STA314: Statistical Methods for Machine Learning I

Midterm Exam – LEC0101

Problem 1 (6 points)

Assume that we analyze the `Carseats` data set. The goal is to predict `Sales`. Based on the following output of R, answer the questions.

```
Call:
lm(formula = Sales ~ Income + Advertising + Price + US
                     + Advertising:US,  data = Carseats)

Coefficients:
                   Estimate   Std.Error  t-value  Pr(>|t|)
(Intercept)       12.205105   0.688920  17.716   <2e-16 ***
Income             0.010972   0.004298   2.553   0.0111 *
Advertising        0.043718   0.122387   0.357   0.7211
Price             -0.053712   0.005069 -10.596   <2e-16 ***
USNo              -0.075873   0.360800  -0.210   0.8335
Advertising:USNo   0.079992   0.124952   0.640   0.5224
---
Residual standard error: 2.387 on 394 degrees of freedom
Multiple R-squared:  0.2945,Adjusted R-squared:  0.2856
F-statistic:  32.9 on 5 and 394 DF,  p-value: < 2.2e-16
```

(1) (1 point) The feature `US` is a factor with two levels: Yes or No. Based on the above output, write down the way the feature `US` is encoded.

SOLUTION: USNo $= 1\{\text{US} = \text{No}\}$.

(2) (1 point) In this linear regression model, how do you interpret the coefficient of `USNo`?

SOLUTION: It represents the unit change of `Sales` for a non-US country comparing to the US when `Advertising` equals 0 and other features are held fixed.

(3) (1 point) How do you interpret the coefficient of `Advertising:USNo`?

SOLUTION: It means the difference of unit-change of `Sales` between non-US countries and the US for one unit increase in `Advertising` with other features held fixed.

(4) (1 point) Based on the R output, can you conclude whether or not `Advertising` is significant for predicting `Sales` at 0.05 significance level? Please state your reasoning.

SOLUTION: No. Even though the coefficient of `Advertising` is not significant, the effect of `Advertising` also depends on the coefficient of the interaction term.

(5) (2 points) Construct the 95% confidence interval for the coefficient of `Income`. (Writing out the expression suffices. You don't need to calculate the exact values.) Interpret the meaning of a 95% confidence interval. (You may assume the estimated coefficient is normal and use $\mathbb{P}\{Z \leq 1.96\} \approx 0.975$ for $Z \sim N(0,1)$)

SOLUTION: The 95% CIs is

$$[0.010972 \pm 1.96 \times 0.004298]\,.$$

It means if we sample 100 times of the training data and repeat our procedure to obtain 100 corresponding 95% confidence intervals as above, there would be around 95% of these CIs contain the true coefficient of `Income`. (PS: other reasonable interpretations are allowed).

Problem 2 (8 points)

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 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 Ordinary Least Squares (OLS) approach. Under each model, we can construct an estimator of the regression function, denoted by $\widehat{f_i}$ for $i \in \{1, 2\}$.

Please compare the two models, and give a short explanation, in terms of the following aspects. (For example, M1 has larger variance than M2 or there is no sufficient information about the comparison.)

(a) (2 points) squared bias of $\widehat{f_i}$

SOLUTION: Both M1 and M2 have no bias. They include all features in the true model.

(b) (2 points) variance of $\widehat{f_i}$

SOLUTION: M2 has a larger variance than M1 as it uses additional features.

(c) (2 points) the test MSE of $\widehat{f}_i$

SOLUTION: M1 yields a smaller test MSE based on (a) and (b).

(d) (2 points) the training MSE of $\widehat{f}_i$

SOLUTION: M2 has a smaller training MSE as adding additional features leads to smaller training MSE (or, equivalently, RSS).

Problem 3 (8 points)

Answer the following questions about the subset selection.

(a) (2 points) Given the following R code and output,

```
summary(regsubsets(Balance ~ Cards + Rating + Limit
   + Income + Student, Credit, method = "exhaustive"))
5 Variables  (and intercept)
Selection Algorithm: exhaustive
         Cards Rating Limit Income StudentYes
1  ( 1 ) " "   "*"    " "   " "    " "
2  ( 1 ) " "   "*"    " "   "*"    " "
3  ( 1 ) " "   "*"    " "   "*"    "*"
4  ( 1 ) "*"   " "    "*"   "*"    "*"
5  ( 1 ) "*"   "*"    "*"   "*"    "*"
```

write down all models with 4 features considered by the above code, and indicate the best one. (You can use $X_1, \ldots, X_5$ to represent the five features in order)

SOLUTION: The possible models use the following set of features: (X_1, X_2, X_3, X_4), (X_1, X_2, X_3, X_5), (X_1, X_2, X_4, X_5),

(X_1, X_3, X_4, X_5), (X_2, X_3, X_4, X_5).

The best one is:

$$\text{Balance} = \beta_0 + \beta_1 X_1 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon.$$

(Full credits are given as long as the four features are correctly identified.)

