STA314: Statistical Methods for Machine Learning I

Midterm Exam – LEC0101

Problem 1 (6 pts)

Let $(y_i, \mathbf{x}_i)$, for $1 \leq i \leq n$, be the training data points where $y_i \in \mathbb{R}$ is the response and $\mathbf{x}_i = (x_{i1}, x_{i2}, x_{i3}) \in \mathbb{R}^3$ contains measurements of three features.

Throughout the exam, let us ignore the intercept term for fitting any model.

(a) (1 pt) Write down the form of linear predictor, fitted by the Ordinary Least Squares (OLS) approach, using all three features. (You need to specify how the coefficients are obtained.) Denote this predictor by $\widehat{f}_1$.

SOLUTION: For any given data point $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)$, we have

$$\widehat{f}_1(\mathbf{x}^*) = \sum_{i=1}^{3} x_i^* \widehat{\beta}_i$$

where the estimated coefficients are obtained from

$$\widehat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}=(\beta_1,\beta_2,\beta_3)}{\arg\min} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i^\top \boldsymbol{\beta}\right)^2.$$

(b) (1 pt) For any unit increment in the first feature, state how the predicted outcome of $\widehat{f}_1$ changes.

SOLUTION: Change by $\widehat{\beta}_1$.

(c) (1 pt) In the case there is a prior belief that the first feature also affects the response quadratically. Write down the form of linear predictor, fitted by OLS, using four features $(x_{i1}, x_{i2}, x_{i3}, x_{i1}^2)$. (Specify how the coefficients are obtained.) Denote this predictor by $\widehat{f}_2$.

SOLUTION: For any given data point $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)^\top$, we have

$$\widehat{f}_2(\mathbf{x}^*) = \sum_{i=1}^{3} x_i^* \widehat{\alpha}_i + x_1^{*2} \widehat{\alpha}_4$$

where

$$\widehat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}=(\alpha_1,\ldots,\alpha_4)}{\arg\min} \sum_{i=1}^{n} \left(y_i - x_{i1}\alpha_1 - x_{i2}\alpha_2 - x_{i3}\alpha_3 - x_{i1}^2\alpha_4\right)^2.$$

(d) (1 pt) For any unit increment in the first feature, state how the predicted outcome of $\widehat{f}_2$ changes.

SOLUTION: Change by $\widehat{\alpha}_1 + (2x_1^* + 1)\widehat{\alpha}_4$ at $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)$.

(e) (2 pts) Can you tell which $\widehat{f}_1$ and $\widehat{f}_2$ has smaller training Mean Squared Error (MSE)? Give your reasoning.

SOLUTION: $\widehat{f}_2$ has smaller training MSE. By the optimality of the OLS solution, we have

$$\begin{aligned}
&\sum_{i=1}^{n} \left(y_i - x_{i1}\widehat{\alpha}_1 - x_{i2}\widehat{\alpha}_2 - x_{i3}\widehat{\alpha}_3 - x_{i1}^2\widehat{\alpha}_4\right)^2 \\
&\leq \sum_{i=1}^{n} \left(y_i - x_{i1}\alpha_1 - x_{i2}\alpha_2 - x_{i3}\alpha_3 - x_{i1}^2\alpha_4\right)^2
\end{aligned}$$

for any $(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$. The left-hand-side (LHS) is the training MSE of $\widehat{f}_2$. In particular, by taking $(\alpha_1, \alpha_2, \alpha_3, \alpha_4) = (\widehat{\beta}_1, \widehat{\beta}_2, \widehat{\beta}_3, 0)$, the RHS becomes the training MSE of $\widehat{f}_1$, concluding the claim.

Problem 2 (5 pts)

Consider the same settings in **Problem 1** as well as the predictors $\widehat{f}_1$ in part (a) and $\widehat{f}_2$ in part (c). Suppose $(y_i, \mathbf{x}_i)$ are i.i.d. realizations of (Y, X) with $X = (X_1, X_2, X_3)$, which satisfy the following model

$$Y = 5X_1 + \frac{1}{2}X_2 + \varepsilon. \tag{1}$$

(a) (1 pt) Is the predictor $\widehat{f}_1$ unbiased at any given test data point $\mathbf{x}^* = (x_1^*, x_2^*, x_3^*)$? If so, prove it. Otherwise, state your reasoning. (You may use the fact that the OLS estimators are unbiased without proving it.)

