7 Learning Theory

Exercise 7.1

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?

Exercise 7.2

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 7.3

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 7.4

Show that the VC dimension of a closed interval $[a, b]$ on $\mathbb{R}$ is 2. Provide a PAC bound with confidence at least $1 - \delta = 1 - 4e^{-7}$ for the previous concept when we have $N = \lfloor e^{11} - 1 \rfloor$ samples and an error on the training set of $L_{train}(h) = \frac{1}{e^{10}}$.

Exercise 7.5

Show that the VC dimension of a plane $\mathcal{P}$ in $3D$ is 4 (you cannot derive it as a corollary of the ND theorem). Hint: showing that $VC(\mathcal{P}) < 5$ requires to check three different cases, one of which is a degenerate case.

Exercise 7.6

Consider the hypothesis space of the decision trees with $n = 4$ binary features with at most $k = 3$ leaves (in this case you have less than $n^{k-1}2^{2k-1}$ different trees) and the problem of binary classification. Assume to have a learning algorithm which is able to perfectly classify the training set of $N = 28$ samples.

1. What is the lowest error ε you can guarantee to have with probability greater than $1 - \delta = 1 - 2^{-5}$?

2. How many samples do you need to halve this error? Do we need some property on the classifier of these new samples so that the error bound is still valid?

Justify your answers properly. Moreover, recall that:

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

and that $\frac{1}{\log_2(e)} \approx 0.694$ and $\frac{1}{\log_7(e)} = 1.94$.

Exercise 7.7

You train a binary classifier y from the hypothesis space $\mathcal{H}$ with finite VC-dimension ν. Training is performed over the dataset $\mathcal{D}_{\text{train}}$ made of N samples using the loss function:

$$\mathcal{L}_{\text{train}} = \frac{1}{N} \sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{train}}} \ell(y(\mathbf{x}), t),$$

with $\ell(y(\mathbf{x}), t) \in [0, L]$.

Tell if the following estimators the true loss of model y (i.e., $\mathcal{L}_{\text{true}} = \mathbb{E}_{\mathbf{x},t}[\ell(y(\mathbf{x}), t)|y]$) lead to unbiased, negatively biased, or positively biased (specify also if they are consistent). For each of them, provide an upper bound of the true loss in terms of the estimator holding with at least probability $1 - \delta$.

1. The training loss $\mathcal{L}_{\text{train}}$.
2. The loss computed over a test set $\mathcal{D}_{\text{test}}$ made of M samples:

$$\mathcal{L}_{\text{test}} = \frac{1}{M} \sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{test}}} \ell(y(\mathbf{x}), t).$$

In which circumstances (if any) the training error is more accurate w.r.t. the test error as an estimator for the true error?

Exercise 7.8

You train two binary classifiers h_1 and h_2 belonging to the hypothesis spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ with VC-dimension $\nu_1 = 2$ and $\nu_2 = 10$, respectively. Training is performed on the same dataset made of $N = 1000$ independent samples. The training accuracy of the two classifier are $\mathcal{L}_1 = 0.6$ and $\mathcal{L}_2 = 0.8$. Answer to the following questions, justifying your answers.

1. Given the training accuracies, can you can say that model h_2 has higher accuracy than model h_1 with at least confidence 0.8?
2. Suppose you evaluate the accuracy of the two models on the same test set (independent and identically distributed as the training set) made of M samples and obtain the test accuracies $\tilde{\mathcal{L}}_1 = 0.55$ and $\tilde{\mathcal{L}}_2 = 0.7$. Which is the minimum size of M such that with confidence at least 0.8 you can say that model h_2 has higher accuracy than model h_1?

Exercise 7.9

You train a supervised learning model $y(\mathbf{x})$ minimizing a loss function ℓ (bounded in $[0, L]$) over a training set $\mathcal{D}_{\text{train}}$ made of N independent samples. You have at your disposal K test sets $\mathcal{D}^{(1)}_{\text{test}}, \ldots, \mathcal{D}^{(K)}_{\text{test}}$ each made of M samples and independent one another and independent of the training set. Tell if the following estimators of the true loss of the learned model (i.e., $\mathcal{L}_{\text{true}} = \mathbb{E}_{\mathbf{x},t}[\ell(y(\mathbf{x}, t))]$) are unbiased, negatively biased, or positively biased. For each of them, provide an upper bound of the true loss in terms of the estimator holding with at least probability $1 - \delta$.

