Solutions to Exercises

Chapter 2

2.1 Two-oracle variant of the PAC model

- Assume that $\mathcal{C}$ is efficiently PAC-learnable using $\mathcal{H}$ in the standard PAC model using algorithm $\mathcal{A}$. Consider the distribution $\mathcal{D} = \frac{1}{2}(\mathcal{D}_- + \mathcal{D}_+)$. Let $h \in \mathcal{H}$ be the hypothesis output by $\mathcal{A}$. Choose δ such that:

$$\mathbb{P}[R_{\mathcal{D}}(h) \leq \epsilon/2] \geq 1 - \delta.$$

From

$$\begin{aligned} R_{\mathcal{D}}(h) &= \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[h(x) \neq c(x)] \\ &= \frac{1}{2}\Big(\mathop{\mathbb{P}}_{x \sim \mathcal{D}_-}[h(x) \neq c(x)] + \mathop{\mathbb{P}}_{x \sim \mathcal{D}_+}[h(x) \neq c(x)]\Big) \\ &= \frac{1}{2}(R_{\mathcal{D}_-}(h) + R_{\mathcal{D}_+}(h)), \end{aligned}$$

it follows that:

$$\mathbb{P}[R_{\mathcal{D}_-}(h) \leq \epsilon] \geq 1 - \delta \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h) \leq \epsilon] \geq 1 - \delta.$$

This implies two-oracle PAC-learning with the same computational complexity.

- Assume now that $\mathcal{C}$ is efficiently PAC-learnable in the two-oracle PAC model. Thus, there exists a learning algorithm $\mathcal{A}$ such that for $c \in \mathcal{C}$, $\epsilon > 0$, and $\delta > 0$, there exist m_- and m_+ polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, such that if we draw m_- negative examples or more and m_+ positive examples or more, with confidence $1 - \delta$, the hypothesis h output by $\mathcal{A}$ verifies:

$$\mathbb{P}[R_{\mathcal{D}_-}(h)] \leq \epsilon \quad \text{and} \quad \mathbb{P}[R_{\mathcal{D}_+}(h)] \leq \epsilon.$$

Now, let $\mathcal{D}$ be a probability distribution over negative and positive examples. If we could draw m examples according to $\mathcal{D}$ such that $m \geq \max\{m_-, m_+\}$, m polynomial in $1/\epsilon$, $1/\delta$, and $size(c)$, then two-oracle PAC-learning would imply standard PAC-learning:

$$\begin{aligned} &\mathbb{P}[R_{\mathcal{D}}(h)] \\ &\qquad \leq \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 0]\,\mathbb{P}[c(x) = 0] + \mathbb{P}[R_{\mathcal{D}}(h)|c(x) = 1]\,\mathbb{P}[c(x) = 1] \\ &\qquad \leq \epsilon(\mathbb{P}[c(x) = 0] + \mathbb{P}[c(x) = 1]) = \epsilon. \end{aligned}$$

If $\mathcal{D}$ is not too biased, that is, if the probability of drawing a positive example, or that of drawing a negative example is more than ϵ, it is not hard to show, using Chernoff bounds or just Chebyshev's inequality, that drawing a polynomial number of examples in $1/\epsilon$ and $1/\delta$ suffices to guarantee that $m \geq \max\{m_-, m_+\}$ with high confidence.

Otherwise, $\mathcal{D}$ is biased toward negative (or positive examples), in which case returning $h = h_0$ (respectively $h = h_1$) guarantees that $\mathbb{P}[R_{\mathcal{D}}(h)] \leq \epsilon$.

To show the claim about the not-too-biased case, let S_m denote the number of positive examples obtained when drawing m examples when the probability of a positive example is ϵ. By Chernoff bounds,

$$\mathbb{P}[S_m \leq (1-\alpha)m\epsilon] \leq e^{-m\epsilon\alpha^2/2}.$$

We want to ensure that at least m_+ examples are found. With $\alpha = \frac{1}{2}$ and $m = \frac{2m_+}{\epsilon}$,

$$\mathbb{P}[S_m > m_+] \leq e^{-m_+/4}.$$

Setting the bound to be less than or equal to $\delta/2$, leads to the following condition on m:

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}$$

A similar analysis can be done in the case of negative examples. Thus, when $\mathcal{D}$ is not too biased, with confidence $1-\delta$, we will find at least m_- negative and m_+ positive examples if we draw m examples, with

$$m \geq \min\{\frac{2m_+}{\epsilon}, \frac{2m_-}{\epsilon}, \frac{8}{\epsilon}\log\frac{2}{\delta}\}.$$

In both solutions, our training data is the set T and our learned concept $L(T)$ is the tightest circle (with minimal radius) which is consistent with the data.

