Theory of Machine Learning: Homework 1

1. 2.1 Given a training set $S = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^m$ with $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$, we wish to show that there exists a polynomial p_S s.t. $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$ where

$$h_S(\mathbf{x}) = \begin{cases} f(\mathbf{x}_i) & \text{if } \exists i \in [m] \text{ s.t. } \mathbf{x}_i = \mathbf{x} \\ 0 & \text{else} \end{cases}$$

Proof. A polynomial, which is continuous, cannot be strictly positive at isolated points. This follows directly from continuity. Therefore, any polynomial that is nonnegative exactly at the positive samples and negative elsewhere must take value 0 at those samples.

We can begin to design such a polynomial using the Euclidean norm:

$$p_S(\mathbf{x}) = - \prod_{i \in [m], f(\mathbf{x}_i)=1} \|\mathbf{x} - \mathbf{x}_i\|^2$$

That is, p_S takes the negated product of the squared norms of the differences between $\mathbf{x}$ and all positive samples in the dataset. Note that the term

$$\|\mathbf{x} - \mathbf{x}_i\|^2 = \sum_{j \in [d]} (\mathbf{x}_j - \mathbf{x}_{i,j})^2$$

is polynomial in $\mathbf{x}$, and a product of these sums will also produce a polynomial, as is the case for p_S.

There are 3 cases for any sample $\mathbf{x}$:

Case 1. $\mathbf{x} \notin S$

By definition, $h_S(\mathbf{x}) = 0$. It is also always true that $p_S(\mathbf{x}) < 0$. Given $\mathbf{x} \notin S$, it must be that $\|\mathbf{x} - \mathbf{x}_i\|^2 > 0$ for all $i \in [m]$, since equivalence with zero would indicate a match of $\mathbf{x}$ with any sample in the dataset and produce a contradiction. Therefore, $p_S(\mathbf{x}) < 0$ via the negated product. Under this case, we have that $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$.

Case 2. $\mathbf{x} \in S$, $f(\mathbf{x}) = 0$

By definition, $h_S(\mathbf{x}) = 0$. It will also always be true that $p_S(\mathbf{x}) < 0$. Given $f(\mathbf{x}) = 0$, and the product iterates only over positive samples from the dataset, it must be that the final product is positive. A zero product would indicate equivalence of $\mathbf{x}_i$ with any positive sample, which would be a contradiction. Therefore, $p_S(\mathbf{x}) < 0$ after negating the product. Under this case, we have that $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$.

Case 3. $\mathbf{x} \in S$, $f(\mathbf{x}) = 1$

By definition, $h_S(\mathbf{x}) = 1$. It will also always be true that $p_S(\mathbf{x}) = 0$. Given $f(\mathbf{x}) = 1$, and the product iterates only over positive samples from the dataset, it must be that the final product is zero, since at least one of the $\|\mathbf{x} - \mathbf{x}_i\|^2$ must produce a zero. That is, for at least one $i \in [m]$, it must be that $\mathbf{x} = \mathbf{x}_i \implies p_S(\mathbf{x}) = 0$, since $\mathbf{x}$ is a positive sample from the dataset. Under this case, we have that $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$.

So, under all cases, we see that $h_S(\mathbf{x}) = 1 \iff p_S(\mathbf{x}) \geq 0$. QED

2. 2.3 An axis aligned rectangle classifier in the plane assigns the value 1 to a point iff it is inside a rectangle defined by two pairs of real numbers $a_1 \leq b_1$ and $a_2 \leq b_2$:

$$h_{(a_1,b_1,a_2,b_2)}(x_1, x_2) = \begin{cases} 1 & a_1 \leq x_1 \leq b_1,\ a_2 \leq x_2 \leq b_2 \\ 0 & \text{else} \end{cases}$$

Denote the class of all such classifiers as $\mathcal{H}^2_{\text{rec}}$. For the following problems, realizability is assumed.

(a) We wish to show that A is an ERM, where A returns the smallest rectangle enclosing all positive samples.

Proof. Under realizability, there is a perfect rectangle $R^\star$ generating an oracle f s.t. $f(\mathbf{x}) = 1 \iff \mathbf{x} \in R^\star$. So, $L_S(f) = 0$, meaning that there is a hypothesis in $\mathcal{H}^2_{\text{rec}}$ achieving zero empirical risk. A hypothesis produced by an ERM would therefore have an empirical risk of zero. For $h_S \in \arg\min_{h \in \mathcal{H}^2_{\text{rec}}} L_S(h)$, it must be that h_S classifies all positive points in S as 1 and all negative points as 0. Therefore, the rectangle must enclose all positive points from S and not contain any negative points.

