COM S 474/574 Spring 2022
Introduction to Machine Learning

Homework 2

1 Directions:

- **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 partial credit.
- Any non-administrative questions must be asked in office hours or (if a brief response is sufficient) Piazza.

2 Problems

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. $X_4(i) = GPA(i) \times IQ(i)$

We are told what the model is. For any new sample, where we are told IQ, GPA, and education level, we can use this model to make a prediction.

$$\begin{aligned}
\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 X_1 + \widehat{\beta}_2 X_2 + \widehat{\beta}_3 X_3 + \widehat{\beta}_4 X_4 + \widehat{\beta}_5 X_5 \\
\widehat{Y} &= \widehat{\beta}_0 + \widehat{\beta}_1 GPA + \widehat{\beta}_2 IQ + \widehat{\beta}_3 Level + \widehat{\beta}_4 (GPA \cdot IQ) + \widehat{\beta}_5 (GPA \cdot Level) \\
\widehat{Y} &= 50 + 20GPA + 0.07IQ + 35Level + 0.01(GPA \cdot IQ) - 10(GPA \cdot Level)
\end{aligned}$$

(a). For this first question, we want to compare the predicted salary for a college student versus a high school student with the same IQ and GPA.
For a college student, $Level = 1$, resulting in a prediction of

$$\widehat{Y}^{\text{college}} = 50 + 20GPA + 0.07IQ + 35 + 0.01(GPA \cdot IQ) - 10GPA.$$

For a high school student, $Level = 0$, resulting in a prediction of

$$\widehat{Y}^{\text{highschool}} = 50 + 20GPA + 0.07IQ + 0 + 0.01(GPA \cdot IQ) - 0.$$

Taking the difference,

$$\begin{aligned}\widehat{Y}^{\text{college}} - \widehat{Y}^{\text{highschool}} &= (50 + 20GPA + 0.07IQ + 35 + 0.01(GPA \cdot IQ) - 10GPA) \\ &\quad - (50 + 20GPA + 0.07IQ + 0 + 0.01(GPA \cdot IQ) - 0) \\ &= 35 - 10GPA\end{aligned}$$

So if $GPA < 3.5$, $\widehat{Y}^{\text{college}} - \widehat{Y}^{\text{highschool}} > 0$ which means $\widehat{Y}^{\text{college}} > \widehat{Y}^{\text{highschool}}$. For higher GPA, the relation is flipped. Thus, iii is correct. (strictly speaking, that is a property of the fitted model, which is based on the data set and model class; with more data, we may get a different model)

(b). For this, we just plug in

$$\begin{aligned}\widehat{Y} &= 50 + 20GPA + 0.07IQ + 35Level + 0.01(GPA \cdot IQ) - 10(GPA \cdot Level) \\ &= 50 + 20(4.0) + 0.07(110) + 35(1) + 0.01(4.0 \cdot 110) - 10(4.0 \cdot 1) \\ &= 137.1 \qquad (1)\end{aligned}$$

(c). From the previous question (part b), we saw that the values of the features are quite different, eg level is just between 0 and 1, but the value of the feature X_4, interaction of GPA and IQ, is in the hundreds. For part (b), $\beta_1 X_1 = 20 * 4.0 = 80$ and $\beta_4 X_4 = (0.01) * (4 * 110) = 4.4$.
The products of the coefficients and the features, which is what contributes to the prediction, end up being on similar scales than the coefficients alone suggest. Thus, we can see from that case that the overall contribution of that term is non-negligible, and the coefficient scale seems to compensate for the scale of the raw features.
[Talking about using statistical assessments of significance, instead of just looking at coefficient magnitude, is also fine]

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

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

For this problem, it is helpful to review Figures 2.9-2.12 in the book, which ran experiments of fitting linear models and more complex models for different underlying functions, and also plotted results on training set fit and test set fit.

$n = 100$ is more samples than we used in some of the Jupyter notebook plots and likely more than in Figs. 2.9 and 2.10 of the book. When we have a lot of data (and relatively few features), overfitting is less of a concern.

(a). We would expect them to be about the same—the cubic function will necessarily do better, but the improvement would likely be small.

(b). Unless the noise is quite large, we would expect them to be about the same. overfitting is less likely to occur here (with 100 samples) than other examples we've seen (with 20-30 samples). It is fine to answer that the linear function will have slightly lower test RSS.

(c). The cubic function will necessarily do better on fitting the data set. The more non-linear the function is, the bigger of a gap in performance between the linear and cubic function.

(d). This is tougher to answer without knowing more, including how big the noise is relative to the underlying trend's values and how non-linear the underlying trend might be.

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

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

Let β_0^* and β_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

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

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

The model is a line, and for any x value the prediction is

$$\widehat{y} = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

We want to show that for a specific x value, namely the average of the data set samples, $x = \bar{x}$, that the prediction has a value of $\widehat{y} = \bar{y}$, the average y value of the data set samples. Using the formula (3.4), the prediction for $x = \bar{x}$ is

$$\begin{aligned}\widehat{y} &= \widehat{\beta}_0 + \widehat{\beta}_1\bar{x}\\ &= (\bar{y} - \widehat{\beta}_1\bar{x}) + \widehat{\beta}_1\bar{x}\\ &= \bar{y}\end{aligned}$$

so indeed, the fitted line does pass through $(\bar{x}, \bar{y})$.

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,

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

  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). Best subset – it will try all possible models with k predictors, including those checked by forward stepwise and backward stepwise. It then selects a model for level k based on training RSS.

(b). Best subset – it will try all possible models with k predictors, including those checked by forward stepwise and backward stepwise. It then selects a model for level k based on training RSS, but among the k models, they have the sample complexity $k+1$ coefficients, and generally among models with the same complexity, the models with smaller training error should have smaller test error. The forward and backward searches will only check a few of the $\binom{p}{k}$ models, constrained by their earlier decisions.

(c). The statements are all similar, but slightly different. Read these carefully.

i. True – for forward stepwise, all models with $k+1$ are the chosen set with k features plus an additional one (the addition varies).

ii. True – for level k, the backward search takes the $k+1$ subset and tries removing one of the features at a time.

iii. False – in general there is no guarantee of sets checked by one heuristic being related to those of another heuristic.

iv. False – in general there is no guarantee of sets checked by one heuristic being related to those of another heuristic.

v. False – In general, there is no guarantee that the best subset with k variables will be related to the best subset with $k+1$.