CPSC 392– Introduction to Data Science

Spring 2020

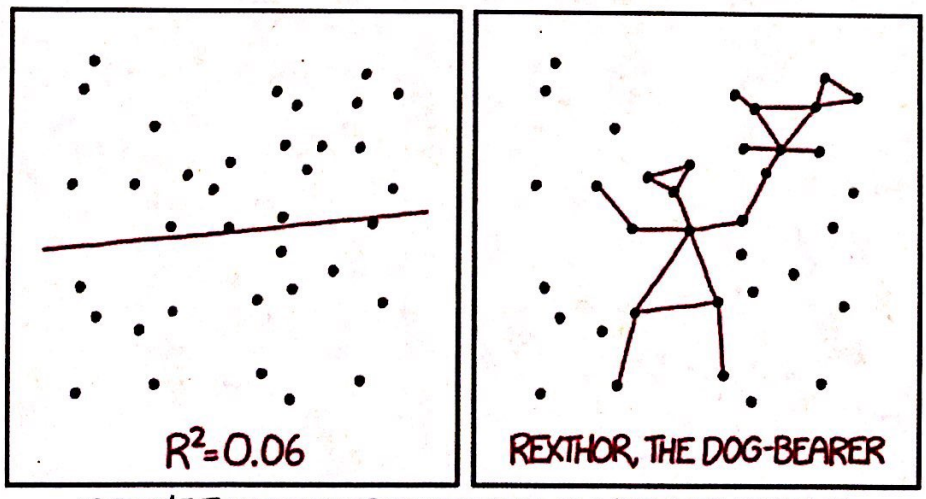
Exam I

This examination is closed book and notes. There are 7 problems and 55 points possible. You have 75 minutes to earn as many points as you can. Good luck!

In the questions below, whenever an explanation is required ("why?"), full credit will not be given if the explanation is not provided.

Name :

Toby Chappell



Academic Integrity Agreement

*failure to sign will result in a 0 for this exam

I certify that I have read and understand Chapman University's policy on academic integrity (https://www.chapman.edu/academics/academic-integrity/_files/academic-integrity-policy.pdf).

In addition to the examples listed in the policy document, I am aware that the following actions also constitute an academic integrity violation:

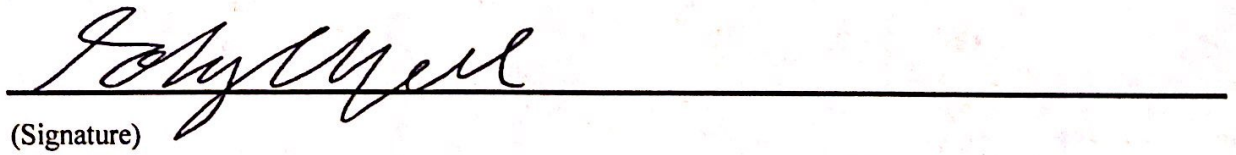
- Copying source code from another individual or the Internet without attribution
- Modifying someone else's code, without attribution, with the intention of claiming it as one's own work
- Referencing solutions to exams or assignments from previous course offerings that have not been made publicly available by the instructor

Furthermore, I understand that any instances of academic misconduct (regardless of circumstances or severity) in computer science or software engineering courses will result in a report to the university Academic Integrity Council with the recommended sanction being one of the following:

a) A grade of "F" in the course
b) A one semester suspension
c) Expulsion

Toby Chappell

(Print Name)



(Signature)

3/16/20

(Date)

1. You are given a dataset consisting of m attributes { $A_1, A_2, A_3, \ldots, A_m$ } as well as a class label, C, for n data points. Your job is to use supervised learning to build a model to predict C given the attributes. After talking to a domain expert, you come to the conclusion that C does not depend on { $A_1, A_2, A_3, \ldots, A_m$ } and in fact, the attributes are useless. In this case, is there anything that can be done to accurately predict C? Support your argument either way. (10)

Since C does not depend on any attributes given, there is no way to accurately predict C using the attributes. In order to build an accurate model, there must be a correlation between the independent variables and dependent variable. Since none exists, a model cannot be created to predict C using any of the attributes. However, it is still possible to predict C without using the attributes. To do this, the model can simply use the average of C to predict C. While this may not be a perfect solution, the model can accurately predict C if C has a low standard deviation (or low spread). If this is not the case, no model can be generated to accurately predict C.