2.2 PAC learning of hyper-rectangles

The proof in the case of hyper-rectangles is similar to the one given presented within the chapter. The algorithm selects the tightest axis-aligned hyper-rectangle containing all the sample points. For $i \in [2n]$, select a region r_i such that $\mathbb{P}_{\mathcal{D}}[r_i] = \epsilon/(2n)$ for each edge of the hyper-rectangle R. Assuming that $\mathbb{P}_{\mathcal{D}}[R - R'] > \epsilon$, argue that R' cannot meet all r_is, so it must miss at least one. The probability that none of the m sample points falls into region r_i is $(1-\epsilon/2n)^m$. By the union bound, this shows that

$$\mathbb{P}[R(R') > \epsilon] \leq 2n(1-\epsilon/2n)^m \leq 2n\exp\Big(-\frac{\epsilon m}{2n}\Big). \tag{E.35}$$

Setting δ to the right-hand side shows that for

$$m \geq \frac{2n}{\epsilon}\log\frac{2n}{\delta}, \tag{E.36}$$

with probability at least $1-\delta$, $R_{\mathcal{D}}(R') \leq \epsilon$.

2.3 Concentric circles

Suppose our target concept c is the circle around the origin with radius r. We will choose a slightly smaller radius s by

$$s := \inf\{s' \colon P(s' \leq ||x|| \leq r) < \epsilon\}.$$

Let A denote the annulus between radii s and r; that is, $A := \{x \colon s \leq ||x|| \leq r\}$. By definition of s,

$$P(A) \geq \epsilon. \tag{E.37}$$

In addition, our generalization error, $P(c \,\Delta\, L(T))$, must be small if T intersects A. We can state this as

$$P(c \,\Delta\, L(T)) > \epsilon \implies T \cap A = \emptyset. \tag{E.38}$$

Using (E.37), we know that any point in T chosen according to P will "miss" region A with probability at most $1-\epsilon$. Defining $error := P(c \,\Delta\, L(T))$, we can combine this with (E.38) to see that

$$P(error > \epsilon) \leq P(T \cap A = \emptyset) \leq (1-\epsilon)^m \leq e^{-m\epsilon}.$$

Setting δ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\epsilon}\log(\frac{1}{\delta})$.

2.4 Non-concentric circles

As in the previous example, it is natural to assume the learning algorithm operates by returning the smallest circle which is consistent with the data. Gertrude is relying on the logical implication

$$error > \epsilon \implies T \cap r_i = \emptyset \text{ for some } i, \tag{E.39}$$

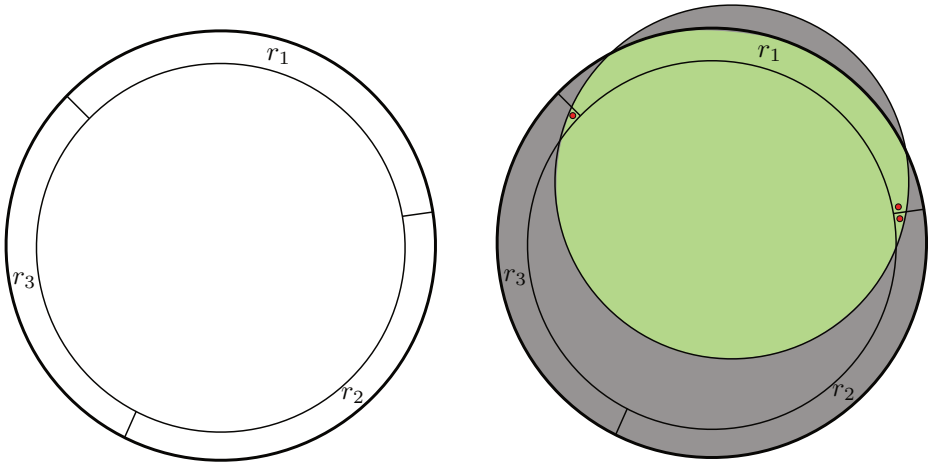

Figure E.5
Counter-example shows error of tightest circle in gray.

which is not necessarily true here. Figure E.5 illustrates a counterexample. In the figure, we have one training point in each region r_i. The points in r_1 and r_2 are very close together, and the point in r_3 is very close to region r_1. On this training data (some other points may be included outside the three regions r_i), our learned circle is the "tightest" circle including these points, and hence one diameter approximately traverses the corners of r_1. In the figure, the gray regions are the error of this learned hypotheses versus the target circle, which has a thick border. Clearly, the error may be greater than ϵ even while $T \cap r_i \neq \emptyset$ for any i; this contradicts (E.39) and invalidates poor Gertrude's proof.

2.5 Triangles

As in the case of axis-aligned rectangles, consider three regions r_1, r_2, r_3, along the sides of the target concept as indicated in figure E.6. Note that the triangle formed by the points $A", B", C"$ is similar to ABC (same angles) since $A"B"$ must be parallel to AB, and similarly for the other sides.

