Stats 598z: Midterm exam 2

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 7 questions, each for 5 points (but not all equally easy)
Attempt all questions, and when appropriate include a brief justification of your answer
Don't spend time polishing your answers as the main idea is more important

---

1. (a) `my_vec2` is a vector of integers. Write a few lines of `R` to return the number of times two consecutive numbers occur in successive positions. E.g. $(5, 6, 7, 2, 4, 5)$ returns `3`. Use vectorization for full points.

   (b) `my_vec1` is a vector consisting of sequence of numbers that are interpreted as $(\texttt{len1}, \texttt{obs}_1, \texttt{obs}_2, \ldots, \texttt{obs}_{\texttt{len1}}, \texttt{len2}, \texttt{obs}_1, \texttt{obs}_2, \ldots, \texttt{obs}_{\texttt{len2}}, \ldots)$. Thus, the first number gives the length of a set of observations that follows it, these in turn are followed by the length of the next set, and so on. Write a few lines of `R` code to return the mean of each set, so that $(3, 1, 2, 3, 2, 5, 3)$ returns $(2, 4)$. You can use loops.

2. You have a vector `my_words`. Write down the `R` regular expression command to find components

   (a) containing `m` followed by one or more `c`'s.

   (b) starting with a vowel.

   (c) starting and ending with a vowel.

   (d) containing ".txt".

   (e) Write `R` code to replace all ".txt" with ".doc".

3. (a) What is coordinate descent? What is gradient descent? Briefly explain possible advantages and disadvantages.

   (b) What is the role of λ in LASSO. Briefly explain how you will use cross-validation to choose λ.

   (c) For a fixed dataset, you want to solve a series of LASSO problems for $\lambda = 0, .1, .2, .5, 1, 2, 5$ and 10. Explain briefly how you can order these problems for efficiency.

4. (a) Briefly describe some of the advantages of object-oriented programming.

   (b) You have `my_list`, a list of vectors of numbers. What does `my_list[1]` return (and what type of data object is this)? How would you use object-oriented programming so that `my_list[i]` returns the mean of the `i`th vector of `my_list`? For full points, provide necessary `R` code.

5. (a) What is tidy data? What are tall and wide formats of dataframes? Give a simple example.

   (b) A data frame `population` has columns `name, city, age, occupation, gender` and `income`, each row giving these quantities for a different individual. Give `R` code to convert it into a new dataframe giving the average income for each occupation and for each gender (averaged across all people, cities, and ages). What if you wanted to do this for each occupation, gender and age decade as well (i.e. 0-9, 10-19 etc.)? To get full points, use `melt()`, `dcast()`, `*pply` or vectorization, but if not sure, use loops.

6. (a) Describe importance sampling at a high-level. Explain (e.g. with an example) how it is useful for estimating quantities involving rare events.

   (b) Briefly describe any Monte Carlo method to estimate $\int_a^b \log(x)\mathrm{d}x$. Provide a few lines of `R` code.

7. (a) Explain why Metropolis-Hastings might be better than importance/rejection sampling.

   (b) Given a 100 coins, you want to calculate the average number of heads given that the number of heads is greater than 70: $\mathbb{E}[\#\texttt{H}|\#\texttt{H} > 70]$. Briefly describe a Metropolis-Hastings algorithm to do this, giving the initialization, the proposal distribution and the acceptance probability.