Assignment 01

Instructions

1. Each assignment can contain both theoretical and practical questions.
2. Use LaTeX (preferred) or Word for theoretical question responses.
3. Practical questions are in the provided Jupyter notebook. Use Google Colab (Preferred) or Jupyter Notebook to complete questions directly in the Jupyter Notebook. Include code changes and reasoning in the Jupyter Notebook. Convert the Jupyter Notebook into an HTML page for submission.
4. Submit a PDF or Word file with reponses to theoretical questions, a Jupyter Notebook, and an HTML page (both files) with completed practical questions.
5. A 25% penalty applies to submissions on the first day after the due date, and a 50% penalty for submissions 24 to 48 hours late. No submissions will be accepted beyond 48 hours past the due date.

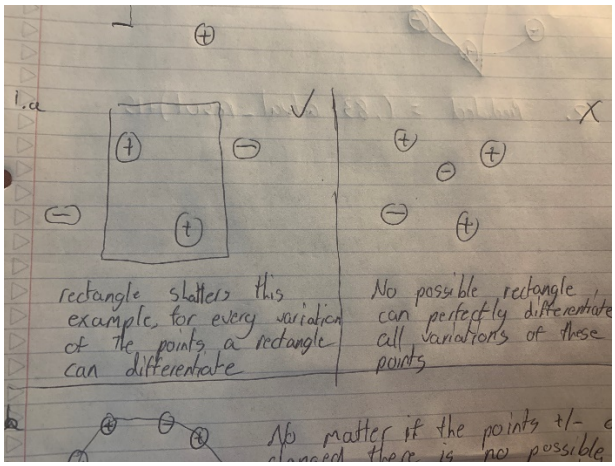
Theoretical Questions

Question 1

Considering the Vapnik-Chervonenkis (VC) Dimension, address the following three questions:

a. Explain the concept of VC Dimension with an illustrative example.

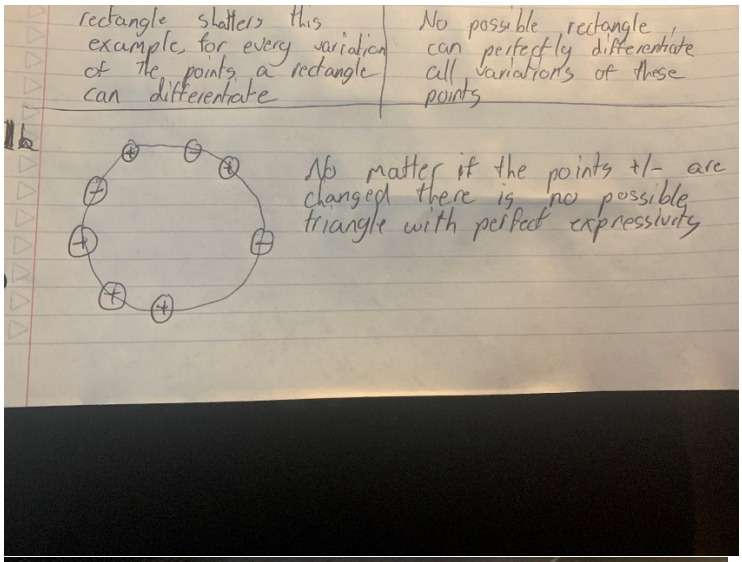
The concept of the VC dimension allows for a deeper understanding of the capabilities and limitations of a given hypothesis shape. A hypothesis can shatter a dimension if there are a set of points that are able to be perfectly separately every variation of points. The VC dimension visually illustrates the expressivity of a model.

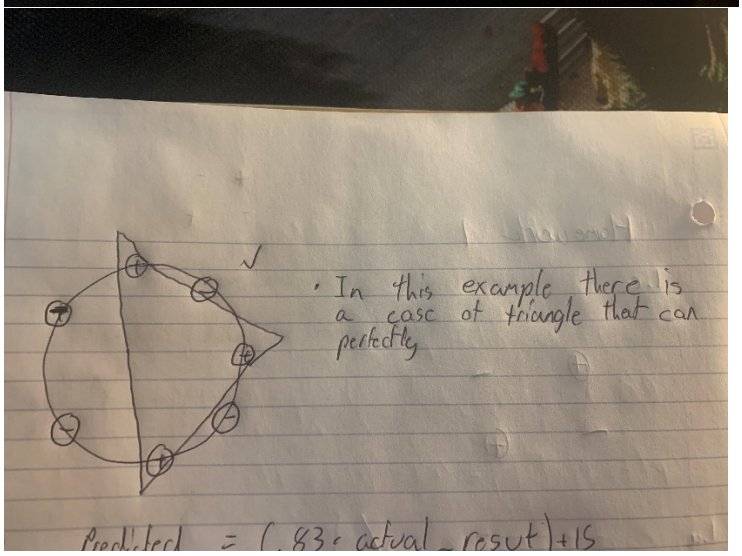


Note: I am not sure if the illustrative example implied a hand drawn example or an image from the internet so I did a hand drawn example.

