ESE 541 Week 8 Practice Problems - Solutions

Problem 1: Consider the following training and testing datasets with $p = 3$ and $N = 4$:

$$\mathcal{D}_{\text{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^4 = \left\{ \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, 0.9 \right), \left(\begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}, 2.1 \right), \left(\begin{bmatrix} 3 \\ 9 \\ 27 \end{bmatrix}, 2.9 \right), \left(\begin{bmatrix} 4 \\ 16 \\ 64 \end{bmatrix}, 4.1 \right) \right\}$$

$$\mathcal{D}_{\text{Te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^4 = \left\{ \left(\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, 1.1 \right), \left(\begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix}, 1.9 \right), \left(\begin{bmatrix} 3 \\ 9 \\ 27 \end{bmatrix}, 3.1 \right), \left(\begin{bmatrix} 4 \\ 16 \\ 64 \end{bmatrix}, 3.9 \right) \right\}$$

a) Find the best linear model using best subset selection. Feel free to use software to find the parameters.

We find that the best model, as selected by best subset selection, is:

$$M_1 = \beta_0 + \beta_1 x_1$$

b) Find a linear model using forward stepwise selection. Feel free to use software to find the parameters.

We find that the best model, as selected by forward stepwise selection, is:

$$M_1 = \beta_0 + \beta_1 x_1$$

c) Find a linear model using backward stepwise selection. Feel free to use software to find the parameters.

We find that the best model, as selected by backward stepwise selection, is:

$$M_2 = \beta_0 + \beta_1 x_1$$

d) Do the models in a), b), and c) coincide?

All 3 models are the same!

Problem 2: Consider the following training and testing datasets with $p = 3$ and $N = 4$:

$$\mathcal{D}_{\text{Tr}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^4 = \left\{ \left(\begin{bmatrix} 1 \\ 1.1 \\ 0.9 \end{bmatrix}, 0.9 \right), \left(\begin{bmatrix} 2 \\ 1.9 \\ 2.1 \end{bmatrix}, 2.1 \right), \left(\begin{bmatrix} 3.1 \\ 3 \\ 2.9 \end{bmatrix}, 2.9 \right), \left(\begin{bmatrix} 3.9 \\ 4 \\ 4.1 \end{bmatrix}, 4.1 \right) \right\}$$

$$\mathcal{D}_{\text{Te}} = \{(\mathbf{x}_i, y_i)\}_{i=1}^4 = \left\{ \left(\begin{bmatrix} 1 \\ 1.1 \\ 0.9 \end{bmatrix}, 1.1 \right), \left(\begin{bmatrix} 2 \\ 1.9 \\ 2.1 \end{bmatrix}, 1.9 \right), \left(\begin{bmatrix} 3.1 \\ 3 \\ 2.9 \end{bmatrix}, 3.1 \right), \left(\begin{bmatrix} 3.9 \\ 4 \\ 4.1 \end{bmatrix}, 3.9 \right) \right\}$$

a) Find the coefficients of the linear model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ and compute the resulting training and testing RSS.

$$\hat{\beta} = (M_X^T M_X)^{-1} M_X^T \mathbf{y}$$

$$M_X = \begin{bmatrix} 1 & 1 & 1.1 & 0.9 \\ 1 & 2 & 1.9 & 2.1 \\ 1 & 3.1 & 3 & 2.9 \\ 1 & 3.9 & 4 & 4.1 \end{bmatrix}$$

$$\mathbf{y} = \begin{bmatrix} 0.9 \\ 2.1 \\ 2.9 \\ 4.1 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

$$RSS_{Tr} = (0.9 - 0.9)^2 + (2.1 - 2.1)^2 + (2.9 - 2.9)^2 + (4.1 - 4.1)^2 = 0$$

$$RSS_{Te} = (0.9 - 1.1)^2 + (2.1 - 1.9)^2 + (2.9 - 3.1)^2 + (4.1 - 3.9)^2 = 4 \cdot 0.04 = 0.16$$

b) Estimate the covariance matrix of X using the formula:

$$\widehat{\Sigma}_X = \frac{1}{4} \sum_{i=1}^{4} (\mathbf{x}_i - \widehat{\mu}_X)(\mathbf{x}_i - \widehat{\mu}_X)^\intercal \in \mathbb{R}^{3\times 3}$$

where $\widehat{\mu}_X = \frac{1}{4}\sum_{i=1}^{4} \mathbf{x}_i \in \mathbb{R}^3$.

