MSAN 623: Multivariate Statistical Analysis

Practice Final Exam

May 3rd, 2017

This was constructed to represent a more general version of the test you will receive next Friday. So, you should see that as a guided exercise to frame your study habits. If you have any questions on any of this material, please be sure to ask me in class on the review day next Wednesday.

1. (*Big Picture of Bayesian Analysis*): Describe what is meant by "Bayesian analysis." Use any notation that you would like to do this. How does "Bayesian analysis" and "Frequentist analysis" compare, and when do they give you the same conclusions?

2. (*Choice of Prior Distribution*): Suppose that we have a density, $f(y \mid \theta)$ that we use to model observed data y. We would like to make inference on θ from its *posterior* distribution $p(\theta \mid y)$, which we calculate using some *prior* distribution $\pi(\theta)$. Describe two ways that one can go about choosing a prior distribution. Discuss the advantages of each type of choice. What difficulties might we run into when trying to calculate or simulate from the posterior $p(\theta \mid y)$? In terms of the density $f(y \mid \theta)$, what is a conjugate family?

3. (*Regression*) A data set contains n training observations $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)$ where $\mathbf{x}_j$ is a $1 \times p$ vector whose entries represent p predictors, and $y_j \in \mathbb{R}$. Consider fitting the following regression model:

$$f(\mathbf{x}_j) = \mathbf{x}_j \beta + \epsilon_j, \qquad j = 1, \ldots, n \tag{1}$$

where $\{\epsilon_j\}$ are independent and identically distributed random variables with mean $\mathbb{E}[\epsilon_j] = 0$. Upon estimating model (1) we obtain

$$\widehat{f}(\mathbf{x}) = \mathbf{x}\widehat{\beta}$$

so that the "best guess" for observation y_j is given by

$$\widehat{y}_j = \widehat{f}(\mathbf{x}_j), \qquad j = 1, \ldots, n$$

   (a) We talked about three primary methods for fitting $\widehat{f}$: ordinary least squares, Lasso, and Ridge Regression. Describe the similarities and differences between each of these three methods. In particular, describe when you would use one method over another and consider the optimization function that you must solve in order to calculate parameter estimates. Furthermore, discuss the Bayesian way of interpreting Lasso and Ridge estimates.

   (b) If the observations $y_1, \ldots, y_n$ were binary observations instead of continuous observations, we may instead use *logistic regression* for classification. Using the data $(\mathbf{x}_1, y_1) \ldots, (\mathbf{x}_n, y_n)$, write out the logistic regression model. If you estimated the best coefficients $\widehat{\beta}$ for your logistic model, how would you go about calculating the probabilities for each observation have a response of 1? (Be explicit about any distributional choices). How can we add a "Bayesian spin" to this model?

4. (*Laws of Total Expectation*) Suppose that Y and θ are jointly distributed random variables, where

   - $Y \mid \theta \sim N(\theta, \theta^2)$
   - $\theta \sim N(0, \lambda)$

   (a) What is the $\mathbb{E}[Y]$?

   (b) What is the $\mathrm{Var}(Y)$?

5. (*Classification*) Many classification methods are Bayesian in nature. In particular, (X, Y) is viewed as a jointly distributed random variable, where X is p-dimensional and, in the simplest case, $Y \in \{-1, 1\}$. We then build a classifier based on the conditional probability derived from

Bayes rule. In particular, we talked about four main types of classification that were directly related to Bayesian analysis. Describe how, in terms of (X, Y), to derive the Bayes classifier. Why do we sometimes refer to the Bayes classifier as the "gold standard" of classification methods? Now using this derivation as a starting point, describe what assumptions must be made in order to obtain (a) the Naïve Bayes classifier, (b) linear discriminants, and (c) quadratic discriminants.

6. (*Markov Chains*) Suppose that the chance of rain tomorrow depends on previous weather conditions only through whether or not it is raining today and not on past weather conditions. If it rains today, it will rain tomorrow with probability α. If it does not rain today, it will rain tomorrow with probability β. Let X_t be the binary variable that describes whether or not it rains on day t.

   (a) Write out the state space for the stochastic process $\{X_t : t \geq 1\}$

   (b) Find the transition probability matrix $\mathbb{T}$ describing the process.

   (c) What is the limiting probability of rain? (*Hint*: use the equation $\pi = \pi^T \mathbb{T}$)

   (d) Where do Markov chains come into play with Markov chain Monte Carlo simulations?

7. (*Sampling and Simulations*) Describe when and why you need to use each of the following sampling methods:

   (a) Monte Carlo simulations

   (b) Rejection sampling

   (c) Importance sampling

   (d) Markov chain Monte Carlo (in particular, Gibbs, Metropolis, and Metropolis Hastings)

8. (*Probability laws*) Write out the laws of total expectation and law of total probability. When do you need to use these things? As a followup to that question, use the law of total probability to prove Bayes rule.