Consider the algorithm A. By definition, it encloses all positive points and cannot be shrunk any further without excluding a positive sample. The rectangle $R(S) = A(S)$ cannot "accidentally" contain a negative sample. If it did, it must occur outside $R^\star$ in some dimension so as to not violate realizability. But for $R(S)$ to extend beyond $R^\star$ in any dimension, there would need to be a positive sample in S that is outside $R^\star$. This contradicts realizability, so it follows instead that $R(S) \subseteq R^\star$ and that $R(S)$ cannot contain negative points. Thus, the algorithm A produces a hypothesis h_S with zero empirical error: $L_S(h_S) = 0$. So A is an ERM. QED

(b) We wish to show that if A receives a training set of size $m \geq \frac{4\ln(4/\delta)}{\epsilon}$, then, with probability at least $1-\delta$, it returns a hypothesis with error no greater than ϵ: $L_{(D,f)}(h_S) \leq \epsilon$.

Proof. Under realizability, we are granted the existence of an oracle $f \in \mathcal{H}^2_{\text{rec}}$ s.t. $L_{(D,f)}(f) = 0$. It is also given that $L_S(f) = 0$. We can denote the rectangle that generates this oracle as $R^\star = \text{rec}(a_1^\star, b_1^\star, a_2^\star, b_2^\star)$. So, for any sample $\mathbf{x}$, it must be that $\mathbf{x}$ contained in $R^\star \iff f(\mathbf{x}) = 1$.

Our algorithm A produces the smallest rectangle, $R(S)$, that encloses all positive (label-1) samples from S (with corresponding hypothesis h_S). Suppose $R(S)$ is itself not necessarily fully enclosed by $R^\star$. So, there could be some $\mathbf{x} \in \mathcal{X}$ s.t. $f(\mathbf{x}) = 1$, $\mathbf{x} \in R(S)$, and $\mathbf{x} \notin R^\star$. This produces a contradiction since we established earlier that $f(\mathbf{x}) = 1 \implies \mathbf{x} \in R^\star$. It therefore must be that $R(S) \subseteq R^\star$.

It follows that the only source of prediction error in h_S is in positive points that are outside $R(S)$ but inside $R^\star$. Assuming samples are distributed according to some continuous distribution D, we can characterize the true risk of h_S:

$$\begin{aligned} L_{(D,f)}(h_S) &= \Pr_{\mathbf{x}\sim D}[h_S(\mathbf{x}) \neq f(\mathbf{x})] \\ &= \Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R^\star \setminus R(S)] \end{aligned}$$

Fix some $\epsilon > 0$. The equivalence above implies that if the probability mass of the region $R^\star \setminus R(S)$ is less than ϵ, we are done, regardless of any new sample drawn from D. This is necessarily true if the mass of $R^\star$ alone is less than ϵ. So, we can assume from this point on that: $\Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R^\star] \geq \epsilon$.

At this point, construct four rectangles R_i for $i \in [4]$. We pick $R_1 = \text{rec}(a_1^\star, a_1, a_2^\star, b_2^\star)$ s.t. $a_1^\star \leq a_1$ and

$$\Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R_1] = \epsilon/4.$$

Similarly, we pick b_1, a_2, b_2 such that the probability masses of $R_2 = \text{rec}(b_1, b_1^\star, a_2^\star, b_2^\star)$, $R_3 = \text{rec}(a_1^\star, b_1^\star, a_2^\star, a_2)$, and $R_4 = \text{rec}(a_1^\star, b_1^\star, b_2, b_2^\star)$ are also all $\epsilon/4$. We can try to justify the existence of these rectangles. We know D permits a density p s.t. $\Pr_{\mathbf{x}\sim D}(\mathbf{x} \in A) = \int_A p(\mathbf{x})\, dA$. In the case of two dimensions, we can let the probability mass of a rectangle with parameters a, b, c, t be

$$F(t) = \int_a^t \int_b^c p(\mathbf{x})\, dA$$

We do not expect that the order of integration matters, so we can choose any of the four parameters of a rectangle to be "floating", i.e., the argument t to F. Pick an arbitrary sequence $t_n \to t$. We expect that

$$\begin{aligned} \lim_{n\to\infty}[F(t) - F(t_n)] &= \lim_{n\to\infty}\left[\int_a^t \int_b^c p(\mathbf{x})\, dA - \int_a^{t_n} \int_b^c p(\mathbf{x})\, dA\right] \\ F(t) - \lim_{n\to\infty}[F(t_n)] &= \lim_{n\to\infty}\left[\int_{t_n}^t \int_b^c p(\mathbf{x})\, dA\right] \\ F(t) - \lim_{n\to\infty}[F(t_n)] &= \lim_{n\to\infty}\left[\int_{t_n}^t \int_b^c p(\mathbf{x})\, dA\right] \\ F(t) - \lim_{n\to\infty}[F(t_n)] &= 0 \\ F(t) &= \lim_{n\to\infty}[F(t_n)] \end{aligned}$$

