MAC0331 - List 1

Matheus T. de Laurentys, 9793714

July 26, 2020

Q 1:
The diagram as a whole shows the concepts, and how they interact, of any supervised machine learning algorithm. The ”probability distribution” box serves to feature the importance of the distribution of the samples generated by the target function and presented by the training examples. The ”Unknown Target Function” box represents any phenomena that one ties to predict with the algorithm. The Ttraining Examples” box represents the a set of points evaluated by the target function, as the arrow suggests, that will be used to train a model. The ”Hypothesis Set” box is used to represent the set of suppositions about the target function, which the learning algorithm uses. The ”Learning Algorithm” is any algorithm that will adjust its parameters based on the examples, and, finally, the ”Final Hypothesis” box indicates that, at the end of the learning algorithm, the model approximates the target function.

Q2.
E_{in}, E_{out} stands for the in-sample, and out-of-samples errors, respectively. The in-sample error is the error present in the learning algorithm in predicting the training examples. On the other hand, the out-of-sample is the error generated when predicting other instances of the problem.

For instance, if, given a sample of 2D points painted red and blue, we create a model M to predict the color of the points. The number of misses divided by the number of points, inside this sample, would be the E_{in} value. The error when painting a different set would be the E_{out}.

Q3.
Minimizing E_{in} as much as possible is not a good idea, because the model would lose generalization. However, if the training set is immense and the sampling is very good, it would be enough to minimize E_{in} in order to generated good results outside the sample.

Q4.
$|E_{in} - E_{out}|$ is the value of the generalization error. It is an important value, which is better the smaller it is, that represents how well the model generalizes. It is good to measure this value to prevent overfitting and underfitting.

Q5.
Mostafa refers to hypothesis as the set of premises about the target function, and, in addition, the set of parameters in the model.

Q6.
The inequality is read as: the probability of the generalization error being greater than a ϵ is smaller or equal to the value $2e^{-2\epsilon^2 N}$. This equality gives a upper bound to the generalization error for a model h.

Q7.
The expression of the previous exercise gives the upper bound to the error of one specific model. This new expression gives a bound to any hypothesis of M. This means that any model $\in M$ has a upper bound.

Q8.
Union-bound is the following lemma:
Let $A_1, ..., A_k$ be k different events.
$P(A_1 \cup \cdots \cup A_k) \leq P(A_1) + \cdots + P(A_k)$

Q9.
Dichotomy is a partition of a set H. That means it is a set C of subsets mutually exclusive whose union is the original set H. The generated dichotomies of a set are all the possible partitions of this set.

Q10.
Growth function is a function that gives the maximum number of ways a set of points can be classified for a hypothesis class. In this context, this function counts the maximum amounts of dichotomies.

Q11.
Counting dichotomies is important because it gives information the complexity of the problem, and this, in turn, gives information on how many perceptrons are required to fit (shatter) the data in a neural network, for example.

Q12.
It is important to know this fact in order to possibly use it in the place of another model M.

Q13.
If the set of points forms a convex shape, then the growth function is exponential. In this case, it is not interesting to use the growth function.

Q14.
VC Dimension is a function that gives how many points can a hypothesis shatter. If this number is larger than the number of dichotomies in a set, then the hypothesis can fit the data, otherwise it cannot,

Q15.
The d_{VC} value for perceptrons is 1 + #dimensions. To prove this one can show that the VC dimension is both larger or equal and smaller or equal to 1 + #dimensions.

Q16.
This statement brings fourth the important aspect of that function. "The more points a hypothesis set can shatter, the more expressive it is" means that it is able to see more differences in sets of points than other, less expressive, sets. For example, a more expressive set might be able to classify very close points differently, while, for other sets of hypothesis, those points are virtually the same, always being classified the same or in fewer ways.

Q17.
The new bound using "Vapnik-Chervonenkis Inequality" is $4m_H(2N)e^{-\frac{1}{8}\epsilon^2 N}$, where m_H is the growth function. However, we can substitute the growth function to one using the d_{VC}:

$$4\sum_{i=0}^{d_{VC}} \binom{2N}{i} e^{-\frac{1}{8}\epsilon^2 N}$$

Q18.
The calculation is possible. In this case, you solve (requires calculating the VC dimension):

$$4\sum_{i=0}^{d_{VC}} \binom{2N}{i} e^{-\frac{1}{8}\epsilon^2 N} = 0.1$$

.

Q19.
Considering a fixed ϵ, the probability of achieving that will increase, however, it is also correct to assume that ϵ decreases as the probability of getting smaller values also increases.

Q20.
A bound to the generalization error is not enough because it does not say anything about the accuracy of the model.

Q21.
I think the VC dimension can be improved depending on special cases, however, in general, I do not see how that is possible. What I mean by this is that, if addition information on point distribution is given, one might be able to improve that bound. it is given that and event does not generate three or more co-linear points, then, a model might fit more points than it is indicated by the VC dimension.

Q22.
VC theory provides bounds for generalization error. If you presume the real data might include points in any possible configuration, receiving any possible classification, then it is important that the model is able to even make that classification (although it might not, based on training). These bounds are based on the number of perceptrons or lines required to divide points in space (any number of dimensions).