2. What is the sigmoid function and why is it used for logistic regression? List the interesting properties of the function. (5)

Sigmoid Function: $\sigma = \frac{1}{1+e^{-x}}$

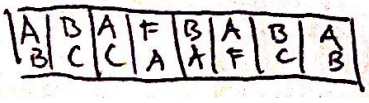
where $x = \theta^T \bar{x}$ for logistic regression.

The sigmoid function is used for logistic regression since it can map values to a number between 0 and 1. Since logistic regression outputs probabilities, this attribute is essential (it is not possible to have a negative probability or a probability over 100%). Another property that is useful with the sigmoid function is that its derrivative is easy to calculate. Since gradient descent/ascent requires for the slope at a point to be calculated (in order to know which direction will minimize/maximize the function), this feature is important and makes calculations less complex and more efficient.

3. What is k-fold cross validation? In what ways is it better than a simple cross validation of dividing the data into one pair of training and validation sets? You may sketch a diagram to support your answer. (10)

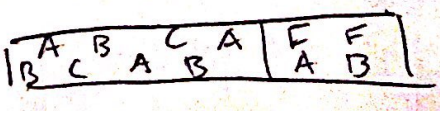
K-fold cross validation randomly divides the data into k bins. It will then take a single bin as a testing set and use the rest as a training set. Next, it will fit the model with the training set, evaluate based on the testing set, and record the score. This process is repeated k times so that each bin is used as a testing set once and only once. When k-fold cross validation is feasible, it is better than a simple cross validation for a few reasons. For one, the data is tested k times providing more accurate information about the model. In addition, if a portion of data exists at a lower frequency (i.e. on this mid few students will recieve an F while a majority will recieve an A, B, or C), it is more likely that a portion of that data will exist in different bins meaning the data will be trained and tested on. With a simple cross validation, it is more likely the data will be in the training set and not the testing set or vice versa.

k-fold



vs.

simple



4. Why is it important to take the log of the Likelihood function when applying gradient descent for Logistic Regression? Support your answer by converting the Likelihood function for Θ into the Log Likelihood function. (10)

It is important to take the log of the likelihood function since it allows us to compute the derrivative far easier. This is required for calculating gradient ascent since it needs the slope to determine which direction and by how much to maximize probability.

$$L(\theta) = \prod_{i=1}^{m} h(x_i)^{y_i} (1-h(x_i))^{(1-y_i)}$$

(Note: this is hard to maximize since it deals with a lot of probabilities being multiplied)

$$L(\theta) = (h(x_1)^{y_1}(1-h(x_1))^{(1-y_1)}) \cdot \ldots \cdot (h(x_m)^{y_m}(1-h(x_m))^{(1-y_m)})$$

$$\log(L(\theta)) = \mathcal{L}(\theta) = \log(h(x_1)^{y_1}(1-h(x_1))^{(1-y_1)}) + \log(h(x_2)^{y_2}(1-h(x_2))^{(1-y_2)}) + \ldots + \log(h(x_m)^{y_m}(1-h(x_m))^{(1-y_m)})$$

$$\therefore \mathcal{L}(\theta) = \sum_{i=1}^{m} \log(h(x_i)^{y_i}(1-h(x_i))^{(1-y_i)})$$