Which suggests that F is continuous in t. Pick a, b, c, t to be the parameters of $R^\star$, with any arbitrary parameter chosen to be t. If we leave t unchanged, $F(t) \geq \epsilon$. If we shrink the rectangle such that it becomes a line, $t = a$ and $F(t) = 0$. At this point, the IVT suggests that there is a choice of t which will produce $F(t) = \epsilon/4$. We can apply this logic to each R_i, swapping the integration order above as needed depending on the parameter chosen to be floating.

Suppose first that all R_i contain positive samples from S. Then it must also be that $R(S)$ intersects all R_i, i.e. the

bounds of $R(S)$ overlap with the bounds of all R_i. So all R_i cover the region of error, $R^\star \setminus R(S)$. Therefore,

$$\begin{aligned}
L_{(D,f)}(h_S) &= \Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R^\star \setminus R(S)] \\
&\leq \Pr_{\mathbf{x}\sim D}\left[\mathbf{x} \in \bigcup_{i\in[4]} R_i\right] \\
&\leq \sum_{i\in[4]} \Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R_i], \text{ via the union bound} \\
&= \sum_{i\in[4]} (\epsilon/4) = \epsilon
\end{aligned}$$

So, if positive samples from S occur in all R_i, we have that $L_{(D,f)}(h_S) \leq \epsilon$.

By contraposition, we expect (for $|S| = m$):

$$\begin{aligned}
\Pr[L_{(D,f)}(h_S) > \epsilon] &= \Pr_{S\sim D^m}\left[\bigcup_{i\in[4]} \{S \cap R_i = \emptyset\}\right] \\
&\leq \sum_{i\in[4]} \Pr_{S\sim D^m}[S \cap R_i = \emptyset] \\
&\leq \sum_{i\in[4]} (1-\epsilon/4)^m, \text{ by the iid assumption} \\
&= 4(1-\epsilon/4)^m \\
&\leq 4(e^{-\epsilon/4})^m, \text{ by Taylor's remainder theorem} \\
&= 4e^{-m\epsilon/4} \leq \delta, \text{ for some } \delta > 0
\end{aligned}$$

We can justify the final inequality via Taylor's remainder theorem (for $x \geq 0$):

$$e^{-x} = 1 - x + \frac{e^{-\xi}}{2}x^2, \xi \in (0, x)$$

Since $e^{-\xi} > 0$, we expect $\frac{e^{-\xi}}{2}x^2 \geq 0$. It follows that $1 - x \leq e^{-x}$. Isolating for m, we get:

$$\begin{aligned}
4e^{-m\epsilon/4} &\leq \delta \\
-m\epsilon/4 &\leq \ln(\delta/4) \\
m &\geq \frac{4\ln(4/\delta)}{\epsilon}
\end{aligned}$$

This gives us a threshold on the sample complexity. We expect that, if $m \geq \frac{4\ln(4/\delta)}{\epsilon}$, then

$$\begin{aligned}
\Pr[L_{(D,f)}(h_S) > \epsilon] &\leq \delta \\
-\delta &\leq -\Pr[L_{(D,f)}(h_S) > \epsilon] \\
1 - \delta &\leq 1 - \Pr[L_{(D,f)}(h_S) > \epsilon] = \Pr[L_{(D,f)}(h_S) \leq \epsilon]
\end{aligned}$$

So $\Pr[L_{(D,f)}(h_S) \leq \epsilon] \geq 1 - \delta$, for some $\epsilon > 0$, $\delta > 0$. QED

(c) We can outline how the proof from part (b) can be extended to $\mathbf{x} \in \mathbb{R}^d$. For $j \in [d]$, we now have d critera defining a rectangle: $\mathbf{x}$ is contained in a rectangle if $a_j \leq x_j \leq b_j$. The algorithm A operates nearly identically, except that the resulting rectangle must contain all positive samples along all d dimensions. Most terms used here refer to the d-dimensional analogs of their 2D counterparts in the previous proof.