1. The loss over the first test set $\mathcal{L}^{(1)}_{\text{test}} = \frac{1}{M} \sum_{(\mathbf{x},t)\in\mathcal{D}^{(1)}_{\text{test}}} \ell(y(\mathbf{x}), t)$
2. The minimum loss among the test sets losses $\mathcal{L}_{\min} = \min_{i\in\{1,\ldots,K\}} \mathcal{L}^{(i)}_{\text{test}}$

3. The maximum loss among the test sets losses $\mathcal{L}_{\max} = \max_{i \in \{1,\dots,K\}} \mathcal{L}_{\text{test}}^{(i)}$
4. The average loss over the test sets losses $\mathcal{L}_{\text{avg}} = \frac{1}{K} \sum_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)}$

Which of the above alternative would you prefer to provide an estimate of the true error of the learned model?

Exercise 7.10

You have at your disposal K independent and identically distributed training datasets $\mathcal{D}_{\text{train}}^{(1)}, \dots, \mathcal{D}_{\text{train}}^{(K)}$, made of N samples each, on which you learn the binary classifier $y_1, \dots, y_K$ all from the same hypothesis space $\mathcal{H}$ with VC-dimension ν with a training accuracy of $\mathcal{L}_1, \dots, \mathcal{L}_K$, respectively. Answer to the following questions, justifying your answers.

- You decide to deploy the classifier with smaller training error y_{i^*} with $i^* \in \arg\max_{i \in \{1,\dots,K\}} \mathcal{L}_i$. Can you provide a lower bound of the true accuracy of y_{i^*}?
- You have at your disposal one validation set $\mathcal{D}_{\text{val}}$ (independent and identically distributed as the training sets) of M samples and you evaluate the performance of the learned models on it, obtaining a validation accuracy of $\hat{\mathcal{L}}_1, \dots, \hat{\mathcal{L}}_K$, respectively. Then, you decide to deploy the classifier with smaller validation error $y_{\hat{i}}$ with $\hat{i} \in \arg\max_{i \in \{1,\dots,K\}} \hat{\mathcal{L}}_i$. Can you provide a lower bound of the true accuracy of $y_{\hat{i}}$?
- You have at your disposal K validation sets $\mathcal{D}_{\text{val}}^{(1)}, \dots, \mathcal{D}_{\text{val}}^{(K)}$ (independent and identically distributed as the training sets and among them) of M samples each and you evaluate the performance of each learned model y_i on one dataset $\mathcal{D}_{\text{val}}^{(i)}$, obtaining a validation accuracy of $\tilde{\mathcal{L}}_1, \dots, \tilde{\mathcal{L}}_K$, respectively. Then, you decide to deploy the classifier with smaller validation error $y_{\tilde{i}}$ with $\tilde{i} \in \arg\max_{i \in \{1,\dots,K\}} \tilde{\mathcal{L}}_i$. Can you provide a lower bound of the true accuracy of $y_{\tilde{i}}$?

Exercise 7.11

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 7.12

You are given with the following class of regression models $\mathcal{M}_\psi$ over the variables $\{x_1, x_2\}$ and an additional feature $\psi(x_1, x_2)$:

$$\mathcal{M}_\psi : \qquad y_{\mathbf{w}}(x_1, x_2) = w_0 + w_1 x_1 + w_2 x_2^2 + w_3 \psi(x_1, x_2), \qquad \mathbf{w} \in \mathbb{R}^4.$$

Consider three possible choices of feature ψ:

$$\psi_1(x_1, x_2) = 1, \qquad \psi_2(x_1, x_2) = x_1 - 3x_2^2, \qquad \psi_3(x_1, x_2) = x_1 x_2.$$

Tell if the following statements are true or false. Motivate your answers.

