HW1

LJHuang

january 24

Question1:

a. The VC dimension of a model is a measure of how complex function it can learn without making mistakes. Imagine you have a two dimensional space and you are using straight lines as your classifier. Start from three points if you make one point negative and the other two positive, a straight line could seperate three points based on their type. However, if you arrange four points in a square. Label the diagonal points the same and the other diagonal differently. No single straight line can seperate these points correctly according to this labeling. Thus, the VC dimension of linear classifiers in a 2D space is 3. This is because they can shatter any set of three points but cannot shatter every possible arrangement of four points.

b. A triangle can at most touch 3 points on its vertices. If the points are queidistant on a circle, then a triangle can only include points that are adjacent to each other. For 8 equidistant points on a circle, it is impossible for a triagle to include a subset of points that are not all adjacent. For instance, you cannot include points 1, 3, and 5, and exclude points 2, 4 and 6. Because the non-adjacent points cannot all be inside the triangle while keeping the intermediate points outside. The correct VC Dimensiion for a triangle in 2D space is actually 7. This is because a triangle can shatter any set of 7 points in general position.

c. This statement suggests that the number of ways to divide N points into two classes by H is limited by 2^d. However, the correct interpretation is that the VC dimension d is the size of the largest set of points that can be shattered by H. If the VC dimension of H is d, then H can shatter any set of d points, but it might not be able to shatter a set of d+1 or more points. The number of ways to divide N points into two classes is actually 2^N (each point can be in class 1 or class 2, independently of the others). So, the corrected statement would be: "For any set of N points, a learning algorithm H can perfectly represent all possible ways of dividing these points into two classes if N is less than or equal to the VC dimension d of algorithm H. For N ¿ d, there is no guarantee that all possible divisions can be represented by H."

Question2:

Predicted Results: [52.35, 68.95, 73.93, 48.2, 78.08, 82.23, 72.27, 88.87, 63.97]

Mean Squared Error: 20.74 The calculated Mean Squared Error (MSE) for the regression model is 20.74. This value represents the average squared deviation of the model's predictions from the actual results. A lower MSE indicates better performance, but without knowing the acceptable error margin for this task, it's hard to definitively say if this performance is good or bad. In my opinion since there are three variable in this table, the current formula is only considering one of the variables. If we could take all three variable into consideration, the predicted result would be more precise and the MSE would be lower.

Question3:

a. In machine learning, dividing a dataset into Training, Validation, and Test sets is crucial for building effective models. The Training set is used to teach the model, the Validation set for tuning hyperparameters and preventing overfitting, and the Test set for evaluating the model's performance on unseen data. The Validation set is particularly important as it allows for model adjustments without compromising the integrity of the Test set, prevent overfitting and ensuring that the final model can generalize well to new data.

b. Using the Training data as both Validation and Test data is problematic. This approach can lead to overfitting, where the model performs exceptionally on familiar data but fails on new, unseen data. It also provides a biased and overly optimistic assessment of the model's performance, as it doesn't get evaluated on independent data. This lack of rigorous testing compromises the model's ability to generalize, which is essential for real-world applications.