SOLUTION: Unbiased, because

$$\begin{aligned}\mathbb{E}[\widehat{f}_1(\mathbf{x}^*)] &= x_1^*\mathbb{E}[\widehat{\beta}_1] + x_2^*\mathbb{E}[\widehat{\beta}_2] + x_2^*\mathbb{E}[\widehat{\beta}_3] \\ &= x_1^*\mathbb{E}[\beta_1] + x_2^*\mathbb{E}[\beta_2] + x_3^* \cdot 0 \\ &= 5x_1^* + \frac{1}{2}x_2^*\end{aligned}$$

which is the true function value at $\mathbf{x}^*$.

(b) (1 pt) Can you tell which $\widehat{f}_1$ and $\widehat{f}_2$ has smaller test MSE under the model in Eq. (1)? Give your reasoning.

SOLUTION: $\widehat{f}_1$. Both are unbiased but $\widehat{f}_1$ has smaller variance because it uses one less feature.

(c) (2 pts) Consider using the backward stepwise selection to select a subset from (X_1, X_2, X_3, X_1^2) for prediction. If the excluded feature is X_3 after the first step, write down all the candidate predictors considered in the second step. (You may write $Y \sim X_1 + X_2 + X_3 + X_1^2$ for a predictor that uses all four features or $Y \sim 0$ for no feature.)

SOLUTION: The candidate predictors considered in the second step are

$$Y \sim X_1 + X_2 \qquad Y \sim X_1 + X_1^2 \qquad Y \sim X_2 + X_1^2$$

(d) (1 pt) Indicate which feature you would expect to be excluded after the second step in part (c), and explain.

SOLUTION: We expect X_1^2 to be excluded as the true model is the first one.

Problem 3 (9 pts)

In a regression problem, assume that the true model is

$$Y = X_2 + X_3 + X_3^2 + \varepsilon,$$

where ε is a random noise. Suppose we fit the following two fitted models by using the training data containing n realizations of (Y, X_1, X_2, X_3)

(M1) $Y \sim X_2 + X_3 + X_3^2$,

(M2) $Y \sim X_1 + X_2 + X_3 + X_1^2 + X_2^2 + X_3^2$.

Here the notation $Y \sim X + X'$ means to regress Y onto X and X' via the OLS approach. For each fitted model, we can construct an estimator of the regression function, denoted by $\widehat{f}_i$ for $i \in \{1, 2\}$.

(a) (3 pts) Describe how to use 2-fold cross-validation (CV) to estimate the expected MSE of $\widehat{f}_1$.

SOLUTION: Key words include:

- Random split (50/50)
- Fit the model on training and compute validation MSE
- Swap and compute the averaged validation MSEs

(b) (2 pts) Let $\widehat{\text{MSE}}_2(\widehat{f}_1)$ be the estimate of the expected MSE of $\widehat{f}_1$ obtained from 2-fold CV. Do you expect to obtain the same value of $\widehat{\text{MSE}}_2(\widehat{f}_1)$ if running 2-fold CV multiple times? If yes, explain why. Otherwise, describe how to understand the different values.

SOLUTION: We will have different values. They are all estimates of the true $\text{MSE}(\widehat{f}_1)$. Their average is close to $\text{MSE}(\widehat{f}_1)$.

(c) (2 pts) Let $\widehat{\text{MSE}}_k(\widehat{f}_1)$ be the estimate of the expected MSE of $\widehat{f}_1$ obtained by k-fold CV. Comment on the role of k.

SOLUTION: The larger k is , the more accurately $\widehat{\text{MSE}}_k(\widehat{f}_1)$ estimates $\text{MSE}(\widehat{f}_1)$. However, the computation cost becomes higher for large k.

(d) (1 pt) Between $\widehat{\mathrm{MSE}}_n(\widehat{f}_1)$ and $\widehat{\mathrm{MSE}}_n(\widehat{f}_2)$, which one do you expect to be smaller? State your reasoning.

