EEE 485-585 FALL 2018 PROBLEM SET 2

Due Date: Wednesday, 14 November 2018, 17:00 (room EE-212)

Question 1 [25 pts]

The true model is given as $y_i = \boldsymbol{x}_i^T \boldsymbol{\beta}^{\text{true}} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, $i = 1, \ldots, n$, and ϵ_i and ϵ_j are independent for all $i \neq j$.

Let $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ be the centered response vector and $\mathbf{X}$ be the centered design matrix, whose ith row is $\boldsymbol{x}_i^T$, where $\boldsymbol{x}_i$ is a p-dimensional predictor (i.e., ith data instance). Let $\hat{\boldsymbol{\beta}} = [\hat{\beta}_1, \ldots, \hat{\beta}_p]^T$ denote the ridge estimator given regularization parameter $\lambda > 0$. Assuming that $\boldsymbol{x}_i$s are fixed

(a) [10 pts] Compute $\mathbb{E}[\hat{\boldsymbol{\beta}}] - \boldsymbol{\beta}^{\text{true}}$. What happens to this quantity as λ increases? What will be this quantity in the limit $\lambda \to \infty$.

(b) [6 pts] Assume that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Compute the covariance matrix of the ordinary least squares estimator.

(c) [6 pts] Assume that $\mathbf{X}^T\mathbf{X} = \mathbf{I}_p$. Compute the covariance matrix of the ridge estimator.

(d) [3 pts] Compare the variance (trace of the covariance matrix) of the ordinary least squares estimator with the ridge estimator for the above cases.

Question 2 [25 pts]

(a)[10 pts] Consider binary logistic regression. Given $\mathcal{D} = (x_i, y_i)_{i=1}^n$ compute the negative log likelihood function $-\log L(\beta_0, \beta_1)$. We will find the minimizer of the negative log likelihood using gradient descent. Compute the update rule for gradient descent in terms of $\boldsymbol{y} = [y_1, \ldots, y_n]^T$, $\mathbf{X} = [1\ x_1; 1\ x_2; \ldots; 1\ x_n]$ and $\boldsymbol{\pi} = [\pi_1, \ldots, \pi_n]^T$, where $\pi_i = \Pr(Y_i = 1 | X_i = x_i)$.

(b)[5 pts] Compute the negative log likelihood of logistic regression for $\mathcal{D} = (\tilde{x}_i, y_i)_{i=1}^n$ for $\tilde{x}_i = 10x_i$. State the relation between the minimizers of the negative log likelihood in parts a and b?

(c)[10 pts] Download student.csv from moodle. Take the first column as the response variable Y and the second column as the predictor variable X. Train your model using gradient descent until convergence. What is $\Pr(Y = 1 | X = x)$ and $\Pr(Y = 0 | X = x)$? Plot $\Pr(Y = 1 | X = x)$ as a function of x from 0 to 800.

Question 3 [25pts]

Download the training dataset PS2Q3train and test dataset PS2Q3test from Moodle. X is the predictor and Y is the response. We will use polynomial regression where the degree of the polynomial, i.e., p is our hyperparameter.

(a)[7 pts] Use least squares to fit the model. Draw on a figure the mean squared training and test error curves as a function of p for $p \in \{0, \ldots, 10\}$. What is the optimal value of p? Justify your reasoning.

(b)[7 pts] For the optimal value of p that you have found in part a, use ridge regression to fit the model. Draw on a figure the mean squared training and test error curves as a function of λ (regularization parameter) for $\lambda \in [0, 100]$. What is the optimal value of λ? Justify your reasoning.

(c)[7 pts] For the optimal value of p that you have found in part a, use lasso regression to fit the model. Draw on a figure the mean squared training and test error curves as a function of λ (regularization parameter) for $\lambda \in [0, 100]$. What is the optimal value of λ? Justify your reasoning.

(d)[4 pts] Report the estimated coefficients found using the training dataset using the procedures described in part b and part c for their corresponding optimal λ values. Compare the coefficients in terms of their sparseness.

Question 4 [25 pts]

(a)[10 pts] Consider Poisson regression. Given $\mathcal{D} = (x_i, y_i)_{i=1}^n$ compute the negative log likelihood function $-\log L(\beta_0, \beta_1)$. We will find the minimizer of the negative log likelihood using gradient descent. Write down the update rule for gradient descent in terms of $\boldsymbol{y} = [y_1, \ldots, y_n]^T$ and $\mathbf{X} = [1\ x_1; 1\ x_2; \ldots; 1\ x_n]$.

(b)[10 pts] Download PS2Q4 from Moodle. Train a Poisson regression model on the entire dataset using gradient descent. Report the MLE estimate of $\hat{\beta}_0$ and $\hat{\beta}_1$.

(c)[5 pts] Plot the probability that $Y \leq 10$ as a function of $X = x$ for x between 0 and 5.