1. Models $\mathcal{M}_{\psi_1}$ and $\mathcal{M}_{\psi_2}$ have the same bias.
2. The bias of model $\mathcal{M}_{\psi_1}$ is larger than the bias of model $\mathcal{M}_{\psi_3}$.
3. The VC dimension of model $\mathcal{M}_{\psi_2}$ is larger than the VC dimension of model $\mathcal{M}_{\psi_3}$.
4. The VC dimension of model $\mathcal{M}_{\psi_1+\psi_2}$ is larger than the VC dimension of model $\mathcal{M}_{\psi_3}$.

Answers

Answer of exercise 7.1

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:

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

Answer of exercise 7.2

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.

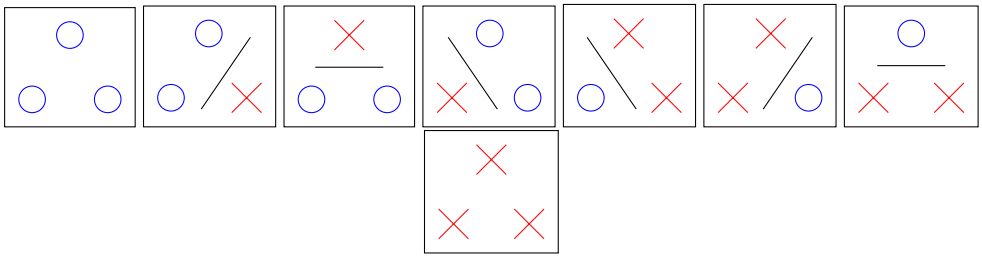
3. FALSE In the NFL theorem we are considering all the sample as likely as any other to be seen, while in real applications some of them have low probability to happen. This is why we are considering specific algorithms for specific tasks.

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 7.3

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

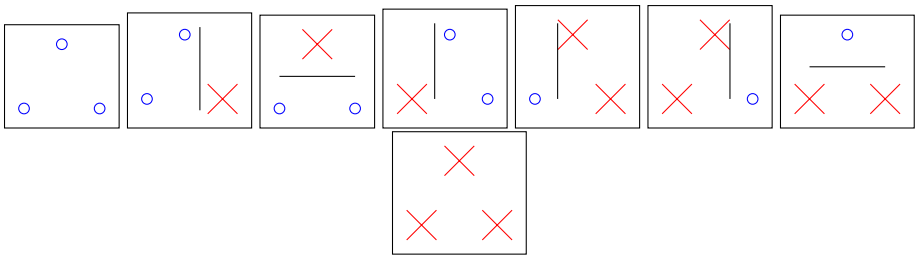
with opposite classes, we cannot shatter them;

- 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 $VC(\mathcal{H}) \leq 3$. □

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 7.4

To show that the classifier H has $VC(H) = 2$ we start by showing that $VC(H) \geq 2$ and then that $VC(H) < 3$. By enumeration (pick any two points $x_1 < x_2$ on the real line) we have that an interval $[a, b]$ can separate any partition of these points, therefore $VC(H) \geq 2$. Select then 3 points on the real line x_1, x_2 and x_3. If two of them coincides you are not able to shatter them. Conversely assume that they are distinct and without loss of generality $x_1 < x_2 < x_3$. If we assign a alternate labels to these points we are not able to shatter them. Thus, $VC(H) < 3$, which concludes the proof.

The PAC bound for infinite continuous hypothesis space is:

$$L_{true} \leq L_{train} + \sqrt{\frac{VC(H)\left(\ln(\frac{2N}{VC(H)}) + 1\right) + \ln\frac{4}{\delta}}{N}}$$

Substituting we have:

$$L_{true} \leq \frac{1}{e^{10}} + \sqrt{\frac{2(\ln(\lfloor e^{11} - 1 \rfloor) + 1) + 7}{\lfloor e^{11} - 1 \rfloor}} \leq \frac{1}{e^{10}} + \sqrt{\frac{31}{e^{10}}}.$$

Answer of exercise 7.5

TODO

Answer of exercise 7.6

TODO

Answer of exercise 7.7

1. NEGATIVELY BIASED but CONSISTENT, since the training set is used for learning the model y and it is statistically dependent to it. Nevertheless, the estimate is consistent since it approaches the true mean as $N \to +\infty$. To provide a confidence interval, we need the VC-dimension bound:

$$\mathcal{L}_{\text{true}} - \mathcal{L}_{\text{train}} \leq L\sqrt{\frac{\nu \log\left(\frac{2eN}{\nu}\right) + \log\left(\frac{4}{\delta}\right)}{N}}.$$

