COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 2

1Directions:

- **Due: Thursday February 10, 2022, at 9pm.** Late submissions will be accepted for 24 hours after that time, with a 15% penalty. (The enforcement is strict, beginning at 9:01pm, except for extreme situations; having a poor wifi connection or minor computer problems is not sufficient for the penalty to be waived.)
- Upload the homework to Canvas as a single pdf file.
- If the graders cannot easily read your submission (writing is illegible, image is too dark, or if the contrast is too low) then you might receive a zero or only partialcredit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2Problems

Problem 1. [13 points total (6,3,4)]

Book problem Chapter 3, problem 3 "Suppose we have a data set with five predictors, $X_1 = GPA$, . . ."

Note: for interactions, use products, e.g., $X_4(i) = GPA(i) \times IQ(i)$

Y^ = 50 + 20*GPA + 0.07*IQ + 35*Level + 0.01*GPA*IQ– 10*GPA*Level
∴For High School: Y^ = 50 + 20*GPA + 0.07*IQ + 0.01*GPA*IQ …*
For College: Y^ = 85 + 10*GPA + 0.07*IQ + 0.01*GPA*IQ…**

(a)When High School>College then,
50 + 20*GPA>85 + 10*GPA
⇔GPA >3.5

∵if GPA is high enough high school graduates earn more
iii is correct

(b)college/IQ=110/GAP=4
∵ **
Y^ = 85 + 10*4 + 0.07*110 + 0.01*4*110
= 85 + 40 + 7.7 + 4.4
= $137.1k

(c)False because to verify it, we need see GPA/IQ has an impact. Also, It is possible to have many evidences although the coefficient is very small.

Problem 2. [12 points total (3 points each)]

Book problem Chapter 3, problem 4 "I collect a set of data . . ."

(a) Here the true relationship is liner, then the cubic regression would have more noise although it is hard to tell without knowing the actual training data.
∴ We would expect the RSS for the liner regression will be lower than the one for cubit regression.

(b) We do not have enough information to conclude this because the test RSS depends on the test data. However, we might be able to expect cubic regression will have a higher test RSS due to the overfit from the training because the true relationship is liner.

(c) We would expect cubic regression will have lower RSS then the liner one because the true relationship is more flexible.

(d) We do not have enough information to conclude this because we do not know how far it is from liner. Therefore, it is not clear enough to know what level of flexibility would fit better.

Problem 3. [5 points]

Suppose we have a data set with one feature X to predict another feature Y. Let n denote the number of samples. Let $\bar{X}$ and $\bar{Y}$ denote the average values of X and Y respectively in the data set

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X(i) \qquad\qquad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y(i).$$

Let θ_0^* and θ_1^* denote the coefficients for the ordinary least squares (O.L.S.) solution,

$$\{\beta_0^*, \beta_1^*\} = \underset{\{\beta_0,\beta_1\}}{\arg\min}\ \frac{1}{n}\sum_{i=1}^{n}\Big(Y(i) - \big(\beta_0 + \beta_1 X(i)\big)\Big)^2.$$

Their values are

$$\beta_0^* = \bar{Y} - \beta_1^*\bar{X} \qquad\qquad \beta_1^* = \frac{\sum_{i=1}^{n}(X(i)-\bar{X})(Y(i)-\bar{Y})}{\sum_{i=1}^{n}(X(i)-\bar{X})^2}$$

Using those formulas, calculate the O.L.S. model's prediction for $X = \bar{X}$
(i.e., the prediction $\hat{Y}$ for a new sample whose X feature has the value $\bar{X}$).

$\beta^*{}_0 = \bar{Y} - \beta^*{}_1\bar{X}$ …*

$Y = \beta^*{}_0 + \beta^*{}_1 X$
Plug in *
$= \bar{Y} - \beta^*{}_1\bar{X} + \beta^*{}_1 X$
$= \bar{Y}\ (\because X = \bar{X})$
$\Leftrightarrow \underline{\bar{Y} = Y}$

Problem 4. [16 points total (3,3,10)]

Book problem Chapter 6, problem 1 "We perform best subset …"

Notes regarding the book's pseudocode for Algorithms 6.1-6.3:

- "RSS" stands for "residual sum of squares" which is the (un-normalized) MSE,

$$\mathrm{RSS} = \sum_{i=1}^{n} \left(Y(i) - \widehat{Y}(i) \right)^2$$

  In the book's pseudo-codes, "RSS" refers to *training set* RSS.

- "Cross-validated prediction error" – you can read this as "Validation set MSE."
- "R^2" and "adjusted R^2" – you can ignore these for this homework

a) The best subset with k predictors is the model with the smallest RSS because it considers all possible models with k predictor.

b) With given information, it is not easy to answer this question. Best subset might overfits if n is smaller than p and other two might choose better on the test set.

c)

i. true

ii. true

iii. false

iv. false

v. false