b. Imagine eight points positioned equidistantly on a circular rim. Can the VC Dimension for a triangle be 8 in 2D space? Justify your answer with an example. If the answer is No, specify the correct VC Dimension for a triangle in 2D space.

The shape of a triangle does not have expressivity to perfectly differentiate between 8 points. Th correct dimension for a triangle is 7 points.





c. For any set of N points, a learning algorithm H can perfectly represent all possible ways of dividing these points into two classes for all values of N less than or equal to 2^d, where d is the VC dimension of algorithm H. Is this statement accurate? Substantiate your answer with clear reasoning.

This statement is incorrect, It is true that for a set of d points there are 2^d different ways into dividing a set of points is because each point can be split into either positive or negative points. A hypothesis H that can shatter this model is able to perfectly differentiate any value of d and below. However, this hypothesis is not able to differentiate n = 2^d it can only differentiate n <= d

Question 2

Consider the dataset below containing information about N students, including their Assignment and Exam marks along with corresponding results. The result is the target variable in the dataset. Assume you are running a regression model to predict the result, and the prediction formula is given as: predicted_result = (0.83 * actual_result) + 15.

Calculate the predicted value for all the students using this formula and subsequently compute the error using the given formula:

$$\square\square\square\square\square = \square\_1 \sum_{\square=1\square} [\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square\square_{\square} - \square\square\square\square\square\square\square\square\square\square\square\square\square\square\square_{\square}]_2$$

Based on the calculated error, assess the performance of our regression model. Do you consider the model's performance to be good or bad? If you were to encounter the same level of error in general for other datasets, would it indicate the model's ability to predict data correctly across

Result	Predicted_Result	Error
45	52.35	54.0225
65	68.95	15.6025
71	73.93	8.5849
40	48.2	67.24
76	78.08	4.3264
81	82.23	1.5129
69	72.27	10.6929
89	88.87	0.0169
59	63.97	24.7009
		AVG Error
		20.74443

Given these error rates, it is difficult to evaluate the performance of this regression model. When evaluating a model, it is important to consider the impact of what the model is doing. In this case if we take the root mean square of the data, we would see that our model would on average deviate from the actual result around 4.55. This error rate can have a vastly different effect depending on what the model is used for. For example, if the model was being used to predict the angle in which a rocket ship needs to take to reach the moon, it would be considered very poor as 4.5 degrees in either direction can vastly change the end destination. In the case that these are used to roughly predict performance of student given some limited information it is ok. It should also be noted that are model is dependent on having the actual result therefore we should expect the model to minimize the error.

Assignment Marks	Exam Marks	Result
1	2	45
2	6	65
3	3	71
3	1	40
4	3	76
4	4	81
7	1	69
5	4	89
6	2	59

Figure 1: Dataset Image

various cases? Justify your response. (Note: There is no absolute right or wrong answer; provide a reasoned explanation for your stance.)

Question 3

a. Explain the rationale behind the practice of dividing a dataset into Training, Validation, and Test sets. Specifically, elaborate on the advantages of incorporating a Validation set into our dataset for machine learning tasks.

   Standard practice in which data sets are split into training, validation, and test sets are helpful in evaluating a model's general ability and if it is overfitting. Training is obviously important to create a model, the validation set is important to be tune the parameters. And the test set is important to evaluate the performance of the final model.

b. Evaluate the decision to use the Training data as both the Validation and Test data. Provide a well-justified response, discussing the potential implications and drawbacks associated with such a choice.

   It is often not good practice to use the training data as validation and test data. This will artificially boost the accuracy of a model as the performance will not be as applicable to real world situations. This could also lead to a model being overfit in which it's too complex and doesn't generalize well. This is the result of a model memorizing a set of data and not the patterns causing it, i.e training to the noise of the data

Practical Questions

Please refer to and answer Question 4 and Question 5 in the provided Jupyter Notebook