Assume that $\mathbb{P}[ABC] > \epsilon$, otherwise the statement would be trivial. Consider a triangle $A'B'C'$ similar to ABC and consistent with the training sample and such that it meets all three regions r_1, r_2, r_3.

Since it meets r_1, the line $A'B'$ must be below $A"B"$. Since it meets r_2 and r_3, A' must be in r_2 and B' in r_3 (see figure E.6). Now, since the angle $\widehat{A'B'C'}$ is equal to $\widehat{A"B"C"}$, C' must be necessarily above $C"$. This implies that triangle $A'B'C'$ contains $A"B"C"$, and thus $error(A'B'C') \leq \epsilon$.

$$error(A'B'C') > \epsilon \implies \exists i \in \{1,2,3\} \colon A'B'C' \cap r_i = \emptyset.$$

Thus, by the union bound,

$$\mathbb{P}[error(A'B'C') > \epsilon] \leq \sum_{i=1}^{3} \mathbb{P}[A'B'C' \cap r_i = \emptyset] \leq 3(1-\epsilon/3)^m \leq 3e^{-3m\epsilon}.$$

Setting δ to match the right-hand side gives the sample complexity $m \geq \frac{3}{\epsilon} \log \frac{3}{\delta}$.

2.8 Learning intervals

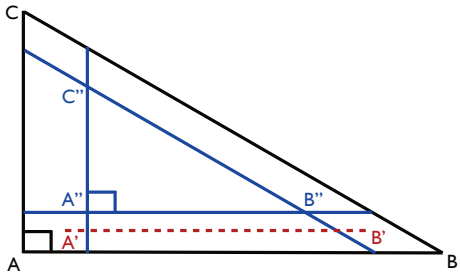

Figure E.6
Rectangle triangles.

Given a sample S, one algorithm consists of returning the tightest closed interval I_S containing positive points. Let $I = [a, b]$ be the target concept. If $\mathbb{P}[I] < \epsilon$, then clearly $R(I_S) < \epsilon$. Assume that $\mathbb{P}[I] \geq \epsilon$. Consider two intervals I_L and I_R defined as follows:

$$I_L = [a, x] \quad \text{with } x = \inf\{x \colon \mathbb{P}[a, x] \geq \epsilon/2\}$$
$$I_R = [x', b] \quad \text{with } x' = \sup\{x' \colon \mathbb{P}[x', b] \geq \epsilon/2\}.$$

By the definition of x, the probability of $[a, x[$ is less than or equal to $\epsilon/2$, similarly the probability of $]x', b]$ is less than or equal to $\epsilon/2$. Thus, if I_S overlaps both with I_L and I_R, then its error region has probability at most ϵ. Thus, $R(I_S) > \epsilon$ implies that I_S does not overlap with either I_L or I_R, that is either none of the training points falls in I_L or none falls in I_R. Thus, by the union bound,

$$\begin{aligned}\mathbb{P}[R(I_S) > \epsilon] &\leq \mathbb{P}[S \cap I_L = \emptyset] + \mathbb{P}[S \cap I_R = \emptyset] \\ &\leq 2(1 - \epsilon/2)^m \leq 2e^{-m\epsilon/2}.\end{aligned}$$

Setting δ to match the right-hand side gives the sample complexity $m = \frac{2}{\epsilon} \log \frac{2}{\delta}$ and proves the PAC-learning of closed intervals. □

2.9 Learning union of intervals

Given a sample S, our algorithm consists of the following steps:

(a) Sort S in ascending order.

(b) Loop through sorted S, marking where intervals of consecutive positively labeled points begin and end.

(c) Return the union of intervals found on the previous step. This union is represented by a list of tuples that indicate start and end points of the intervals.

This algorithms works both for $p = 2$ and for a general p. We will now consider the problem for $\mathcal{C}_2$. To show that this is a PAC-learning algorithm we need to distinguish between two cases.

The first case is when our target concept is a disjoint union of two closed intervals: $I = [a, b] \cup [c, d]$. Note, there are two sources of error: false negatives in $[a, b]$ and $[c, d]$ and also false positives in (b, c). False positives may occur if no sample is drawn from (b, c). By linearity of expectation and since these two error regions are disjoint, we have that $R(h_S) = R_{\text{FP}}(h_S) + R_{\text{FN},1}(h_S) + R_{\text{FN},2}(h_S)$, where

$$\begin{aligned}R_{\text{FP}}(h_S) &= \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[x \in h_S, x \notin I], \\ R_{\text{FN},1}(h_S) &= \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[x \notin h_S, x \in [a, b]], \\ R_{\text{FN},2}(h_S) &= \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[x \notin h_S, x \in [c, d]].\end{aligned}$$