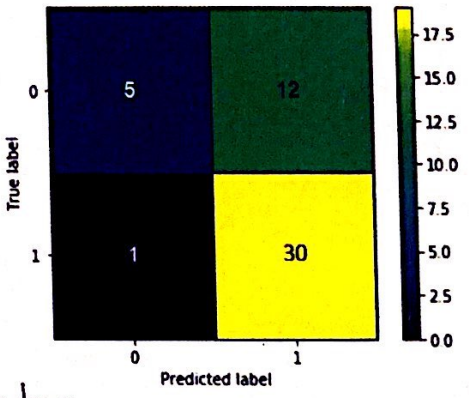
$$\mathcal{L}(\theta) = \sum_{i=1}^{m} (\log(h(x_i)^{y_i}) + \log((1-h(x_i))^{(1-y_i)}))$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} (y_i \log(h(x_i)) + (1-y_i)\log(1-h(x_i)))$$

$$\mathcal{L}(\theta) = \sum_{i=1}^{m} y_i \log(\sigma(\theta^T \bar{x})) + (1-y_i)\log(1-\sigma(\theta^T \bar{x}))$$

$\mathcal{L}(\theta)$ is far easier to maximize since it is a summation

5. Below is a confusion matrix generated from fitting a logistic regression model on a data set of students opting to take a class as pass/no pass (pass:1,no pass:0). Interpret the matrix, identify the True Positive, True Negative, False Positive, and False Negative counts. Calculate the Recall and Accuracy. If the model accuracy isn't good, discuss why the model could still be useful. (10)



The data indicates that 31 students opted for pass while 17 opted no pass. The model correctly predicted 30 who opted pass and 5 who opted no pass. However, the model predicted pass for 12 who actually opted no pass. In addition, it predicted no pass for 1 wh opted pass.

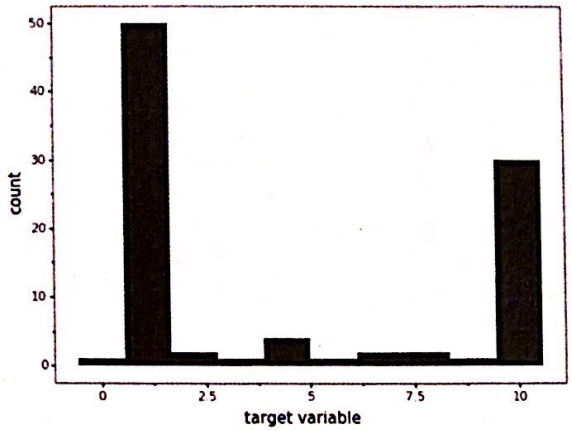
True Positives: 30
True Negatives: 5
False Positives: 12
False Negatives: 1

$$\text{Recall} = \frac{TP}{TP+FN} = \frac{30}{30+1} = 0.9677$$

$$\text{Accuracy} = \frac{TP+TN}{\text{Total}} = \frac{30+5}{30+5+1+12} = 0.7292$$

If the accuracy is not good enough for whatever application, the model can still be used to accurately predict if a student opted pass. However, an accuracy of 0.7292 is fairly good and as such this should not be an issue.

6. You are given a data set which includes a continuous target variable for which the histogram is given below. Is it possible to repurpose this attribute to fit a Logistic Regression model? Support your argument either way. (5)



Yes, the data can be seperated into categories making the data discrete. For instance, it can be seperated into "low" (values less than 5) and "high" (values greater than or equal to 5). We can then use this new column to predict if a value is "low" or "high" using logistic regression. Moreover, this model should be fairly accurate since most data lies towards the extremas of the data (close to 0 or 10 but not in the middle).

7. (NO PENALTY FOR CONSTRUCTIVE CRITICISM) (5)

What is your impression of the course so far? Is there anything I can do to help improve your learning? Given the transition to online classes, in what ways do you think you will be able to get most from the class?

I've really enjoyed the couse so far. The content is interesting and has many practical applications. One thing I'd think would be useful is to assign with homework on theory or problems we do not have the ability to program.