2. UNBIASED and CONSISTENT, since the test set is independent and identically distributed as the training set. The estimate is consistent since it approaches the true mean as $N \to +\infty$. To provide a confidence interval, it suffices to use Höeffding's inequality:

$$\mathcal{L}_{\text{true}} - \mathcal{L}_{\text{test}} \leq L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}}.$$

The training error is more accurate w.r.t. the test error as an estimator for the true error, when the size M of the test set is significantly smaller than the size of the training set N. More formally:

$$L\sqrt{\frac{\nu \log\left(\frac{2eN}{\nu}\right) + \log\left(\frac{4}{\delta}\right)}{N}} > L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} \implies M < \frac{N \log\left(\frac{1}{\delta}\right)}{2\left(\nu \log\left(\frac{2eN}{\nu}\right) + \log\left(\frac{4}{\delta}\right)\right)} \approx \frac{N}{2\nu},$$

where the last approximation ignores the logarithmic terms.

Answer of exercise 7.8

The rationale behind the solution of this exercise is to construct an inequality for the true losses $\mathcal{L}_{\text{true},1}$ and $\mathcal{L}_{\text{true},2}$ expressed in terms of the training and/or test losses that hold with the prescribed probability. Specifically, we search for inequalities of the form $\mathcal{L}_{\text{true},1} \leq \text{bound}_1$ and $\mathcal{L}_{\text{true},2} \geq \text{bound}_2$. Then, if $\text{bound}_1 \leq \text{bound}_2$, we are guaranteed that $\mathcal{L}_{\text{true},1} \leq \mathcal{L}_{\text{true},2}$.

1. We know from the VC-dimension learning bound that we can upper bound the true accuracy with the training accuracy:

$$\mathcal{L}_{\text{true},1} \leq \mathcal{L}_1 + \sqrt{\frac{\nu_1 \log\left(\frac{2eN}{\nu_1}\right) + \log\left(\frac{4}{\delta}\right)}{N}} \qquad \text{w.p. } 1-\delta \tag{7.1}$$

$$\mathcal{L}_{\text{true},2} \geq \mathcal{L}_2 - \sqrt{\frac{\nu_2 \log\left(\frac{2eN}{\nu_2}\right) + \log\left(\frac{4}{\delta}\right)}{N}} \qquad \text{w.p. } 1-\delta \tag{7.2}$$

Since, each of the inequality hold w.p. (with probability) at least $1-\delta$, to compute the probability that both hold simultaneously, we need to apply Bool's inequality (a.k.a. union bound) $\Pr(A \vee B) \leq \Pr(A) + \Pr(B)$:

$$\begin{aligned}\Pr(\text{(7.1) holds} \wedge \text{(7.2) holds}) &= 1 - \Pr(\text{(7.1) not holds} \vee \text{(7.2) not holds}) \\ &\geq 1 - \Pr(\text{(7.1) not holds}) + \Pr(\text{(7.2) not holds}) \\ &\geq 1 - \delta - \delta.\end{aligned}$$

Thus, both hold simultaneously w.p. at least $1-2\delta$. Thus, w.p. at least $1-2\delta$ we are guaranteed that $\mathcal{L}_{\text{true},1} \leq \mathcal{L}_{\text{true},2}$, whenever:

$$\mathcal{L}_1 + \sqrt{\frac{\nu_1 \log\left(\frac{2eN}{\nu_1}\right) + \log\left(\frac{4}{\delta}\right)}{N}} < \mathcal{L}_2 - \sqrt{\frac{\nu_2 \log\left(\frac{2eN}{\nu_2}\right) + \log\left(\frac{4}{\delta}\right)}{N}}$$

By setting δ such that $1-2\delta = 0.8$, we have $0.7397 < 0.5418$ which is false.

