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?

Answers

Answer of exercise 10.1

1. FALSE Given a specific task we are able to find an algorithm which is likely to perform better than a random guess. This does not mean that the algorithm will perform well also on a generic task.

2. TRUE This is exactly how we are able to prove the NFL theorem, i.e., by showing that on a specific concept an algorithm performs arbitrarily bad and, therefore, on average it is not able to beat a random guess.

3. FALSE In the NFL theorem we are considering all the sample as likely as any other to be seen, while in real applications some of them have low probability to happen. This is why we are considering specific algorithms for specific tasks.

4. FALSE The NFL theorem does not allow that, since we can always built an instance where we perform arbitrarily bad. That is why we usually consider PAC-Learning which allows to get a limited amount of mistakes with a given probability.

Answer of exercise 10.2

1. FALSE, according to the No Free Lunch theorem we can expect all the learning algorithms to perform equally bad on the *average* over the concepts. Instead, the specific structure of a given concept can favor an algorithm over another.

2. FALSE, the opposite is true. The value of δ upper bounds the probability of training a model with a generalization loss greater than ϵ.

3. FALSE, the VC dimension is a finer measure of the expressivity of the hypothesis space w.r.t. its cardinality, and it can be finite even if the hypothesis space is not.

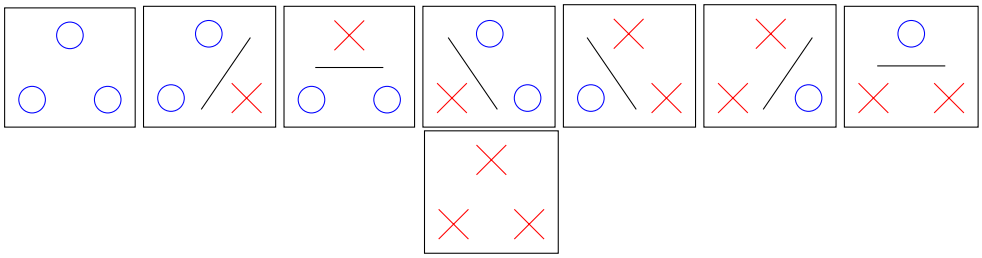
4. TRUE. Let us take two points on the axis `y = 0`, say `x1 = a` and `x2 = b` (`a < b`). For every meaningful combination of labels `(+,+)`, `(-,-)`, `(-,+)`, we can always obtain a perfect classifier with a linear model `x < a`, `x > b`, `a < x < b`, respectively. Instead, if we add a third point `x3 = c` (`b < c`), we cannot correctly classify the instance `(+, -, +)`.

Answer of exercise 10.3

1. Consider 4 points. It is possible to show by enumeration that all the possible labeling are shattered by the rectangle. Consider 5 points. Consider the set of points

with maximum and minimum x coordinate and maximum and minimum y coordinates. If all the points are on the rectangle, we consider the labeling which assign alternate labels to the points if you follow the rectangle perimeter. Otherwise, there are at most 4 points in this set. If we label them $+$ and label $-$ the other, it is not possible to shatter this labeling.

2. Let us call the class of all the linear classifier in 2D $\mathcal{H}$. The proof consists in two steps: $VC(\mathcal{H}) \geq 3$ and $VC(\mathcal{H}) \leq 3$. Let us consider the first step. We need to show that it exists a set of 3 points that can be shattered by the considered hypothesis space $\mathcal{H}$. By considering a set of three non–aligned points, it is possible to show by enumeration that it is possible to shatter them with a linear classifier:



Thus, $VC(\mathcal{H}) \geq 3$. Let us consider the second step. We need to show that it does not exists a set of 4 points which can be shattered by a linear classifier. There are different cases:

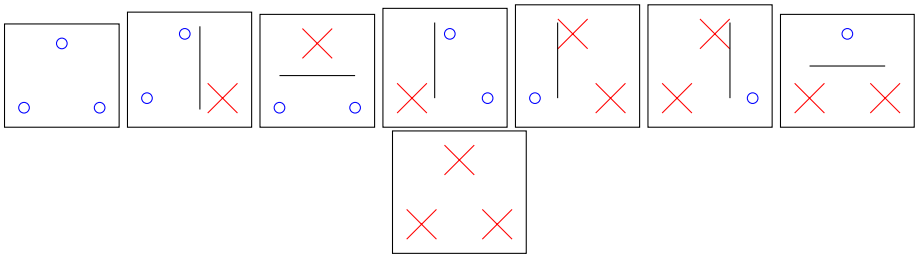
- Four aligned points: if we alternate instances coming from the positive and negative classes we cannot shatter them;
- Three aligned points and a fourth on an arbitrary position: if we alternate instances coming from the positive and negative classes for the three aligned points, we cannot shatter them;
- Four points on a convex hull: if we label the points on the two diagonals with opposite classes, we cannot shatter them;
- Three points on a convex hull (triangle) and one inside the hull: if we label the three points on the triangle with a label and the last one with the other class, we cannot shatter them.

Since there does not exists a configuration where we can shatter the points, we have that $VC(\mathcal{H}) \leq 3$. $\square$

3. If we consider a set of points on a circle it is possible to show by enumeration that a triangle ia able to shatter all of them.

4. Since the decision stumps in 2D are a model which is less flexible than linear boundaries, which have $VC = 3$, they should have $VC(\mathcal{H}) \leq 3$. The proof that

$VC(\mathcal{H}) = 3$ is by enumeration:



Answer of exercise 10.4

We have that we need a number of points:

$$
\begin{aligned}
N &\geq \frac{1}{\epsilon}\left(4\log_2\left(\frac{2}{\delta}\right) + 8VC(H)\log_2\left(\frac{13}{\epsilon}\right)\right) \\
N &\geq \lceil 10(4\log_2(10) + 32\log_2(75)\rceil \\
N &\geq \lceil 40 \cdot 4 + 32 \cdot 7\rceil \\
N &\geq 384
\end{aligned}
$$

Answer of exercise 10.5

Since we have a learner in the version space, we are able to resort to the theorem which states that the error on a hypothesis of the version space ε follows:

$$\mathbb{P}(\exists h \in H,\ L_{true}(h) > \varepsilon) \leq |H|e^{-N\varepsilon} = \delta.$$

Thus:

$$\varepsilon \geq \frac{1}{N}\left(\ln|H| + \ln\left(\frac{1}{\delta}\right)\right) = 0.0286.$$

By looking at this inequality, to halve this error we need to double the number of samples we had originally. Clearly we still require to have have a perfect classifier on the new training set.

In the case we are not allowed to use the previous bound, since the classifier is not able to perfectly classify all the points in the training set. Thus, we might consider an agnostic approach and rely on the fact that:

$$\mathbb{P}(L_{train}(h) - L_{true}(h) > \varepsilon) \leq |H|e^{-2N\varepsilon^2} = \delta.$$

Thus the error bound is:

$$err = L_{train}(h) + \varepsilon \leq L_{train}(h) + \sqrt{\frac{\ln|H| - \ln(\delta)}{2N}} = 0.1397.$$

If we want to halve the error bound we should have $err' = \frac{err}{2}$ and we have:

$$L_{train}(h) + \sqrt{\frac{\ln|H| - \ln(\delta)}{2N'}} \leq err'$$
$$N' \geq \frac{\ln|H| - \ln(\delta)}{2(\frac{err}{2} - L_{train})^2} = 5766.34 \approx 5767.$$