SOLUTION: $\widehat{\mathrm{MSE}}_n(\widehat{f}_1)$ should be smaller as M_1 corresponds to the true model.

(e) (1 pt) Suppose we first sort the n data points according to their response values in non-decreasing order, and then perform n-fold CV to obtain $\widehat{\mathrm{MSE}}_n(\widehat{f}_1)$ and $\widehat{\mathrm{MSE}}_n(\widehat{f}_2)$. Would you expect the same conclusion as in part (d)?

SOLUTION: LOOCV does not change as we shuffle the data.

Problem 4 (3 pts)

Suppose we have the data $(y_i, \mathbf{x}_i)$ for $1 \le i \le n$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}$. Let $\widehat{\boldsymbol{\beta}}$ be the estimated coefficients by regressing $(y_1, \ldots, y_n)$ onto $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ via OLS. For any given $\lambda > 0$, let $\widehat{\boldsymbol{\beta}}_\lambda^R$ be the estimated coefficients by regressing $(y_1, \ldots, y_n)$ onto $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$ via the ridge.

Denote by $\mathrm{RSS}(\widehat{\boldsymbol{\beta}})$ the residual sum of squares (RSS) of the linear predictor that uses $\widehat{\boldsymbol{\beta}}$. Similarly, we write $\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_\lambda^R)$.

(a) (1 pt) Show that

$$\mathrm{RSS}(\widehat{\boldsymbol{\beta}}) \le \mathrm{RSS}(\widehat{\boldsymbol{\beta}}_\lambda^R).$$

SOLUTION: Note that

$$\mathrm{RSS}(\widehat{\boldsymbol{\beta}}) = \sum_{i=1}^n \left(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}\right)^2, \quad \mathrm{RSS}(\widehat{\boldsymbol{\beta}}_\lambda^R) = \sum_{i=1}^n \left(y_i - \mathbf{x}_i^\top \widehat{\boldsymbol{\beta}}_\lambda^R\right)^2$$

The claim follows by the optimality of $\widehat{\boldsymbol{\beta}}$, that is, $\mathrm{RSS}(\widehat{\boldsymbol{\beta}}) \le \mathrm{RSS}(\boldsymbol{\beta})$ for any $\boldsymbol{\beta} \in \mathbb{R}^p$, in particular, for $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}_\lambda^R$.

(b) (2 pts) Prove that, for all $\lambda > 0$,

$$\|\widehat{\boldsymbol{\beta}}_\lambda^R\|_2 \leq \|\widehat{\boldsymbol{\beta}}\|_2.$$

(For a vector $v \in \mathbb{R}^p$, we write $\|v\|_2^2 = \sum_{j=1}^p v_j^2$.)

SOLUTION: Since

$$\widehat{\boldsymbol{\beta}}_\lambda^R = \arg\min_{\boldsymbol{\beta}} \mathrm{RSS}(\boldsymbol{\beta}) + \lambda\|\boldsymbol{\beta}\|_2^2,$$

we have

$$\mathrm{RSS}(\widehat{\boldsymbol{\beta}}_\lambda^R) + \lambda\|\widehat{\boldsymbol{\beta}}_\lambda^R\|_2^2 \leq \mathrm{RSS}(\widehat{\boldsymbol{\beta}}) + \lambda\|\widehat{\boldsymbol{\beta}}\|_2^2.$$

Rearranging terms gives

$$\begin{aligned} \lambda\left(\|\widehat{\boldsymbol{\beta}}_\lambda^R\|_2^2 - \|\widehat{\boldsymbol{\beta}}\|_2^2\right) &\leq \mathrm{RSS}(\widehat{\boldsymbol{\beta}}) - \mathrm{RSS}(\widehat{\boldsymbol{\beta}}_\lambda^R) \\ &\leq 0 \qquad \text{by part (a).} \end{aligned}$$

Since $\lambda > 0$, the claim follows.

