Statistics 700 Homework 1

Bayes Models and R Coding

Due date: 6:00 pm (EST) Sept. 19, 2017

Let's practice coding! Find a data set, preferably the same as the one you use for your final project, and perform the following.

1. Read the data into R/Python. Examine the following:
   (a) dimension of the data, number of entries;
   (b) summary of each variable/dimension;
   (c) visualize each variable using boxplot/histogram/table;
   (d) calculate summary statistics of the data as a whole or for each variable.
2. If simple linear regression model is applicable, test it out.
3. If k-means clustering is applicable, test it out.
4. Calculate pairwise correlations across all variables. What do you find?
5. Perform any other exploratory data analysis you have on mind.

Finally, summarize the features of the data set. What is interesting or unique about the data? What might be a potential difficult feature of the data? Are you going to use this data set for your final project? Why or why not?

Guideline for Submission: submit R markdown (or jupyter notebook) with annotated code followed by results. Discussions about the results should follow the results.

Try out FLUX.

1. Upload your data to your home directory.
2. Create a folder called "Hw2_YourUniqueName" under your home directory.
3. Write an R/Python script conducting any appropriate exploratory data analysis, with outputs in the format of '.txt' (for numbers) and '.pdf' (for images), which are saved to the folder you just created. Submit your job to Flux.
4. Cut your R script into at least 3 independent parts, e.g. one part for summary statistics, one part for histogram visualization. Submit a job array on Flux to run these independent parts on parallel cores.

Guideline for Submission: submit R markdown (or jupyter notebook) with annotated code followed by results. The document should at least include: R/Python script for submission, commands for Flux job submission, and results of runtime.

Optional Reading. Read one of the following papers and post your summary and thoughts on Canvas. Bonus points up to 5 will be rewarded.

1. Biostatistics and Bayes, Norman Breslow, Statist. Sci., Volume 5, Number 3 (1990), 269-284.
2. Bayesian Methods in Practice: Experiences in the Pharmaceutical Industry, A. Racine, A. P. Grieve, H. Fluhler and A. F. M. Smith, Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 35, No. 2 (1986), pp. 93-150.
3. Induction and deduction in Bayesian data analysis. Andrew Gelman, 2011.
4. Philosophy and the practice of Bayesian statistics (with discussion), A. Gelman and C. Shalizi, British Journal of Mathematical and Statistical Psychology, 66, 8-80.