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 as MAP

In this problem, we work through the maximum *a posteriori* (MAP) interpretation of ridge regression. Suppose $x_1, \ldots, x_n \in \mathbb{R}^d$ are fixed feature vectors. Assume the linear model, where we observe

$$y_i = \beta^\top x_i + \varepsilon_i,\ i = 1, \ldots, n,$$

where $\varepsilon_i \sim N(0, \sigma^2)$ are independent of each other, and $\beta \in \mathbb{R}^d$ and $\sigma^2 > 0$ are unknown.

Let $y = (y_1, \ldots, y_n), \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$, and let X denote the matrix whose i-th row is equal to x_i. Using this notation, we may more succinctly write the linear model as

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

We model the regression weights as a random variable with the following prior distribution:

$$\beta \sim N(0, \sigma_\beta^2 I_d).$$

where $\sigma_\beta^2 > 0$ is hyperparameter we choose.

(a) Write the posterior distribution for β after observing the data, $p(\beta|X, y)$. *Hint:* use Bayes' rule and the probability density functions of multivariate Gaussians. Also use the fact that for a vector z, $z^T z = \|z\|_2^2$, where $\|z\|_2$ is the Euclidean norm of z.

(b) Show that the MAP estimator of β,

$$\hat{\beta}_{MAP} := \arg\max_\beta p(\beta|X, y)$$

solves the regularized least-squares problem,

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

with $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$. *Hint:* use part (a).

(c) In the regularized least-squares problem, λ is the regularization term: large values of λ penalize weight vectors with large norms. Since $\hat{\beta}_{MAP}$ is the solution to the regularized least-squares problem with $\lambda = \frac{\sigma^2}{\sigma_\beta^2}$, explain how our modeling decision (*i.e.*, choice of σ_β^2) influences $\hat{\beta}_{MAP}$.

[1] In Jupyter, you can download as PDF or print to save as PDF

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) 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) 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) 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. Gibbs Sampling

Graphical models are often useful for modeling phenomena involving multiple variables. In this problem, you'll formulate a graphical model, then demonstrate how to sample from the posterior using Gibbs sampling.

(a) 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, and is known to be:

Z_b	Z_i	$\mathbb{P}(X = 1 \mid Z_b, Z_i)$
0	0	0
0	1	0.05
1	0	0.85
1	1	0.90

Draw the graphical model depicting the direct relationships between π_b, π_i, Z_b, Z_i, and X.

(b) Suppose you know the parameters π_b and π_i, as well as $\mathbb{P}(X = 1 \mid Z_b, Z_i)$ as specified in Part (a). X is the observed variable, and Z_i and Z_b are the latent (unobserved) variables. We want to sample from $\mathbb{P}(Z_i, Z_b \mid X, \pi_b, \pi_i)$, the posterior over the latent variables conditioned on everything else. We'll use Gibbs sampling to do this:

(i) Suppose we are running Gibbs sampling, and on each iteration we sample Z_b first and then sample Z_i. We observed $X = 0$, and the values of Z_b and Z_i from iteration t are $Z_b^{(t)} = 0$ and $Z_i^{(t)} = 1$.

Derive the distribution used for the Gibbs sampling update of $Z_b^{(t+1)}$. Your solution should be in terms of π_b, π_i, and constants.

(ii) Now, suppose we draw $Z_b^{(t+1)} = 1$ from the distribution derived in Part (b.i). Derive the distribution used for the Gibbs sampling update of $Z_i^{(t+1)}$. Your solution should be in terms of π_b, π_i, and constants.