression. Please standardize the

2 variables, introduce the dummy variable and then write a function to perform gradient descent on this data set. You'll repeat gradient descent for a range of α values. Please use $\alpha = (0.001, 0.01, 0.1, 1.0)$ as your choices. For each value of α perform about 500 SGD iterations with 5 randomly picked samples in each iteration and compute $J(\theta)$ at the end of each iteration. When you perform SGD, you randomly pick a mini-batch (in this case 5 samples), use that mini-batch to compute the gradient, and then take a step to improve the objective function. You repeat this process in each iteration. It is very important to randomly pick samples in each iteration – otherwise SGD will not work. Please plot $J(\theta)$ versus number of iterations for each of the 4 α choices.

Once you have your final gradient descent solution, compare this with regular linear regression (using the built-in function in R). Please document both solutions in your submission. How does the SGD solution differ from the Linear Regression solution? Are they different? If so, why? If not, why not?