Problem 5 (7 pts)

In the following problem we will see the benefit of shrinking a given unbiased estimator.
Let $\widehat{\beta}$ be an unbiased estimator of $\beta \in \mathbb{R}$ and suppose that $\mathrm{Var}(\widehat{\beta}) = \sigma^2$ for some $\sigma^2 > 0$. For a given $\lambda \geq 0$, define the shrinkage version of $\widehat{\beta}$ as

$$\widehat{\beta}_\lambda := \frac{\widehat{\beta}}{1+\lambda}.$$

Denote the expected MSE of $\widehat{\beta}$ by $\mathrm{MSE}(\widehat{\beta}) := \mathbb{E}[(\widehat{\beta} - \beta)^2]$ and similarly, $\mathrm{MSE}(\widehat{\beta}_\lambda) := \mathbb{E}[(\widehat{\beta}_\lambda - \beta)^2]$.

(a) (1 pt) For any estimator $\widetilde{\beta}$ of β, prove that

$$\mathrm{MSE}(\widetilde{\beta}) = \mathrm{Var}(\widetilde{\beta}) + \Big(\mathbb{E}[\widetilde{\beta}] - \beta\Big)^2.$$

(You may assume the first two moments of $\widetilde{\beta}$ exist.)

SOLUTION:

$$\begin{aligned}
\mathrm{MSE}(\widetilde{\beta}) &= \mathbb{E}[(\widetilde{\beta} - \beta)^2] \\
&= \mathbb{E}\left[\Big(\widetilde{\beta} - \mathbb{E}[\widetilde{\beta}] + \mathbb{E}[\widetilde{\beta}] - \beta\Big)^2\right] \\
&= \mathbb{E}\left[\Big(\widetilde{\beta} - \mathbb{E}[\widetilde{\beta}]\Big)^2\right] + \mathbb{E}\left[\Big(\mathbb{E}[\widetilde{\beta}] - \beta\Big)^2\right] \\
&\quad + 2\mathbb{E}\left[\Big(\widetilde{\beta} - \mathbb{E}[\widetilde{\beta}]\Big)\Big(\mathbb{E}[\widetilde{\beta}] - \beta\Big)\right] \\
&= \mathbb{E}\left[\Big(\widetilde{\beta} - \mathbb{E}[\widetilde{\beta}]\Big)^2\right] + \Big(\mathbb{E}[\widetilde{\beta}] - \beta\Big)^2.
\end{aligned}$$

By definition, the first term is $\mathrm{Var}(\widetilde{\beta})$.

(b) (2 pts) Use part (a) to deduce the expressions of $\mathrm{MSE}(\widehat{\beta})$ and $\mathrm{MSE}(\widehat{\beta}_\lambda)$.

SOLUTION: By part (a), we have

$$\mathrm{MSE}(\widehat{\beta}) = \mathrm{Var}(\widehat{\beta}) + \Big(\mathbb{E}[\widehat{\beta}] - \beta\Big)^2 = \sigma^2$$

and

$$\begin{aligned}\mathrm{MSE}(\widehat{\beta}_\lambda) &= \mathrm{Var}(\widehat{\beta}_\lambda) + \Big(\mathbb{E}[\widehat{\beta}_\lambda] - \beta\Big)^2 \\ &= \frac{\mathrm{Var}(\widehat{\beta})}{(1+\lambda)^2} + \left(\frac{1}{1+\lambda}\mathbb{E}[\widehat{\beta}] - \beta\right)^2 \\ &= \frac{\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2\beta^2}{(1+\lambda)^2}.\end{aligned}$$

(c) (1 pt) Comment on the effect of λ.

SOLUTION: For larger λ, the bias of $\widehat{\beta}_\lambda$ gets larger while the variance gets smaller.

(d) (3 pts) Find the best λ that minimizes $\text{MSE}(\widehat{\beta}_\lambda)$ and its corresponding $\text{MSE}(\widehat{\beta}_\lambda)$. Comment on the choice of the best λ and compare the best $\text{MSE}(\widehat{\beta}_\lambda)$ with $\text{MSE}(\widehat{\beta})$.

