10 Learning Theory

Exercise 10.1

Are the following statement regarding the *No Free Lunch* (NFL) theorem true or false? Explain why.

1. On a specific task all the ML algorithms perform in the same way;
2. It is always possible to find a set of data where an algorithm performs arbitrarily bad;
3. In a real scenario, when we are solving a specific task all the concepts f belonging to the concept space $\mathcal{F}$ have the same probability to occur;
4. We can design an algorithm which is always correct on all the samples on every task.

Exercise 10.2

Tell if the following statements about learning theory are true or false. Provide adequate motivations for your answers.

1. We can expect all the learning algorithms to perform equally bad on a given learning concept.
2. In the theory of PAC learning, the value of ϵ controls the probability of incurring in a generalization loss greater than δ on the target concept.
3. The VC dimension of an hypothesis space with infinite cardinality cannot be finite.
4. The VC dimension of a linear classifier in a 1-dimensional space is exactly 2.

Exercise 10.3

1. Show that the VC dimension of an axis aligned rectangle is 4.
2. Show that the VC dimension of a linear classifier in 2D is 3.

3. Show that the VC dimension of a triangle in the plane is at least 7.

4. Show that the VC dimension of a 2D stump, i.e., use either a single horizontal or a single vertical line in 2D to separate points in a plane, is 3.

Exercise 10.4

The VC-dimension of the class H of axis-aligned rectangles is $VC(H) = 4$. How many samples do we need to guarantee that this classifier provides an error larger than $\epsilon = 0.1$ with probability smaller than $\delta = 0.2$?

Exercise 10.5

Consider the hypothesis space of the decision trees with attributes with $n = 4$ binary features with at most $k = 10$ leaves (in this case you have less than $n^{k-1}2^{2k-1}$ different trees) and the problem of binary classification.

Suppose you found a learning algorithm which is able to perfectly classify a training set of $N = 1000$ samples. What is the lowest error ε you can guarantee to have with probability greater than $1 - \delta = 0.95$? How many samples do you need to halve this error?

Another classifier is able only to get an error of $L_{train}(h) = 0.02$ on your original training set. It is possible to use the same error bound derived in the first case? If not, derive a bound with the same probability for this case? How many samples do we need to halve the error bound?