IEOR242 HW2 solution Fall 2021

October 15, 2021

Please self-grade problems 1 and 2 and attach the rubric form in your HW3 Submission

Problem 1 (20 points)

a)

$$P(Y=1 \mid X=30, Z=1) = \frac{1}{1+\exp(-(-3.50+0.18*30+1.24*1))} \approx 0.9585$$

The predicted probability of getting a good grade for a team that spends 30 hrs working on the project using Python is 0.9585

b) (a) Notice that we are training two different sub-models: If the team is using Python, then the probability to get a good grade follows

$$P(Y=1 \mid X, Z) = \frac{1}{1+\exp(-(\alpha_0+\alpha_1 X))} \tag{1}$$

On the other hand, if the team uses R, the probability of getting a good grade follows

$$P(Y=1 \mid X, Z) = \frac{1}{1+\exp(-(\beta_0+\beta_1 X))} \tag{2}$$

(b) First, we split the training data into two groups: group 0 uses R and group 1 uses Python.

Second, for group 1, we fit equation (1) and for group 0, we fit equation (2).

Then, we obtain the desired model:

$$P(Y=1 \mid X, Z) = Z\frac{1}{1+\exp(-(\alpha_0+\alpha_1 X))} + (1-Z)\frac{1}{1+\exp(-(\beta_0+\beta_1 X))}$$

(c)

$$P(Y=1 \mid X=30, Z=1) = 1 + \frac{1}{1+\exp(-4.02+0.24*30)} + 0 \sim 0.9601$$

The predicted probability is 0.9601

(d) Notice that there is a significant difference between the two groups: 60% Python getting good grade and 52.25% R getting good grade. The model from b) will have more discriminitive power.

(e) Since we have smaller number of training set (for both Python and R), we may overfit (i.e., biased w.r.t. small dataset) when we train the model separately. So it is more suitable to put all data into one regression model.

Problem 2 (10 points)

Expected loss when our prediction is 0:

$$\mathbb{E}[L(Y = 0 \mid X = x)] = 0 \times P(Y = 0 \mid X = x) + L_{FN}P(Y = 1 \mid X = x) = pL_{FN}$$

Expected loss when our prediction is 1:

$$\mathbb{E}[L(Y = 1 \mid X = x)] = 0 \times P(Y = 1 \mid X = x) + L_{FP}P(Y = 0 \mid X = x) = (1 - p)L_{FP}$$

Since the classifier should give prediction 1 if the expected cost of it is less than the expected loss of giving prediction 0; and give the prediction 0 if vice versa, i.e.,

$$h^*(x) = \begin{cases} 1 & \text{if } (1-p)L_{FP} \leq pL_{FN} \\ 0 & \text{if } (1-p)L_{FP} > pL_{FN} \end{cases}$$

Solving $(1 - p)L_{FP} = pL_{FN}$, we have the threshold $\bar{p} = \frac{L_{FP}}{L_{FP}+L_{FN}}$

Problem 3(70 points)

Problem 3-a(40 points)

part-(i)(7 points)

The logistic regression would be like:

$$log(Odds) = \hat{\beta}_0 + \sum_{i=1}^{p} \hat{\beta}_i x_i$$

Following is the table of coefficients.

Variables	Coefficient
intercept	-9.2740
education[T.High school/GED] (binary)	-0.1053
education[T.Some college/vocational school] (binary)	-0.1025
education[T.Some high school](binary)	0.0610
male(binary)	0.5621
age(continuous)	0.0689
currentSmoker (binary)	0.1539
cigsPerDay (integer)	0.0155
BPMeds (binary)	0.1528
prevalentStroke (binary)	0.8209
prevalentHyp (binary)	0.2075
diabetes (binary)	-0.2975
totChol (continuous)	0.0020
sysBP(continuous)	0.0181
diaBP(continuous)	-0.0045
BMI(continuous)	0.0136
heartRate(continuous)	-0.0046
glucose(continuous)	0.0096

