Overview

Submit your writeup including any code as a PDF via gradescope.[1] We recommend reading through the entire homework beforehand and carefully using functions for testing procedures, plotting, and running experiments. Taking the time to reuse code will help in the long run!

Data science is a collaborative activity. While you may talk with others about the homework, please write up your solutions individually. If you discuss the homework with your peers, include their names on your submission. Please make sure any handwritten answers are legible, as we may deduct points otherwise.

1. Ridge Regression and Bayesian MAP

In this problem, we'll compute the posterior distribution for a Gaussian linear regression model, and explore the relationship between the maximum *a posteriori* (MAP) estimate and the result of Ridge regression. You may find it helpful to review Chapter 18 of the Data 100 textbook, which covers multiple linear regression.

Recall that in linear regression, we're trying to predict a (scalar) number y using a vector of (fixed) features $x \in \mathbb{R}^d$. Suppose we have n data points (pairs of X and y). We can write

$$y_i = \beta^T x_i + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where each $\varepsilon_i \sim N(0, \sigma^2)$ are independent of each other, and $\beta \in \mathbb{R}^d$ and $\sigma^2 > 0$ are unknown. Our goal is to estimate β, which tells us about the relationship between X and y.

Let $y = (y_1, \ldots, y_n)$ be the vector of y-values; $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ be the vector of ε values, and let X denote the $n \times d$ matrix whose i-th row is equal to x_i. Using this notation and linear algebra, we can write the model as

$$y = X\beta + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2 I_n), \tag{2}$$

where I_n denotes the $n \times n$ identity matrix.

(a) (3 points) We'll think about the problem from a Bayesian perspective and assume that β is random. We put a zero-mean normal prior on β, i.e. $\beta \sim N(0, \sigma_0^2 I_d)$. To simplify the algebra, we'll also assume that $\sigma = 1$ so that $\varepsilon \sim N(0, I_n)$.

Show that the posterior distribution for β given the data y satisfies

$$p(\beta|y) \propto \exp\left\{-\frac{1}{2} q(\beta)\right\}. \tag{3}$$

where $q(\beta) = \beta^T A \beta - b^T \beta + c$, where A is a $d \times d$ square matrix, and $b \in \mathbb{R}^d$ is a vector. What are A and b (you may ignore c)? **You must show your work!**

Hint: use Bayes' rule and linear algebra.

[1] In Jupyter, you can download as PDF via LaTeX or print as html and then save as PDF

Taking your result from part (a) and completing the square, you can show that $q(\beta) = (\beta - \beta^*)^T \Sigma^{-1} (\beta - \beta^*) + c$. Because the terms involving β have the form of a normal distribution, we can ignore c, and conclude that the posterior distribution for β is multivariate normal with mean β^* and covariance Σ. The mean β^* and the inverse covariance Σ^{-1} are given by:

$$\Sigma^{-1} = \frac{1}{\sigma^2}\left(\frac{\sigma^2}{\sigma_0^2} I_d + X^T X\right) \tag{4}$$

$$\beta^* = \Sigma \frac{X^T y}{\sigma^2}. \tag{5}$$

(b) Using the file `election_data.csv`, we'll try to predict the outcome of the 2020 election using information from previous elections. Specifically, for each Congressional district, we'll try to predict how much Democrats won by (`HouseDem20 margin`) using the current officeholder's ideology score[2] (`Govtrack ideology`) and the result from the 2018 election (`HouseDem18 Margin`).

Because we're predicting using only two variables, we can easily visualize each value for the 2-dimensional β. When making visualizations below, your x-axis should have values for the coefficient that corresponds to `Govtrack ideology`, and the y-axis should have values for the coefficient that corresponds to `HouseDem18 Margin`.

(i) (3 points) Using the information given in Equations (4) and (5), compute the posterior distribution for β given the data as described above. Make three contour plots showing the posterior distributions for different values of $\sigma_0^2 = \{1, 0.1, 0.01\}$, assuming that $\sigma^2 = 1$. You should use the provided `plot_contour` function.

Comment on any differences you see between the three contour plots.

(ii) (2 points) For each value of σ_0^2 from part (i), find the MAP estimate for β. Plot your three estimates on a scatterplot using crosses, where the x and y axes are the same as above.

Hint: Remember, each estimate should be a single point, so your scatterplot should have exactly three crosses on it.

(iii) (2 points) On the same scatterplot, plot three estimates for β using Ridge regression. Recall that Ridge regression solves the following problem:

$$\arg\min_{\beta} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2$$

Your three estimates should use values $\lambda = 1$, $\lambda = 10$, and $\lambda = 100$, and should be plotted using circles.

Hint: Remember, each estimate should be a single point, so your scatterplot should have exactly three circles on it.

(iv) (1 point) You should find that the two methods produce the same results. Intuitively, why might that be?

[2] Data used is from govtrack. For more information on this dataset, see this notebook.

Hint: This question shouldn't necessarily involve any algebra. If you're seeing different results, check your code for calculating the MAP estimate against Equations (4) and (5), and make sure you check the default parameters being used in Ridge regression!

Hint: You may need to use alpha transparency to properly display all six points (three from MAP estimation and three from Ridge regression): see discussion 8 for a reminder on how to do this.

(c)

(i) (3 points) Take a closer look at the crosses in your plot from 1(b)(ii) for $\sigma_0 = 1$ and $\sigma_0 = 0.1$. How does the choice of σ_0 affect the results of MAP estimation? What are the different assumptions being made about the effect of ideology and 2018 results on predictions for 2020 results?

(ii) (2 points) Your friend claims that based on the results of 1(b)(ii), there's no point in doing the math to find the Bayesian posterior for this model when you can just use the estimate of $\hat{\beta}$ from Ridge regression. Do you agree? What extra benefits might the Bayesian approach provide?

2. Rejection Sampling

Consider the function

$$g(x) = \cos^2(12x) \times |x^3 + 6x - 2| \times \mathbb{1}_{x \in (-1,-.25) \cup (0,1)}.$$

In this problem, we use rejection sampling to generate random variables with pdf $f(x) = cg(x)$.

(a) (2 points) Plot g over its domain. What is a uniform proposal distribution q that covers the support of f? What is a constant M such that the scaled target distribution $p(x) = Mg(x)$ satisfies $p(x) \leq q(x)$ for all x?

(b) (4 points) Suppose you run rejection sampling with target p and proposal q from part (a) until you generate n samples and your sampler runs a total of $N \geq n$ times, including n acceptances and $N - n$ rejections. Explain how you can use n, N and M to estimate c.

(c) (4 points) Use rejection sampling to generate a sample of size 10^3 from f and overlay a line plot of f atop a normalized histogram of your samples. Repeat this step with 10^6 samples. *Hint:* to plot f, first use your values of n, N and M to estimate c using your answer from part (b).

3. Graphical models (3 points)

Graphical models are often useful for modeling phenomena involving multiple variables. Consider the following scenario: suppose the probability that a burglar breaks into your car is π_b, and the probability that an innocent passerby accidentally touches your car is π_i. Let Z_b be a binary random variable that is 1 if there is a burglar, and 0 otherwise. Likewise, let Z_i be a binary random variable that is 1 if there is an innocent passerby, and 0 otherwise. Suppose Z_b and Z_i are independent of each other. Let X be a binary random variable that is 1 if your car alarm goes off. The probability your car alarm goes off depends on Z_b and Z_i. Draw the graphical model depicting the direct relationships between π_b, π_i, Z_b, Z_i, and X.