MATH 4322 - Introduction to Data Science & Machine Learning

Exam 1 - A Solution
Fall 2019

September 29, 2019

Part 1(40 points)

Problem 1. (4 points) The response variable is $Accept$ which is a **quantitative** variable. Thus this is a **regression** task with **linear** regression.

Problem 2. (4 points) The predictor that is qualitative is $Private$, we would use a **dummy variable** to represent it.

Problem 3a. (6 points) The p-value $< 2.2e^{-16} \approx 0$.
The percentage of variation is the $Multiple\ R - squared$ 74.5%.
$H_0 : \beta_{PrivateYes} = \beta_{F.Undergrad} = \beta_{Grad.Rate} = \beta_{Top25perc} = 0$. (Can use numbers instead).

Problem 3b. (4 points)

$$Ac\hat{c}ept = \begin{cases} -2233 + 1269 + 0.6875 \times F.Undergrad + 14.78 \times Grad.Rate + 4.416 \times Top25perc, & \text{if Private University} \\ -2233 + 0.6875 \times F.Undergrad + 14.78 \times Grad.Rate + 4.416 \times Top25perc, & \text{if Public University} \end{cases}$$

Problem 3c. (4 points) Significant predictors are $Private$ and $F.Undergrad$.

Problem 3d. (4 points) $H_0 : \beta_{F.Undergrad} = 0$, given $\beta_{F.Undergrad}$ and $\beta_{Private}$ and $\beta_{Top25perc}$ are in the model.

Problem 3e. (8 points) $\hat{\beta}_{Private} = 1269$ that implies that the number of accepted applications will increase on average by 1,269 if the university is Private.
$\hat{\beta}_{F.Undergrad} = 0.6875$ For each additional full time undergrad, the number of accepted applications will increase on average by 0.6875.

Problem 3f. (6 points) **Confidence Interval**: The universities that have a 65% graduation rate and 15,000 full time undergrads will accept on average between 8,433 and 10,343 applications.
Prediction Interval: For one university that has a 65% graduation rate and 15,000 full time undergrads is predicted to accept between 5,634 and 13,142 applications.
The **prediction interval** is wider.

Part 2 (30 Points)

Problem 1. (9 points) The response is $student(yes/no)$ thus this is a **classification** task. We will use **logistic regression**.

Problem 2. (8 points)

$$\left(\log \frac{\hat{p(x)}}{1-\hat{p(x)}}\right) = -11.65 - 0.6095 \times student + 0.00525 \times balance + 0.0000535 \times income$$

or

$$\hat{p(x)} = \frac{exp(-11.65 - 0.6095 \times student + 0.00525 \times balance + 0.0000535 \times income)}{1 + exp(-11.65 - 0.6095 \times student + 0.00525 \times balance + 0.0000535 \times income)}$$

Where

$$student = \begin{cases} 0, & \text{if no} \\ 1, & \text{if yes} \end{cases}$$

Problem 3. (4 points) $balance$ is significant, (if they also say $income$ that is fine).

Problem 4. (5 points) As the $balance$ increases, the chance of a default increases.
As the $income$ increases, the chance of a default increases. (Again this is only needed if they say that income is significant).

Problem 5. (4 points) The **chance** or **probability** of a person defaulting on the credit card is at 1.2% for a customer that has a balance of \$1,000.

Part 3 (30 Points + 10 Bonus)

Problem 1. (20 points)

- Linear relationship
- The expected values of the errors (residuals) are zero.
- Constant variance of the errors (residuals).
- Error have a Normal distribution for each value of x.

Problem 2. (Only 2 are necessary for points)(2 bonus points)

- Scatterplot or "pairs" to see linearity
- Residual plot to see constant variance, linearity and outliers
- Q-Q plot to see Normality

Problem 3a. (2 points) This is not a linear relationship.

Problem 3b. (8 points) This would be a polynomial regression. **Cubic**

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Problem 3c. (4 bonus points) If we have too many degrees for a polynomial regression, we will have a problem of **overfitting** the training data. Then the results will not work best for any other test data that we would use from the same population.

Problem 4. (Only need 2 for points)(4 bonus points)

- C_p - small
- AIC - small
- BIC - small
- adjusted R^2 - large (it has to say adjusted).