Proof. Like in the two-dimensional case, we expect by contradiction that $R(S) \subseteq R^\star$. This is true again because a positive sample in $R(S)$ but not $R^\star$ would break the assumption that $R^\star$ is perfect, which follows from the realizability assumption.

It follows that sources of error lie in regions outside of $R(S)$ but inside $R^\star$ along each of the d axes: $L_{(D,f)}(h_S) = Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R^\star \setminus R(S)]$. Now, if the probability mass of the region $R^\star \setminus R(S) < \epsilon$, then PAC learnability is shown for any sample from D. So, we can assume that the mass of $R^\star \setminus R(S)$ is at least ϵ, which requires that the mass of $R^\star \geq \epsilon$.

The proof diverges when we attempt to bound the error of $R(S)$. Given $R^\star$, for each dimension, there is an upper and lower bound: $a_j^\star, b_j^\star$. For each $j \in [d]$, and for each lower and upper bound, we create a new rectangle R_i where all parameters match those of $R^\star$ except for the selected bound, which may be modified such that the resulting R_i has shrunken. For instance, R_1 may replace $a_1^\star$ by some real number $b_1 \leq b_1^\star$. R_2 may replace $b_1^\star$ by $a_1 \geq a_1^\star$, and so on for subsequent dimensions. The replacement parameters are selected such that $\Pr_{\mathbf{x}\sim D}[\mathbf{x} \in R_i] = \epsilon/(2d)$. We reason that rectangles satisfying this construction exist according to nearly the same reasoning as in the 2D case. Note that the 2D argument used a double-integral for $F(t)$. In d dimensions, $F(t)$ will be defined using d nested integrals for each dimension. For any limit of any integral, we can once again select it to be "floating" and reorder the integral to pull its parent integral to the outside of the expression. Continuity of $F(t)$ follows from essentially the same logic as before, and the IVT establishes the existence of a t producing a probability mass of $\epsilon/(2d)$.

This process produces $2d$ total rectangles R_i. Under the assumption that the union of all R_i cover the error region $R^\star \setminus R(S)$, we expect an error of at most ϵ via the union bound over all R_i, like before. The probability that all R_i will not cover the error region follows again from contraposition:

$$\begin{aligned}
\Pr[L_{(D,f)}(h_S) > \epsilon] &= \Pr_{S\sim D^m}\left[\bigcup_{i\in[2d]} \{S \cap R_i = \emptyset\}\right] \\
&\leq \sum_{i\in[2d]} \Pr_{S\sim D^m}[S \cap R_i = \emptyset] \\
&= \sum_{i\in[2d]} (1 - \epsilon/(2d))^m, \text{ by the iid assumption} \\
&= 2d(1 - \epsilon/(2d))^m \\
&\leq 2de^{-m\epsilon/(2d)}, \text{ since } 1 - x \leq e^{-x} \\
&\leq \delta, \text{ for some } \delta > 0
\end{aligned}$$

This gives us the threshold for the sample complexity: $m \geq \frac{2d}{\epsilon}\ln(2d/\delta)$. With m samples constrained by this inequality, we can expect that $\Pr[L_{(D,f)}(h_S) \leq \epsilon] \geq 1 - \delta$, for some $\epsilon > 0$, $\delta > 0$.

QED

(d) We can outline how the runtime of A is bounded by a polynomial in d, $1/\epsilon$, and $\ln(1/\delta)$. Recall that A produces the smallest rectangle enclosing all positive examples from the training set. This requires iterating over each sample in a dataset of m samples. For each sample, assuming it is labeled as positive, we may need to expand our rectangle $R(S)$ in up to d dimensions to accomodate the new positive sample. This results in roughly $O(md)$ operations. We established previously that $m \geq \frac{2d}{\epsilon}\ln(2d/\delta)$, so we can expect a runtime of roughly $O(\frac{d^2}{\epsilon}\ln(2d/\delta))$, which is $O(\frac{d^2}{\epsilon}(\ln(d) + \ln(1/\delta)))$. This is already polynomial in $1/\epsilon$ and $\ln(1/\delta)$. The $d^2\ln(d)$ term can be bounded by a polynomial in d: $d^2\ln(d) = O(d^3)$. So we expect a runtime of $O(\frac{d^3}{\epsilon}\ln(1/\delta))$, which will be polynomial in d, $1/\epsilon$, and $\ln(1/\delta)$.