(b) (2 points) Given the following R code and output,

```
summary(regsubsets(Balance ~ Cards + Rating + Limit
    + Income + Student, Credit, method = "forward"))
5 Variables  (and intercept)
Selection Algorithm: forward
         Cards Rating Limit Income StudentYes
1  ( 1 ) " "   "*"    " "   " "    " "
2  ( 1 ) " "   "*"    " "   "*"    " "
3  ( 1 ) " "   "*"    " "   "*"    "*"
4  ( 1 ) " "   "*"    "*"   "*"    "*"
5  ( 1 ) "*"   "*"    "*"   "*"    "*"
```

write down all models with 4 features considered by the above code, and indicate the best one. (You can use $X_1, \ldots, X_5$ to represent the five features in order)

SOLUTION: (X_1, X_2, X_4, X_5), (X_2, X_3, X_4, X_5).

The best one is:

$$\text{Balance} = \beta_0 + \beta_1 X_2 + \beta_2 X_3 + \beta_3 X_4 + \beta_4 X_5 + \varepsilon.$$

(Full credits are given as long as the four features are correctly identified.)

(c) (2 points) Let's denote the model you find in (a) by M_1, and denote the model in (b) by M_2. Which model (M_1 or M_2) has smaller training MSE (or there is no sufficient information to tell)? Please briefly explain the answer.

SOLUTION: M1 has smaller training MSE. This is because for models using the same number of features, the best subset selection considers all possible models and selects the one with the smallest training MSE.

(d) (2 points) Which model would you expect to have smaller test MSE (or there is no sufficient information to tell)? Please briefly explain the answer.

SOLUTION: M1. As both models have the same number of features (i.e. the same model complexity), a smaller training MSE also indicates a smaller test MSE.

Problem 4 (8 points)

Based on the following output of R, answer the following questions.

```
> library(ISLR)
> library(boot)
> set.seed(1)
> glm.fit = glm(mpg ~  poly(horsepower, 2), data=Auto)
> cv.glm(Auto, glm.fit, K=5)$delta[1]
[1] 19.14336
```

(a) (1 point) Write down the model corresponding to line 4.

SOLUTION: With X being the horsepower,

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon.$$

(b) (1 point) Briefly explain the meaning of the number 19.14336 in the above R output.

SOLUTION: 5-fold CV error.

(c) (2 points) If we change `set.seed(1)` to `set.seed(2)` in the above R code, do you expect the same value 19.14336 as the output? Please briefly explain.

SOLUTION: No. A different seed leads to a different split of the data. Hence the result CV error would change.

Now consider the following R output.

```
> set.seed(1)
> glm.fit2 = glm(mpg ~ poly(horsepower, 2), data=Auto)
> cv.glm(Auto, glm.fit2, K = nrow(Auto))$delta[1]
[1] 19.24821
```

(d) (2 points) If we change `set.seed(1)` to `set.seed(2)` in the above R code, do you expect the same value 19.24821 as the output? Please briefly explain.

SOLUTION: Yes, as the LOOCV is not affected by randomly shuffling the data.

(e) (2 points) Which of 19.14336 and 19.24821 do you expect to be closer to the expected MSE of the model you write in part (a)?

SOLUTION: 19.24821 as the LOOCV yields more accurate estimate of the expected MSE than 5-fold CV.

Problem 5 (10 points)

(a) (3 points) Consider the ridge regression with tuning parameter λ. Draw a picture which contains three curves (squared bias, variance, test MSE) of the fitted model. Use the x-axis to indicate values of λ which start from a large number and decrease until 0. The y-axis should represent the values of squared bias, variance and test MSE of the fitted models.

In your picture, indicate which curves correspond to the squared bias, variance and test MSE, respectively. (Pay attention to the trend and relative magnitudes of the three metrics.) Also indicate the points correspond to the OLS.

(b) Consider the estimator which solves the following problem

$$\min_{\beta=(\beta_1,\beta_2)} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2, \quad \text{subject to} \quad |\beta_1| + |\beta_2| \leq s. \qquad (1)$$

(b1) (1 point) In general, can the solution of Eq. (1) give us a sparse model?

SOLUTION: It can yield a sparse model provided that s is chosen sufficiently small.

(b2) (2 points) Suppose the least squares estimate of (β_1, β_2) in this example is $(\widehat{\beta}_1^{LS}, \widehat{\beta}_2^{LS}) = (-1, 1/2)$. Suppose we choose $s = 2$ in Eq. (1). If the solution to Eq. (1) is unique for $s = 2$, does it contain zeros? Please briefly explain your reasoning.

SOLUTION: Since the LS estimate $(\widehat{\beta}_1^{LS}, \widehat{\beta}_2^{LS})$ lies in the feasible region, i.e.,

$$|\widehat{\beta}_1^{LS}| + |\widehat{\beta}_2^{LS}| \leq 2,$$

the solution to (1), by definition, is $(\widehat{\beta}_1^{LS}, \widehat{\beta}_2^{LS})$, which does not contain 0.

