EEE 485-585 SPRING 2017 PROBLEM SET 2

Due Date: March 27 2017, 17:30

Question 1 [20 pts]

Consider Bayesian linear regression with a Gaussian prior on $\boldsymbol{\beta} \sim N(0, \tau^2 \mathbf{I}_p)$ and Gaussian likelihood $\boldsymbol{y} \sim N(\mathbf{X}\beta, \sigma^2 \mathbf{I}_n)$, where $\boldsymbol{y}$ denotes the response vector, $\mathbf{X}$ denotes the design matrix and $\mathbf{I}_j$ denotes the j by j identity matrix.

(a) [15 pts] Prove that the mode of the posterior distribution over $\boldsymbol{\beta}$ is equivalent to the ridge regression estimate.

(b) [5 pts] Find the relationship between the regularization parameter λ of ridge regression and the variances σ^2 and τ^2.

Question 2 [20 pts]

Consider n by p centered design matrix $\mathbf{X}$ with p-dimensional rows, where each row corresponds to a data instance. Consider the n dimensional response as a column vector $\boldsymbol{y}$. Let

$$\tilde{\mathbf{X}} = \begin{bmatrix} \mathbf{X} \\ 4\mathbf{I}_p \end{bmatrix}$$

and

$$\tilde{\boldsymbol{y}} = \begin{bmatrix} \boldsymbol{y} \\ \mathbf{0} \end{bmatrix}$$

where $\mathbf{I}_p$ denotes the p by p identity matrix and $\mathbf{0}$ denotes the p dimensional column vector of zeros.

(a) [5 pts] Compute the ordinary least squares estimate for the extended dataset $\tilde{\mathbf{X}}$ and $\tilde{\boldsymbol{y}}$.

(b) [5 pts] Compute the ridge regression solution for the original dataset $\mathbf{X}$ and $\boldsymbol{y}$ given the regularization parameter $\lambda > 0$.

(c) [10 pts] Compare the results in parts a and b. What can you say about the relation between these two results? Are they equivalent?

Question 3 [20 pts]

This question compares the behavior of Lasso and Ridge regression on a synthetic dataset. The dataset consists 1000 independent samples of a 9-dimensional feature vector $x = [x_1, \ldots, x_9]$ drawn from a uniform distribution on the interval $[-1, 1]$, along with the response

$$\begin{aligned} y &= -4x_1 - 3x_2 - 2x_3 - 1x_4 + 0x_5 + 1x_6 + 2x_7 + 3x_8 + 4x_9 + n \\ &= w^T x + n, \end{aligned}$$

where n is a standard Normal random variable. The file $question1data.txt$ consists of 1000 lines of 10 values. The first 9 columns represent $[x_1, \ldots, x_9]$, and the last column represents y. Each row consists of one sample. This problem explores the behavior of the estimated weights as the strength of the regularization (λ) varies. First centralize both the predictors and the response.

(a) [3 pts] Estimate the weights w using linear regression.

(b) [5 pts] Estimate the weights w using linear regression with Ridge regularization for various choices of λ. For each of the weights, plot the weight as a function of λ (start with $\lambda = 0$ and increase λ by 0.01 up to 3). Report the weights when $\lambda = 3$.

(c) [5 pts] Estimate the weights w using linear regression with Lasso regularization for various choices of λ. For each of the weights, plot the weight as a function of λ (start with $\lambda = 0$ and increase λ by 0.01 up to 3). Report the weights when $\lambda = 3$.

(d) [7 pts] As regularization parameter varies, how many of the estimated weights are exactly zero with Lasso regression? How many of the estimated weights are exactly zero with Ridge regression?

Question 4 [20 pts]

Consider $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ where y_is are assumed to be independent and

$$p(y_i|\boldsymbol{x}_i, \boldsymbol{\beta}) = \frac{1}{\sqrt{2}b} \exp\left(-\frac{\sqrt{2}|y_i - \boldsymbol{\beta}^t \boldsymbol{x}_i|}{b}\right)$$

for $b > 0$. Also assume that in the prior distribution β_j, $j = 1, \dots, p$ are independent and are distributed according to

$$p(\beta_j) = \frac{1}{\sqrt{2}a} \exp\left(-\frac{\sqrt{2}|\beta_j|}{a}\right)$$

for $a > 0$.

(a) [10 pts] Compute the log of the posterior distribution over $\boldsymbol{\beta}$.

(b) [5 pts] Propose a method to compute the MAP estimate of $\boldsymbol{\beta}$.

(c) [5 pts] As you have already observed, the likelihood in this case is not Gaussian as we have studied in the class. What is the advantage of using this likelihood function compared to the Gaussian likelihood function? Clearly explain your reasoning.

Question 5 [20 pts]

Consider the linear model $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^* + \epsilon_i$, $i = 1, 2, \dots, n$, where $\boldsymbol{x}_i \in \mathbb{R}^p$ are fixed predictor vectors, $\boldsymbol{\beta}^*$ is the fixed unknown parameter vector and ϵ_i are i.i.d. zero mean Gaussian errors with variance σ^2.

(a) [5 pts] Prove that $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$ is an unbiased estimate of $\boldsymbol{\beta}^*$, where $\mathbf{X}$ represents the design matrix.

(b) [5 pts] Also show that for a given predictor vector $\boldsymbol{x}_0$, the estimate $\hat{y}_0 = \boldsymbol{x}_0^T \hat{\boldsymbol{\beta}}$ is unbiased.

(c) [5 pts] For the estimates $\hat{y}_i = \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}$, $i = 1, \dots, n$, prove that the average variance $\frac{1}{n}\sum_{i=1}^n \text{Var}(\hat{y}_i)$ is equal to $\sigma^2 p/n$.

(d) [5 pts] As one can see from part c, the variance can be very large if $p > n$. This might result in high MSE on the test data. However, if one seek to reduce the MSE by reducing the variance, the bias will increase. Explain how this problem can be solved.