STA314: Statistical Methods for Machine Learning I

Midterm Exam – LEC0101

Problem 1 (11 points)

Assume that we analyze the `Carseats` data set. The goal is to predict `Sales` based on several predictors.

(a) Based on the following output of R,

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

Coefficients:
            Estimate   Std.Error  t-value  Pr(>|t|)
(Intercept) 12.173572   0.686644  17.729  < 2e-16 ***
Income       0.011070   0.004292   2.579   0.0103 *
Advertising  0.120463   0.024630   4.891 1.46e-06 ***
Price       -0.053833   0.005062 -10.635  < 2e-16 ***
USYes       -0.004623   0.342948  -0.013   0.9893
---
Residual standard error: 2.385 on 395 degrees of freedom
Multiple R-squared: 0.2938, Adjusted R-squared: 0.2866
F-statistic: 41.08 on 4 and 395 DF,  p-value: < 2.2e-16
```

answer the following questions.

(a1) (1 point) What variables other than the intercept are significant in the model (using 0.05 as the significance level)?

(a2) (2 point) Can you deduce how large is the sample size (i.e. n)? Please explain.

(a3) (2 points) Construct the 95% confidence interval for the coefficient of `Income` and also 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 $Z \sim N(0,1)$)

(b) Based on the following output of R,

```
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 ***
USYes             -0.075873   0.360800  -0.210   0.8335
Advertising:USYes  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
```

answer the following questions. The feature `US` is a factor with two levels: Yes or No.

(b1) (2 points) In this linear regression model, how do you interpret the coefficient of `USYes`?

(b2) (2 points) How do you interpret the coefficient of `Advertising:USYes`?

(b3) (2 points) Only based on the R output of part (b), can you conclude whether or not `Advertising` is significant for predicting `Sales` at 0.05 significance level? Please state your reasoning.

Problem 2 (8 points)

In a regression problem, assume that the true model is $Y = f(X_1, X_2, X_3) + \varepsilon$, where

$$f(X_1, X_2, X_3) = X_2 + X_3 + X_3^2,$$

and 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 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 $\widehat{f}_i$

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

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

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

Problem 3 (6 points)

Answer the following questions about the subset selection.

(a) (1 point) Given the following R code and output,

```
> summary(regsubsets(mpg~cylinders+displacement
                          +horsepower, Auto))
Subset selection object
Call: regsubsets.formula(mpg ~ cylinders + displacement
                               + horsepower, Auto)
3 Variables  (and intercept)
              Forced in  Forced out
cylinders        FALSE      FALSE
displacement     FALSE      FALSE
horsepower       FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: exhaustive
       cylinders displacement horsepower
1  ( 1 ) " "       "*"          " "
2  ( 1 ) "*"       " "          "*"
3  ( 1 ) "*"       "*"          "*"
```

write down the best model with 2 predictors.

(b) (1 point) Given the following R code and output,

```
> summary(regsubsets(mpg~cylinders+displacement
                         +horsepower, Auto, method="forward"))
Subset selection object
Call: regsubsets.formula(mpg ~ cylinders + displacement
                          + horsepower, Auto, method = "forward")
3 Variables  (and intercept)
              Forced in  Forced out
cylinders        FALSE      FALSE
displacement     FALSE      FALSE
horsepower       FALSE      FALSE
1 subsets of each size up to 3
Selection Algorithm: forward
      cylinders displacement horsepower
1  ( 1 ) " "       "*"           " "
2  ( 1 ) " "       "*"           "*"
3  ( 1 ) "*"       "*"           "*"
```

write down the best model with 2 predictors.

(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 RSS (or there is no sufficient information to tell)? Please briefly explain the answer.

(d) (2 points) Which model you would choose in terms of the adjusted R^2 (or there is no sufficient information to tell)? Please briefly explain the answer.
(Hint: adjusted $R^2 = 1 - \frac{RSS/(n-p-1)}{TSS/(n-1)}$, where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$ with $\bar{y}$ being the sample average of y_i.)

Problem 4 (5 points)

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

```
> library(ISLR)
> set.seed(1)
> glm.fit=glm(mpg~poly(horsepower,2),data=Auto)
> cv.glm(Auto,glm.fit,K=10)$delta[1]
[1] 19.14336
```

(a) (1 points) Please briefly explain the meaning of the number 19.14336 in the above R output.

(b) (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.

Now consider the following R output.

```
> set.seed(1)
> glm.fit2=glm(mpg~poly(horsepower,3),data=Auto)
> cv.glm(Auto,glm.fit2)$delta[1]
[1] 19.24821
```

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

Problem 5 (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 Linear regression is a classification method.

   C Regression and classification problems are both supervised learning problems.

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

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

3. In which case, we usually prefer the nonparametric method rather than the parametric method (the number of features 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.

4. Which of the following statements are true

   A The nonparametric methods may have large variance.

   B The parametric methods may have large bias.

   C The parametric methods may have large expected test MSE.

   D The nonparametric methods may overfit data.

5. Which of the following statements are true in a linear regression problem

   A Collinearity between predictors may lead to a higher variance of the prediction

   B We can look at the residual plot to check the existence of collinearity.

   C We can look at the residual plot to check the heteroscedasticity.

   D We can look at the studentized residuals to detect outliers.

6. 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 be larger than 1.

7. Qualitative predictors in regression

   A Can be incorporated using dummy variables.

   B Can not be incorporated since the model would become nonlinear.

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

   D None of the above statements is correct.

8. Which of the following statements are true

   A Forward selection starts from the model including all variables.

   B Best subset selection can be used when $p > n$.

   C Backward selection can be used when $p > n$.

   D None of the above statements is correct.

9. Which of the following statements are true

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

   B We can compare two logistic regressions using AIC or BIC.

   C We can compare two logistic regressions using adjusted R^2.

   D We can compare two logistic regressions using Mallow's C_p.

10. Which of the following statements are true

    A Lasso can have 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 non-sparse.

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