Stats 598z: Midterm exam 2

Important:
Write you name and PUID on all sheets, and include the number of sheets
There are 5 questions, each for 10 points.
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) A dataframe `menu` has columns `dish`, `cuisine`, `meat`, `price` and `calories`. In the following, use commands from `tidyverse` for full points, but you will get most points if you just use base R.
   - `cuisine` has values `American, Chinese, Thai` etc. Write code to get the number of dishes of each cuisine.
   - `meat` has values `veg, beef, fish` etc. Get the number of dishes of each `(cuisine, meat)` pair, as well as the minimum and maximum price for each pair, sorting by the minimum price.
   - Write code to plot `calories` (y-axis) vs `price` (x-axis), for each `(cuisine,meat)` pair.
   - Convert the dataframe to a new one with columns `cuisine` followed by all unique values of `meat`, i.e. `cuisine, veggie, beef, fish, ....` Each row corresponds to one cuisine, and gives the average price of `veggie`, `beef`, `fish` etc dishes in that cuisine.

2. `countries` is a vector of names of countries, all in lower case. Write down an `R` regular expression commands to find:
   (a) countries whose names contain an `i`, and end with '`land`'
   (b) countries whose names have an '`i`' as the second letter, and '`a`' as second last
   (c) countries whose names contain an '`a`' AND an '`i`' in them
   (d) countries whose names contain an '`a`' OR an '`i`' in them
   (e) countries whose name contains a non-alphabetical character (e.g. `'south korea'` contains a space).
   (f) countries whose name starts and ends with the same letter

3. (a) Briefly describe some of the advantages of object-oriented programming.
   (b) `my_scores` is a vectors of numbers. Use object-oriented programming so that `print(my_scores)` prints the minimum and maximum of this vector.
   (c) What is a generic function? What is an infix function?
   (d) For the dataframe `menu`, use `ggvis` to plot `calories` vs `price` as points, but printing the dish name and cuisine whenever you hover the mouse over a point.

4. (a) Write down the LASSO loss-function, briefly explaining why it returns sparse solutions.
   (b) What is a subgradient? What is the soft-threshold function?
   (c) In the context of optimization, give the update rule for Newton's method. Explain its intuition.
   (d) Briefly describe a Monte Carlo method to estimate $\int_0^5 |\sin(x)| \mathrm{d}x$. Provide a few lines of `R` code.

5. (a) Let `x` and `y` be Gaussians with mean 0 and variance 1. Provide `R` code for a Monte Carlo estimate of the probability that $|\mathtt{x}| > \mathtt{k}|\mathtt{y}|$ or $|\mathtt{y}| > \mathtt{k}|\mathtt{x}|$ for a given `k`.
   (b) When might the above sampler be inefficient? At a high level, describe how to fix it using importance sampling.
   (c) Explain the Metropolis-Hastings algorithm at a high level. What does 'burn-in' refer to?