2. Since the test set is independent on the training set, we know from Höeffding's inequality that:

$$\mathcal{L}_{\text{true},1} \leq \tilde{\mathcal{L}}_1 + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} \qquad \text{w.p. } 1-\delta \tag{7.3}$$

$$\mathcal{L}_{\text{true},2} \geq \tilde{\mathcal{L}}_2 - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} \qquad \text{w.p. } 1-\delta \tag{7.4}$$

Using the same union bound argument, we have that, w.p. at least $1-2\delta$ we are guaranteed that $\mathcal{L}_{\text{true},1} \leq \mathcal{L}_{\text{true},2}$, whenever:

$$\tilde{\mathcal{L}}_1 + \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} < \tilde{\mathcal{L}}_2 - \sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} \implies M > \frac{2\log\left(\frac{1}{\delta}\right)}{(\tilde{\mathcal{L}}_2 - \tilde{\mathcal{L}}_1)^2} = 204.67.$$

having set δ such that $1-2\delta = 0.8$. So, we have to enforce $M \geq 205$.

Answer of exercise 7.9

1. UNBIASED, since the test set $\mathcal{D}_{\text{test}}^{(1)}$ is independent and identically distributed as the training set $\mathcal{D}_{\text{train}}$. The confidence interval can be computed using standard Höeffding's inequality since all samples $\ell(y(\mathbf{x}), t)$ are independent:

$$\mathcal{L}_{\text{true}} \leq \mathcal{L}_{\text{test}}^{(1)} + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2M}} \qquad \text{w.p. } 1-\delta.$$

2. NEGATIVELY BIASED, since we are using as evaluation the best among loss a set of independent ones and, thus, it will result in a negatively biased estimator. To obtain a confidence interval, we need to proceed more carefully with a union bound[1] followed by an application of Höeffding's inequality. Let $\epsilon > 0$:

$$\begin{aligned}
\Pr\left(\min_{i\in\{1,\dots,K\}} \mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) &= \Pr\left(\bigcup_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) \\
&\overset{\text{union bound}}{\leq} \sum_{i=1}^{K} \Pr\left(\mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) \\
&\overset{\text{Höeffding's inequality}}{\leq} K\exp\left(-\frac{2\epsilon^2 M}{L^2}\right).
\end{aligned}$$

   By setting the last expression equal to δ and solving for ϵ, we obtain:

$$\mathcal{L}_{\text{true}} \leq \mathcal{L}_{\min} + L\sqrt{\frac{\log\left(\frac{K}{\delta}\right)}{2M}} \qquad \text{w.p. } 1-\delta.$$

[1] One could avoid the application of the union bound by exploiting the independence between the tests sets at the price of a more complex expression:

$$\begin{aligned}
\Pr\left(\bigcup_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) &= 1 - \Pr\left(\bigcap_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)} > \mathcal{L}_{\text{true}} - \epsilon | y\right) \\
&\overset{\text{independence}}{=} 1 - \prod_{i=1}^{K} \Pr\left(\mathcal{L}_{\text{test}}^{(i)} > \mathcal{L}_{\text{true}} - \epsilon | y\right) \\
&= 1 - \prod_{i=1}^{K} \left(1 - \Pr\left(\mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right)\right) \\
&\overset{\text{Höeffding's inequality}}{\leq} 1 - \left(1 - \left(-\frac{2\epsilon^2 M}{L^2}\right)\right)^K.
\end{aligned}$$

   From which, setting the latter expression equal to δ and solving for ϵ, we obtain:

$$\mathcal{L}_{\text{true}} \leq \mathcal{L}_{\min} + L\sqrt{\frac{\log\left(\frac{1}{1-(1-\delta)^{1/K}}\right)}{2M}} \qquad \text{w.p. } 1-\delta.$$

   Notice that $1-(1-\delta)^{1/K} \sim \delta/K$ when $\delta \to 0$, recovering the expression obtained with the union bound.

Notice how the value of K appears inside the logarithm.

