# FINAL EXAM

IS 605 FUNDAMENTALS OF COMPUTATIONAL MATHEMATICS - FALL 2014

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. These will be followed by 5 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 point each. The next 5 questions are worth 5 points each. The mini-project will have a maximum of 15 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 - 10 points

(1) What is the rank of the following matrix? You'll need to show your calculations and not simply return the rank information provided by R.

$$\begin{bmatrix} -1 & 1 & 3 & 5 \\ 2 & -1 & 5 & 7 \\ 6 & -10 & -1 & 3 \end{bmatrix} \tag{1}$$

(2) What is the determinant of the above matrix?

(3) Define orthonormal basis vectors. Please write down at least one orthonormal basis for the 5-dimensional vector space $R^5$.

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

$$A = \begin{bmatrix} 2 & -1 & 4 \\ -1 & -2 & 6 \\ 1 & 0 & -3 \end{bmatrix} \tag{2}$$

(5) What are its eigenvectors and eigenvalues? Please show your calculations. Do not simply use the eigenvalue function available in R.

(6) When would a model be said to have a *high bias* and when would it be said to have a *high variance*?

(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 average value of each set?

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

(9) What is the derivative of $\log(\sin(2x))$?

(10) What is $\int x^2 \sin(x)dx$? Compute this value for the interval $x \in [0, \frac{\pi}{2}]$. Assume x is in radians.

## 2. Mini-coding assignments - 25 points

2.1. **Bayes Rule.**

- You are working for a credit card company and you know that about 70% of your customers have a *good credit score* . People with good credit have a loan default rate of 0.5%. Whereas people with *bad credit* default on their loans at the rate of 10%. Given that you are looking at a defaulted loan, what is the probability that it belongs to someone with a *bad credit* score?
- Suppose there are two full bowls of cookies. Bowl 1 has 5 Chocolate Chip and 35 Oatmeal Raisin cookies, while bowl 2 has 20 of each. Our friend Fred picks a bowl at random, and then picks a cookie at random. We may assume there is no reason to believe Fred treats one bowl differently from another, likewise for the cookies. The cookie turns out to be a Chocolate Chip one. How probable is it that Fred picked it out of Bowl 2?

2.2. **Central Limit Theorem.**

- Assume a sample of size 100 is selected from a log-normally distribution population with mean 50 and standard deviation 10. What will be the mean of this sample? What will be its standard error?
- A random sample of 100 observations is to be drawn from a population with a mean of 40 and a standard deviation of 25. The standard deviation of the mean's distribution will be?
- A random sample of 100 observations is to be drawn from a population with a mean of 40 and a standard deviation of 25. The probability that the mean of the sample will exceed 45 is approximately? (please be precise to within 2 decimal places).

l

2.3. **Sampling from function.** Assume that you have a function that generates integers between 0 and 10 with the following probability distribution: $P(x = k) = \binom{10}{k} p^k q^{1-k}$ where $p = 0.1$ and $q = 1 - p = 0.9$ and $x \in [0, 10]$. 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 $x \in [0, 10]$. Sampling from a discrete distribution with finite values is very simple but it is not the same as sampling from a continuous distribution.

2.4. **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.

2.5. **Sampling in Bootstrapping.** As we discussed in one of the lectures, 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.

## 3. Mini-project - 15 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 $\mu$ and $\sigma$ 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 stochastic gradient descent on this data set. You'll repeat stochastic gradient descent for a range of $\alpha$ values. Please use $\alpha = (0.001, 0.01, 0.1, 1.0)$ as your choices. For each value of $\alpha$ perform about 5000 SGD iterations and compute $J(\theta)$ at the end of each 100 iterations. Please plot $J(\theta)$ versus number of iterations for each of the 4 $\alpha$ choices.

Once you have your final gradient descent solution, compare this with regular linear regression (using the built-in function in R). Also solve using Ordinary Least Squares approach. Please document all 3 solutions in your submission.