Table 1: Coefficients

Grading Rubrics

- (-3) Missing coefficient.
- (-4) Didn't present results in a clear manner.

part-(ii)(7 points)

According to p-value, independent variable(p-value) *male(0.000), age(0,000), sysBP(0.000), cigsperDay(0.038), glucose(0.001)* might be considered important. Here, we present age as an example. "Keeping everything else constant, an extra year in age increases the odds of developing CHD in the next 10 years by a (multiplicative) factor of exp(0.0689) ≈1.07133"

Grading Rubrics

- (-4) Didn't consider $p - value$ while choosing important variables list. Also, at least one of variables of the above mentioned should be included.
- (-3) Wrong explanation for the variable's odd. (-1 for incorrect odds value and -2 for incorrect explanation.)

part-(iii)(6 points)

Our threshold value $\bar{p}$ should satisfy the following equation (Expected loss when we prescribe

medication = Expected loss when we don't prescribe medication.)

$$775000 \times 0.15\bar{p} + 75000 \times (1 - 0.15\bar{p}) = 700000 \times \bar{p} + 0 \times (1 - p) \Rightarrow \bar{p} = \frac{15}{119} \approx 0.126$$

Grading Rubrics

- (-4) Wrong equation.
- (-2) Wrong $\bar{p}$ value.

part-(iv)(5 points)

Our confusion matrix looks like this:

	Predict as 0	Predict as 1
Real response is 0	569	354
Real response is 1	56	119

Table 2: Confusion Matrix

With this data, we can calculate accuracy, TPR, and FPR

$$accuracy = \frac{569 + 119}{569 + 119 + 56 + 354} = \frac{688}{1098} \approx 0.6266$$
$$TPR = \frac{119}{56 + 119} = \frac{119}{175} = 0.68$$
$$FPR = \frac{354}{569 + 354} \approx 0.3835$$

So we can say that our model can correctly predict 63% of the patients. 68% of group of patients who would have CHD in ten years, are correctly predicted as the patient would contract CHD in ten years. Also 38% of group of patients who would not have CHD in ten years, are falsely predicted as Ten Year CHD.

Grading Rubrics

- (-1) For each wrong value of ACC, TPR, FPR.
- (-2) Didn't present non-technical explanation. (Avoid using 'positive' or 'negative' term.)

part-(v)(5 points)

We can use the confusion matrix that we have created in the previous question. Let's first consider the first case where we are calculating expected cost per patient under assumption that CHD outcomes in the test set are not affected by the treatment decision. The expected cost per patient in this case would be,

$$700000 \times 56(\#\text{ of FN}) + 75000 \times 354(\#\text{ of FP}) + 775000 \times 119(\#\text{ of TP}) = 157,975,000$$
$$157,975,000/1098 \approx 143875.227 \text{ (per patient)}$$

Now let's assume that the treatment decision impacts a patient's risk of developing CHD. So 80% of TP cases would move to FP cases. The expected cost per patient would be

$700000 \times 56(\#\text{ of FN}) + 75000 \times 455.15(\#\text{ of FP} + 0.85^*\#\text{ TP}) + 775000 \times 17.85(0.15^*\#\text{ of TP}) = 87170000$
$87170000/1098 \approx 79389.799$ (per patient)

Grading Rubrics

- (-2) Wrong Calculation for the first case
- (-3) Wrong calculation for the second case (Some rounding errors are allowed)

part-(vi)(5 points)

Since majority of response is '0', our baseline model would always predict as 0.

