Homework 1: Probability Review and Priors

Instructions: You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below. Include an honor pledge with your submission. Submit on-line and in pdf. This homework is worth 100 points and the point totals for each question are shown in parentheses.

1. (15) You are a data scientist and are choosing between three approaches, A, B, and C, to a problem. With approach A you will spend a total of four days coding and running an algorithm and it will not produce useful results. With approach B you will spend a total of three days coding and running an algorithm and it will not produce useful results. With approach C you will spend one day coding and running an algorithm and it will give the results you are looking for. You are starting your project and do not know which approach will work, and so are equally likely to choose among unselected options but if your selected approach does not work you will select a new one and continue until you get useful results. What is the expected time in days for you to obtain the results you are looking for? What is the variance on this time?

2. (15) Suppose if it is sunny or not in Charlottesville depends on the weather of the last three days. Show how this can be modeled as a Markov chain by displaying a diagram and transition matrix.

3. (15) Assume a Gaussian distribution for observations, $X_i, i = 1, \ldots, N$ with unknown mean, M, and known variance 5. Suppose the prior for M is Gaussian with variance 10. How large a random sample must be taken (i.e., what is the minimum value for N) to specify an interval having unit length of 1 such that the probability that M lies in this interval is 0.95? [1]

4. (20) You have started an online business selling books that are of interest to your customers. A publisher has just given you a large book with photos from famous 20^{th} century photographers. You think this book will appeal to people who have bought art books, history books and coffee table books. In an initial offering of the new book you collect data on purchases of the new book and combine these data with data from the past purchases (see ArtHistBooks.csv).

   Use Bayesian analysis to give the posterior probabilities for purchases of art books, history books and coffee table books, as well as, the separate probabilities for purchases of the new book given each possible combination of prior purchases of art books, history books and coffee table books. Do this by first using beta priors with values of the hyperparameters that represent lack of prior information. Then compute these probabilities again with beta priors that show strong weighting for low likelihood of a book purchase. Compare your results.

5. (20) The data set CHDdata.csv contains cases of coronary heart disease (CHD) and variables associated with the patient's condition: systolic blood pressure, yearly tobacco use (in kg), low density lipoprotein (ldl), adiposity, family history (0 or 1), type A personality score (typea), obesity (body mass index), alcohol use, age, and the diagnosis of CHD (0 or 1).

   Perform a Bayesian analysis of these data that finds the posterior marginal probability distributions for the means for the data of patients with and without CHD. You should first standard scale (subtract the mean and divide by the standard deviation) all the numeric variables (remove family history and do not scale CHD). Then separate the data into two sets, one for patients with CHD and one for patients without CHD.

   Your priors for both groups should assume means of 0 for all variables and a correlation of 0 between all pairs of variables. You should assume all variances for the variables are 1. Use a prior alpha equal to one plus the number of predictor variables. Compute and compare the Bayesian estimates for the posterior means for each group.

   For 5 extra credit points, compute the probability of observing a point at least as extreme as the posterior mean of patients without coronary heart disease under the posterior distribution for the patients with coronary heart disease. Then compute the probability of observing a point at least as extreme as the posterior mean of patients with coronary heart disease under the posterior distribution for the patients without coronary heart disease.

6. (15) Using the Python Notebook https://www.kaggle.com/billbasener/pt2-probabilities-likelihoods-and-bayes-theorem, complete the challenge question from Section 6: Modify the code from Section 5 to and add the ability to use the `posterior_from_conjugate_prior` function to output the posterior probability parameters given parameters and for a Gaussian Likelihood with known variance σ^2, and use your modified function to create the Prior, Likelihood, Posterior plots as in Section 5 of the notebook.

References

[1] DeGroot, Morris H, *Optimal Statistical Decision*, New York: McGraw-Hill, 1970.