ECE 592 – Topics in Data Science

Final Exam – Fall 2016

December 9, 2016

Please remember to justify your answers carefully, and to staple your test sheet and answers together before submitting.

Name: _____ Student ID: _____

Question 1

You are interviewing at a large bank that makes car loans to customers. In the past, decisions about car loans were made by individual bankers who met customers in person, but the bank is concerned that bankers may be prone to biases. The bank wants you to develop guidelines for making loans.

(a) Suppose that the bank only *wants to know whether to make a loan or not* (they want a binary output). You have access to X, a matrix whose rows correspond to customers and whose columns correspond to factors about each customer that may impact the loan making decision. You also have access to a vector Y that indicates whether the customer paid back the loan. How will you use X and Y to make loan decisions?

(b) The bank can profit from risky customers by requiring that they make larger payments on their loans. To compute the payments, the bank *wants to estimate the percentage of the loan that will be paid* (a continuous valued output). How will you estimate this percentage?

(c) The bank knows that if its algorithm is complicated, its bankers will be uncomfortable using it. To reduce this problem, the bank wants to perform some form of feature selection or subset selection on its data. How would you approach this problem algorithmically?

Question 2

Many of us are familiar with the formula for the sum of the first n natural numbers, $f(n) = \sum_{i=1}^{n} i = \frac{n(n+1)}{2}$. In this question, you will prove that the sum of the squares of the first n natural numbers obeys

$$g(n) = \sum_{i=1}^{n} i^2 = \frac{1}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n.$$

You will prove this result using induction.
(a) Basis case – provide some small value of n for which you can show that the formula holds.
(b) Inductive step – prove that if the formula holds for $1, 2, \ldots, n$, then it also holds for $n+1$. (Recall that $(n+1)^2 = n^2 + 2n + 1$ and $(n+1)^3 = n^3 + 3n^2 + 3n + 1$.)
(c) Prove that $g(n) = \Theta(n^3)$.

Question 3

Consider a communication system that modulates a binary input vector $x \in \{-1, +1\}^N$, which has values -1 or $+1$, each appearing with probability 50%. These values are multiplied by columns of a matrix $A \in \mathbb{R}^{M \times N}$ whose individual entries are zero mean Gaussian with variance $1/M$. The sum of the modulated columns, Ax, is transmitted to a Gaussian channel, which adds noise $z/\sqrt{\gamma}$, where $z \in \mathbb{R}^M$ is zero mean unit norm Gaussian, and $\gamma > 0$ is the signal to noise ratio (SNR). The resulting noisy measurements are $y = Ax + z/\sqrt{\gamma}$.
Our goal is to decode or reconstruct x given y and A. To do so, we will use the approximate message passing (AMP) framework. Similar to Homework 4, you will develop the denoising function, which will be applied to the output of the scalar channel.
(a) Develop a denoising function, $E[X|V]$. You can assume that V, the output of the scalar channel, obeys $V = X + W$, where W is a Gaussian random variable with zero mean and variance σ^2. (Recall that the probability density function of W satisfies $f_W(w) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{w^2}{2\pi\sigma^2}}$.)
(b) Write the denoising function in Matlab. You can assume that v and $varw$ are existing variables that correspond to the output of the scalar channel and the variance of the Gaussian noise. Your output should be assigned to the variable $xhat$.

Question 4

Recall the rod cutting problem from the slides on optimization, where we cut a rod of length 4 in a way that maximizes its value. The prices charged for rods of lengths 1 to 4 are $p_1 = 1$, $p_2 = 5$, $p_3 = 8$, and $p_4 = 9$.
(a) Develop a dynamic programming formulation for this problem. In particular, show how $\Psi(n)$, the optimal cost for cutting a rod of length n, depends on $\Psi(1), \Psi(2), \ldots, \Psi(n-1)$.
(b) How should you cut a rod of length $n = 5$ in an optimal way?