	Predict as 0	Predict as 1
Real response is 0	923	0
Real response is 1	175	0

Table 3: Baseline model Confusion Matrix

TPR and FPR are both equal to 0 and accuracy would be $\frac{923}{1098} \approx 0.84062$ Expected cost per patient would be $175 \times 700000/1098 = 111566.4845$

Grading Rubrics

- (-1) For each wrong TPR, FPR and Accuracy
- (-2) Wrong expected cost per patient

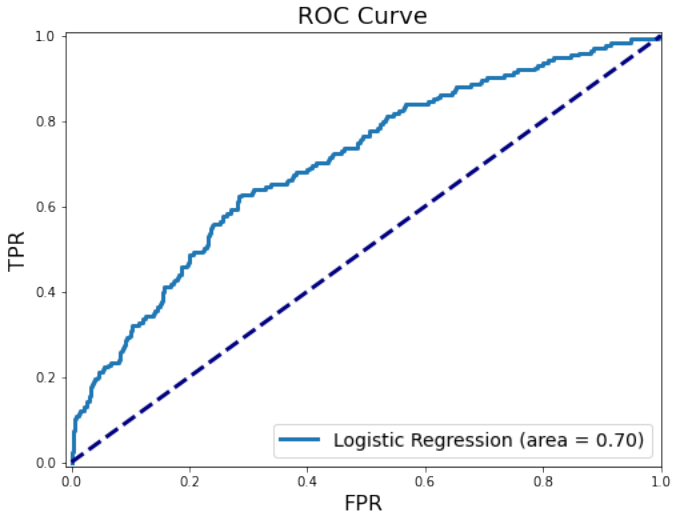
part-(vii)(5 points)

We input our new observation data to our fitted logistic regression model. Then the probability of predicting as TenYearCHD = 1 with the given data is 0.137213. And since $0.137213 > 15/119$, we should prescribe the preventive medication.

Grading Rubrics

- (-3) Wrong Probability.
- (-2) Wrong decision

Problem 3-b(15 points)



Here, the AUC is 0.70. The ROC curve is helpful because it can help us understand the trade off between true positive rate and false positives rate, and help choose the best threshold value. We have different costs associated with these true positive samples and false positive samples, and we want to select the point on the curve that can balance these two costs. For example, here we have 75000 loss for false positive and 775000 costs for true positive. Therefore we want a cut-off value where the true positive rate is high and false positive rate is low, which means the left upper area of the plot.

One interesting observation is that the ROC curve is above the baseline curve so it performs better than the baseline model. Our breakeven point, TPR of 0.68 and a FPR of 0.38, is far away from the naive baseline, which is a sign that this model have good discriminative ability

Grading Rubrics

- (-4) Didn't present ROC curve.
- (-4) Wrong AUC.
- (-4) No explanation of usefulness of ROC curve.
- (-3) Didn't present at least one interesting observation.

Problem 3-c(10 points)

Now the cost of not receiving medication but getting CHD is equal to 500k and the cost of receiving medication but getting CHD is equal to 500k $+C$ and the cost of receiving medication but not getting CHD is equal to C. So we should have C such that expected loss is equal in both cases (receiving medication or not) under probability $\bar{p}$. So this would be,

- Loss from meds: $(C + 500000) \times 0.15\bar{p} + C(1 - 0.15\bar{p})$
- Loss from no mdes: $500000\bar{p}$

$$0.15\bar{p} \times (C + 500000) + (1 - 0.15\bar{p}) \times C = 500000 \times \bar{p}$$
$$\iff C = 425000 \times \bar{p} \approx 53571.4286$$

Grading Rubrics

- (-1) Wrong cost for each case.(Total three cases' cost have been changed)
- (-4) Wrong equation for getting C value.
- (-3) Wrong C value.

Problem 3-d(5 points)

There is no fixed answer for this question. If you present reasonable answer, you will get full credit.

One possible answer could be: We chose $\bar{p}$ by comparing expected loss. This could be interpreted as we only treat patients who could be beneficial. Also, this does not reflect patient's willingness for the treatment. To remedy this, we could deliberately lower the value of $\bar{p}$ so that we could cover patients who have low probability of CHD though it would be costly.

Grading Rubrics

- (+2) Identifying at least one ethical concern.
- (+3) Proposing at least one way of addressing this concern.