(c) Suppose we consider another estimator which solves the following problem

$$\min_{\beta=(\beta_1,\beta_2)} \sum_{i=1}^{n} (y_i - \beta_1 x_{i1} - \beta_2 x_{i2})^2, \quad \text{subject to} \quad \sqrt{\beta_1^2 + \beta_2^2} \leq s. \tag{2}$$

Consider the two predictors based on estimators computed from Eq. (1) and Eq. (2), respectively.

(c1) (2 points) Suppose we choose the same s in Eq. (1) and Eq. (2). In general, which predictor has smaller training MSE, the one corresponding to Eq. (1) or that corresponding to Eq. (2)? (or there is no sufficient information to tell). Please briefly explain the reason.

SOLUTION: Since the feasible region of (2) is larger than that of (1), the training MSE of (2) is no greater than that of (1).

(c2) (2 points) Assume the true model is

$$Y = \alpha_1 X_1 + \alpha_2 X_2 + \varepsilon.$$

Suppose you can choose the best s for both Eq. (1) and Eg. (2). When do we expect the predictor in Eq. (1) to have smaller test MSE (in terms of the true coefficients α_1 and α_2)? And when do we expect the predictor in Eq. (2) to have smaller test MSE?

SOLUTION: We expect the predictor in Eq. (1) to have smaller test MSE when α_1 or (and) α_2 is (are) 0. Otherwise, we expect the predictor in Eq. (2) to have smaller test MSE.

Problem 6 (10 points, 1 point 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 are true

   A Finding the clusters of data is usually a supervised learning problem.

   B Both regression and classification problems belong to supervised learning problems.

   C Linear regression is an example of parametric methods for estimating the regression function.

   D Ordinary Least Squares approach (OLS) can only be deployed under linear models.

   SOLUTION: BC

2. Assume that the model $Y = f(X) + \varepsilon$ holds, where ε is a random noise with mean 0 and independent of X, then

   A The regression function $f(x)$ minimizes the training mean squared error (MSE) at $X = x$.

   B The regression function $f(x)$ minimizes the expected mean-squared prediction error at $X = x$.

   C The expected mean-squared prediction error of $f(x)$ is $\mathrm{Var}(\varepsilon)$.

   D None of the above statements is correct.

   SOLUTION: BC

3. In which case, we usually prefer the nonparametric method rather than the parametric method (the number of features we use to fit the model is p and the sample size is n)

   A When p is large and n is small

   B When p is small and n is large

   C When p is small and n is small

   D None of the above statements is correct.

   SOLUTION: B

4. In a linear regression problem,

   A The unknown regression coefficients can be estimated by the Ordinary Least Squares (OLS) approach.

   B A small value of the residual sum of squares (RSS) means that the model is correct.

   C A large value of R^2 means that the model is correct.

   D R^2 can never be greater than 1.

   SOLUTION: AD

5. Qualitative predictors in regression

   A Can be incorporated using dummy variables.

   B Can be interpreted despite the model is nonlinear in them.

   C Can not be incorporated since qualitative predictors lead to a classification problem.

   D None of the above statements is correct.

   SOLUTION: AB

6. Which of the following statements are true

   A Forward selection starts from the model only including the intercept.

   B Best subset selection always finds the best model.

   C Backward selection compares fewer models than forward selection.

   D Backward selection might find the best model.

   SOLUTION: ABD

7. Which of the following statements are true

   A BIC usually selects a model with fewer features than AIC.

   B AIC and BIC require to specify the distribution of the observed data.

   C AIC or BIC is more applicable than cross-validation for selecting different models.

   D Cross-validation is always preferred over AIC or BIC.

   SOLUTION: AB

8. Which of the following statements are true

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

   B Lasso can possibly produce a sparse model.

   C Ridge can produce a biased estimator.

   D Ridge can have a smaller test MSE than Lasso when the true model is sparse.

   SOLUTION: BC

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

   A A larger size of the neighborhood usually yields a fitted model with smaller variance.

   B A larger size of the neighborhood usually yields a fitted model with smaller bias.

   C A larger size of the neighborhood usually yields a fitted model with smaller expected MSE.

   D Local regression method suffers from curse of dimensionality.

   SOLUTION: AD

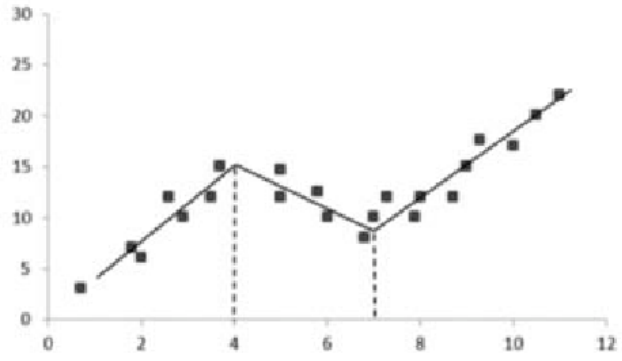
10. In the following plot, what approach did we most likely use to fit the data?

    A Polynomial regression

    B Step function approach

    C Local regression

    D Linear spline



SOLUTION: D