$$\mu_X = \begin{bmatrix} 2.5 \\ 2.5 \\ 2.5 \end{bmatrix}$$

$$\widehat{\Sigma}_X = \frac{1}{4} \sum_{i=1}^{4} \left(\mathbf{x}_i - \begin{bmatrix} 2.5 \\ 2.5 \\ 2.5 \end{bmatrix} \right) \left(\mathbf{x}_i - \begin{bmatrix} 2.5 \\ 2.5 \\ 2.5 \end{bmatrix} \right)^\intercal = \begin{bmatrix} 1.205 & 1.200 & 1.270 \\ 1.200 & 1.205 & 1.270 \\ 1.27 & 1.270 & 1.360 \end{bmatrix}$$

c) Compute the largest eigenvalue and the associated eigenvector $\mathbf{v}_1$ (feel free to use computer software). Using $\mathbf{v}_1$, generate a reduced order dataset

$$\mathcal{D}_1 = \{(z_i, y_i)\}_{i=1}^4 \text{ where } z_i = \mathbf{v}_1^\intercal \mathbf{x}_i \in \mathbb{R}$$

$$\lambda_1 = 3.753$$

$$\mathbf{v}_1 = \begin{bmatrix} 0.565 \\ 0.565 \\ 0.600 \end{bmatrix}$$

$$z_i = \mathbf{v}_1^\intercal \mathbf{x}_i$$

$$\mathcal{D}_1 = \{(1.727, 0.9), (3.466, 2.1), (5.190, 2.9), (6.928, 4.1))\}$$

d) Using the dataset $\mathcal{D}_1$ in c), find the coefficients of the linear model $Y = \theta_0 + \theta_1 Z$ and compute the resulting training and testing RSS.

$$\hat{\beta} = (M_Z^T M_Z)^{-1} M_Z^T \mathbf{y}$$

$$M_Z = \begin{bmatrix} 1 & 1.727 \\ 1 & 3.466 \\ 1 & 5.190 \\ 1 & 6.928 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} -0.098 \\ 0.600 \end{bmatrix}$$

$$y_{pred} = \begin{bmatrix} 0.938 \\ 1.981 \\ 3.016 \\ 4.059 \end{bmatrix}$$

$$RSS_{Tr} = (0.9 - 0.938)^2 + (2.1 - 1.981)^2 + (2.9 - 3.016)^2 + (4.1 - 4.059)^2 = 0.030$$

$$RSS_{Te} = (1.1 - 0.938)^2 + (1.9 - 1.981)^2 + (3.1 - 3.016)^2 + (3.9 - 4.059)^2 = 0.065$$

e) Compare the training and testing RSS computed in a) with the one computed in d).

We find that we achieve a worse training RSS after PCA, but yield a better test RSS.

Problem 3: Which of the following statements about regularization are true? Check all that apply.

a) Using a very large value of λ cannot hurt the performance of your hypothesis. We just do not set λ to be too large to avoid computational problems.

False. Using a large value of λ can result in underfitting the data, and thus, hurt performance.

b) Using too large a value of λ can cause your hypothesis to underfit the data.

True. A large value of λ results in a large regularization penalty and thus a strong preference for simpler models which can underfit the data.

c) Because logistic regression outputs values between 0 and 1, its range of outputs are always shrunk with an increasing λ.

False. Having an increasing λ results in greater penalty for having large coefficients. However, this doesn't necessarily mean the outputs get smaller as well. For instance, shrinking a negative coefficient would result in greater output.

Problem 4: You are training a classification model with logistic regression. Which of the following statements are true? Mark all that apply.

a) Adding many new features to the model helps prevent overfitting on the training set.

False. Adding many new features gives us more expressive models which are able to better fit out training set. If too many new features are added, this can lead to overfitting of the training set.

b) Introducing regularization to the model always results in equal or better performance on examples not in the training set.

False. If we introduce too much regularization, we can underfit the training set and this can lead to worse performance even for examples not in the training set.

c) Introducing regularization to the model always results in equal or better performance on the training set.

False. If we introduce too much regularization, we can underfit the training set and have worse performance on the training set.

d) Adding a new feature to the model always results in equal or better performance on the training set.

True. Adding many new features gives us more expressive models which are able to better fit out training set. If too many new features are added, this can lead to overfitting of the training set.