3. POSITIVELY BIASED, since using as evaluation the worst among loss a set of independent ones and, thus, it will result in a positively biased estimator. To obtain a confidence interval, we need to proceed more carefully exploiting the independence of the tests sets, followed by an application of Höeffding's inequality.

$$\begin{aligned}\Pr\left(\max_{i\in\{1,\dots,K\}} \mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) &= \Pr\left(\bigcap_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) \\ &\overset{\text{independence}}{=} \prod_{i=1}^{K} \Pr\left(\mathcal{L}_{\text{test}}^{(i)} \leq \mathcal{L}_{\text{true}} - \epsilon | y\right) \\ &\overset{\text{Höeffding's inequality}}{\leq} \exp\left(-\frac{\epsilon^2 M}{L^2}\right)^K = \exp\left(-\frac{2\epsilon^2 K M}{L^2}\right).\end{aligned}$$

By setting the last expression equal to δ and solving for ϵ, we obtain:

$$\mathcal{L}_{\text{true}} \leq \mathcal{L}_{\max} + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2KM}} \qquad \text{w.p. } 1-\delta.$$

4. UNBIASED, since we are averaging unbiased estimators of the form of $\mathcal{L}_{\text{test}}^{(i)}$. Since all of them are computed over same-size independent test sets, we can look at the resulting estimator as the loss over a test set $\mathcal{D}_{\text{test}} = \cup_{i=1}^{K} \mathcal{D}_{\text{test}}^{(i)}$ given by the union of the test sets $\mathcal{D}_{\text{test}}^{(i)}$:

$$\mathcal{L}_{\text{avg}} = \frac{1}{K}\sum_{i=1}^{K} \mathcal{L}_{\text{test}}^{(i)} = \frac{1}{K}\sum_{i=1}^{K} \frac{1}{M} \sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{test}}^{(i)}} \ell(y(\mathbf{x}),t) = \frac{1}{KM} \sum_{(\mathbf{x},t)\in\mathcal{D}_{\text{test}}} \ell(y(\mathbf{x}),t).$$

Thus, we can apply standard Höeffding's inequality:

$$\mathcal{L}_{\text{true}} \leq \mathcal{L}_{\text{avg}} + L\sqrt{\frac{\log\left(\frac{1}{\delta}\right)}{2KM}} \qquad \text{w.p. } 1-\delta.$$

Answer of exercise 7.10

TODO

Answer of exercise 7.11

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 7.12

1. TRUE. Indeed, they are the very same model after a renaming of the weights:

$$\begin{aligned}\mathcal{M}_{\psi_1}: \quad y_{\mathbf{w}}(x_1,x_2) &= w_0 + w_1x_1 + w_2x_2^2 + w_3 \\ &= \underbrace{w_0 + w_3}_{w_0^{(1)}} + \underbrace{w_1}_{w_1^{(1)}} x_1 + \underbrace{w_2}_{w_2^{(1)}} x_2^2, \qquad \mathbf{w}^{(1)} \in \mathbb{R}^3,\end{aligned}$$

$$\begin{aligned}\mathcal{M}_{\psi_2}: \quad y_{\mathbf{w}}(x_1,x_2) &= w_0 + w_1x_1 + w_2x_2^2 + w_3(x_1 - 3x_2^2) \\ &= \underbrace{w_0}_{w_0^{(2)}} + \underbrace{(w_1 + w_3)}_{w_1^{(2)}} x_1 + \underbrace{(w_2 - 3w_3)}_{w_2^{(2)}} x_2^2, \qquad \mathbf{w}^{(2)} \in \mathbb{R}^3.\end{aligned}$$

2. TRUE. Indeed, models $\mathcal{M}_{\psi_3}$ are a super-set of models $\mathcal{M}_{\psi_1}$. Formally:

$$\begin{aligned}\mathcal{M}_{\psi_1}: \quad & y_{\mathbf{w}}(x_1,x_2) = w_0 + w_1x_1 + w_2x_2^2 + w_3, \\ \mathcal{M}_{\psi_3}: \quad & y_{\mathbf{w}}(x_1,x_2) = w'_0 + w'_1x_1 + w'_2x_2^2 + w'_3x_1x_2.\end{aligned}$$

   To get $\mathcal{M}_{\psi_1}$ from $\mathcal{M}_{\psi_3}$, simply set $w'_0 = w_0 + w_3$, $w'_1 = w_1$, $w'_2 = w_2$, and $w'_3 = 0$.

3. FALSE. For the same reason as the previous question, recalling that models $\mathcal{M}_{\psi_2}$ are the same as models $\mathcal{M}_{\psi_1}$.

4. FALSE. Indeed, models $\mathcal{M}_{\psi_1+\psi_2}$ are the same as models $\mathcal{M}_{\psi_1}$ (or models $\mathcal{M}_{\psi_2}$).