NYU Center for Data Science: DS-GA 1003

Machine Learning and Computational Statistics (Spring 2018)

Brett Bernstein

March 27, 2018

Instructions: Following most lab and lecture sections, we will be providing concept checks for review. Each concept check will:

- List the lab/lecture learning objectives. You will be responsible for mastering these objectives, and demonstrating mastery through homework assignments, exams (midterm and final), and on the final course project.
- Include concept check questions. These questions are intended to reinforce the lab/lectures, and help you master the learning objectives.

You are strongly encourage to complete all concept check questions, and to discuss these (and related) problems on Piazza and at office hours. However, problems marked with a ($\star$) are considered optional.

Bayesian Methods and Regression: Concept Check

Bayesian Methods and Regression

Bayesian Methods and Regression Learning Objectives

- (Recap) Recall the basic Bayesian setup (likelihood and prior), and be able to write the posterior distribution using proportionality – (see slide 15 for Gaussian Example).
- Explain the difference between the posterior predictive distribution function and the MAP or posterior mean estimator.
- Be able to show the relationship between Gaussian regression and ridge regression.
- Explain what a predictive distribution is, and how it gives additional information (relative to the prediction functions we've learned in our ridge/lasso homework, for example).

Bayesian Methods and Regression Concept Check Questions

1. (From DeGroot and Schervish) Let θ denote the proportion of registered voters in a large city who are in favor of a certain proposition. Suppose that the value of θ is unknown, and two statisticians A and B assign to θ the following different prior PDFs $\xi_A(\theta)$ and $\xi_B(\theta)$, respectively:

$$\begin{aligned} \xi_A(\theta) &= 2\theta \quad \text{for } 0 < \theta < 1, \\ \xi_B(\theta) &= 4\theta^3 \quad \text{for } 0 < \theta < 1. \end{aligned}$$

   In a random sample of 1000 registered voters from the city, it is found that 710 are in favor of the proposition.

   (a) Find the posterior distribution that each statistician assigns to θ.

   (b) Find the Bayes estimate of θ (minimizer of posterior expected loss) for each statistician based on the squared error loss function.

   (c) Show that after the opinions of the 1000 registered voters in the random sample had been obtained, the Bayes estimates for the two statisticians could not possibly differ by more than 0.002, regardless of the number in the sample who were in favor of the proposition.

2. Two statistics students decide to compute 95% confidence intervals for the distribution parameter θ using an i.i.d. sample $X_1, \ldots, X_n$. Student B uses Bayesian methods to find a 95% credible set $[L_B, R_B]$ for θ. Student F uses frequentist methods to find a 95% confidence interval $[L_F, R_F]$ for θ. Both conclude that parameter θ is in their respective intervals with probability at least .95. Who is correct? Explain.

3. Suppose θ has prior distribution Beta(a, b) for some $a, b > 0$. Given θ, suppose we make independent coin flips with heads probability θ. Find values of a, b and the coin flips so that the posterior variance is larger than the prior variance. [Hint: Recall that a Beta(a, b) random variable has variance given by

$$\frac{ab}{(a+b)^2(a+b+1)}.$$

   Try $b = 1$.]

4. Fix $\sigma^2 > 0$. Let w, taking values in $\mathbb{R}^d$, have prior distribution $\mathcal{N}(\mu_0, \Sigma_0)$. Conditional on w and $x_1, \ldots, x_n \in \mathbb{R}^2$ suppose that $y_1, \ldots, y_n$ are i.i.d. with $y_i \sim \mathcal{N}(w^T x_i, \sigma^2)$. Let $\mathcal{N}(\mu_1, \Sigma_1)$ denote the posterior distribution of w given the data $\mathcal{D} = \{(x_1, y_1), \ldots, (y_n, y_n)\}$.

   (a) Given a new x-value you want to forecast y to minimize the expected square loss. That is, we want to find

$$\hat{y} = \arg\min_y \mathbb{E}_{y'}(y - y')^2,$$

   where y' has the predictive distribution given x and $\mathcal{D}$. What is $\hat{y}$, and what is the associated expected loss $\mathbb{E}_{y'}(\hat{y} - y')^2$?

(b) What types of values for σ, Σ_0, n will lead to the prior exerting a lot of influence on our prediction?

(c) We saw that the Bayesian approach to Gaussian linear regression corresponds to ridge regression. What values in the Bayesian approach correspond to a large amount of regularization?

5. Suppose you are using Bayesian techniques to fit a Poisson regression model. Conditional on x, w, we have $y \sim \text{Pois}(e^{w^T x})$. A colleague, working with his own data set and prior, has given you a function f that returns i.i.d. samples from his posterior distribution on w. Give pseudocode that, given x, lets you sample from the predictive distribution of y given x.