SOLUTION: Minimizing $\text{MSE}(\widehat{\beta}_\lambda)$ over $\lambda \geq 0$ gives that

$$\lambda^* = \frac{\sigma^2}{\beta^2}$$

and the corresponding mimimal MSE

$$\text{MSE}(\widehat{\beta}_{\lambda^*}) = \frac{\sigma^2\beta^2}{\beta^2 + \sigma^2}.$$

Fix σ^2. The smaller β is, the larger λ^* becomes, while for fixed β, λ^* gets larger as σ^2 increases.

Clearly,

$$\text{MSE}(\widehat{\beta}) - \text{MSE}(\widehat{\beta}_{\lambda^*}) = \sigma^2 \frac{\sigma^2}{\beta^2 + \sigma^2} \geq 0$$

and the difference becomes larger as β^2 gets smaller or σ^2 gets larger.

Problem 6 (10 pts, 1 pt for each subquestion)

Be sure to mark your answers on the answer sheet of multiple choice questions. There can be **one to four** correct answers to each question. One point is assigned to a multiple choice question **if and only if** all correct answers to this question are checked and no incorrect answer to this question is checked.

1. Which of the following statements about model evaluation are true?

   A The training MSE is always higher than the test MSE.

   B Cross-validation helps in assessing prediction performance of the model on test data.

   C Overfitting occurs when the model is too simple.

   D A lower training error guarantees better test performance.

   SOLUTION: B

2. Which of the following statements about regression are true?

   A The coefficients in a multiple linear regression model indicate the effect of one predictor while holding others constant.

   B Multicollinearity can lead to unstable estimates of regression coefficients.

   C Adding more predictors to a model will always improve its prediction performance.

   D The adjusted R-squared value accounts for the number of predictors in the model.

   SOLUTION: ABD

3. Regarding regularization techniques in regression, which of the following statements are true?

   A Lasso regression can shrink estimated coefficients to exactly zero.

   B Ridge regression penalizes the sum of the absolute values of coefficients.

   C Regularization can help mitigate overfitting.

   D Ridge tends to have better prediction performance than lasso when the true regression coefficients are non-sparse

   SOLUTION: ACD

4. Which of the following statements regarding model selection criteria are true?

   A AIC adjusted the RSS by additively penalizing the number of parameters.

   B BIC typically results in more complex models compared to AIC.

   C Cross-validation is more applicable than AIC and BIC.

   D Lower values of AIC and BIC indicate a better fitted model.

   SOLUTION: ACD

5. Which of the following statements are true

   A Best subset selection is guaranteed to find the best model.

   B Forward stepwise selection is a greedy approach.

   C Backward stepwise selection considers the same set of models as the forward stepwise selection.

   D Best subset selection can always be used in practice.

   SOLUTION: AB

6. Which of the following statements are true

   A Image object detection is a supervised learning problem.

   B Regression problems belong to supervised learning while classification problems belong to unsupervised learning.

   C k-nearest neighbor regression is an example of non-parametric method for estimating the regression function.

   D Shrinkage regression is an example of parametric method.

   SOLUTION: ACD

7. Which of the following statements are true

   A Ridge can yield a smaller training MSE than the OLS estimator.

   B Ridge estimator has smaller variance than the OLS estimator.

   C Ridge can outperform Lasso in terms of prediction.

   D Ridge with a selected regularization parameter is computationally much more expensive than the OLS estimator.

   SOLUTION: BC

8. In which case, we usually prefer nonparametric approaches rather than linear approaches

   A When the number of features is large.

   B When the sample size is large.

   C When the data has little noise.

   D When we want to model the trend between the features and the response.

   SOLUTION: BC

9. Which of the following statements about local regression are true?

   A A smaller size of the neighborhood usually yields a more flexible fitted model.

   B A smaller size of the neighborhood usually yields a fitted model with smaller test MSE.

   C Choosing the neighborhood size is a model selection problem.

   D Local regression method can only be used when the number of features is small.

   SOLUTION: ACD

10. Which of the following statements are true?

    A OLS can be used to fit polynomial regression

    B Step function approach needs to specify cutoff points

    C The fitted step function is piecewise constant

    D Linear splines are piecewise continuous linear functions

    SOLUTION: ABCD

(You may use this page as scratch paper if needed)

(You may use this page as scratch paper if needed)

(You may use this page as scratch paper if needed)

(You may use this page as scratch paper if needed)