Since we need to have that at least one of $R_{\text{FP}}(h_S)$, $R_{\text{FN},1}(h_S)$, $R_{\text{FN},2}(h_S)$ exceeds $\epsilon/3$ in order for $R(h_S) > \epsilon$, by union bound

$$\begin{aligned}\mathbb{P}(R(h_S) > \epsilon) &\leq \mathbb{P}(R_{\text{FP}}(h_S) > \epsilon/3 \text{ or } R_{\text{FN}(h_S),1} > \epsilon/3 \text{ or } R_{\text{FN}(h_S),2} > \epsilon/3)\\ &\leq \mathbb{P}(R_{\text{FP}}(h_S) > \epsilon/3) + \sum_{i=1}^{2} \mathbb{P}(R_{\text{FN}(h_S),i} > \epsilon/3) \qquad \text{(E.40)}\end{aligned}$$

We first bound $\mathbb{P}(R_{\text{FP}}(h_S) > \epsilon/3)$. Note that if $R_{\text{FP}}(h_S) > \epsilon/3$, then $\mathbb{P}((b,c)) > \epsilon/3$ and hence

$$\mathbb{P}(R_{\text{FP}}(h_S) > \epsilon/3) \leq (1-\epsilon/3)^m \leq e^{-m\epsilon/3}.$$

Now we can bound $\mathbb{P}(R_{\text{FN}(h_S),i} > \epsilon/3)$ by $2e^{-m\epsilon/6}$ using the same argument as in the previous question. Therefore,

$$\mathbb{P}(R(h_S) > \epsilon) \leq e^{-m\epsilon/3} + 4e^{-m\epsilon/6} \leq 5e^{-m\epsilon/6}.$$

Setting, the right-hand side to δ and solving for m yields that $m \geq \frac{6}{\epsilon}\log\frac{5}{\delta}$.

The second case that we need to consider is when $I = [a,d]$, that is, $[a,b] \cap [c,d] \neq \emptyset$. In that case, our algorithm reduces to the one from exercise 2.8 and it was already shown that only $m \geq \frac{2}{\epsilon}\log\frac{2}{\delta}$ samples is required to learn this concept. Therefore, we conclude that our algorithm is indeed a PAC-learning algorithm.

Extension of this result to the case of $\mathcal{C}_p$ is straightforward. The only difference is that in (E.40), one has two summations for $p-1$ regions of false positives and $2p$ regions of false negatives. In that case sample complexity is $m \geq \frac{2(2p-1)}{\epsilon}\log\frac{3p-1}{\delta}$.

Sorting step of our algorithm takes $O(m\log m$ time and steps (b) and (c) are linear in m, which leads to overall time complexity $O(m\log m)$.

2.10 Consistent hypotheses

Since PAC-learning with L is possible for any distribution, let $\mathcal{D}$ be the uniform distribution over $\mathcal{Z}$. Note that, in that case, the cost of an error of a hypothesis h on any point $z \in \mathcal{Z}$ is $\mathbb{P}_{\mathcal{D}}[z] = 1/m$. Thus, if $R_{\mathcal{D}}(h) < 1/m$, we must have $R_{\mathcal{D}}(h) = 0$ and h is consistent. Thus, choose $\epsilon = 1/(m+1)$. Then, for any $\delta > 0$, with probability at least $1-\delta$ over samples S with $|S| \geq P((m+1), 1/\delta)$ points (where P is some fixed polynomial) the hypothesis h_S returned by L is consistent with Z since $R_{\mathcal{D}}(h_S) \leq 1/(m+1)$.

2.11 Senate laws

(a) The true error in the consistent case is bounded as follows:

$$R_{\mathcal{D}}(h) \leq \frac{1}{m}(\log|\mathcal{H}| + \log\frac{1}{\delta}). \qquad \text{(E.41)}$$

For $\delta = .05$, $m = 200$ and $|\mathcal{H}| = 2800$, $R_{\mathcal{D}}(h) \leq 5.5\%$.

(b) The true error in the inconsistent case is bounded as:

$$R_{\mathcal{D}}(h) \leq \widehat{R}_{\mathcal{D}}(h) + \sqrt{\frac{1}{2m}(\log 2|\mathcal{H}| + \log\frac{1}{\delta})}. \qquad \text{(E.42)}$$

For $\delta = .05$, $\widehat{R}_{\mathcal{D}}(h) = m'/m = .1$, $m = 200$ and $|\mathcal{H}| = 2800$, $R_{\mathcal{D}}(h) \leq 27.05\%$.

2.12 Bayesian bound. For any fixed $h \in \mathcal{H}$, by Hoeffding's inequality, for any $\delta > 0$,

$$\mathbb{P}\left[R(h) - \widehat{R}_S(h) \geq \sqrt{\frac{\log\frac{1}{p(h)\delta}}{2m}}\right] \leq p(h)\delta. \qquad \text{(E.43)}$$