# 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 $\varepsilon$ 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 $\varepsilon$ 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 $\nu$. 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 error 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 errors, can you can say that model $h_2$ has larger error than model $h_1$ with at least confidence $0.8$?
2. Suppose you evaluate the errors 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 errors $\tilde{\mathcal{L}}_1 = 0.55$ 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 larger errors than model $h_1$?

## Exercise 7.9

You train a supervised learning model $y(\mathbf{x})$ minimizing a loss function $\ell$ (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 $\nu$ 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 $\epsilon$ controls the probability of incurring in a generalization loss greater than $\delta$ 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 $\psi$:

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

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 $\mathcal{M}_{\psi_3}$.