Jared Yu
Assignment 1

1. **Overfitting of polynomial matching:** We have shown that the predictor defined in Equation (2.3) leads to overfitting. While this predictor seems to be very unnatural, the goal of this exercise is to show that it can be described as a thresholded polynomial. That is, show that given a training set $S = \{(\mathbf{x}_i, f(\mathbf{x}_i))\}_{i=1}^{m} \subseteq (\mathbb{R}^d \times \{0,1\})^m$, there exists a polynomial p_S such that $h_S(\mathbf{x}) = 1$ if and only if $p_S(\mathbf{x}) \geq 0$, where h_S is as defined in Equation (2.3). It follows that learning the class of all thresholded polynomials using the ERM rule may lead to overfitting.

Ans: (used reference [1], [7])

The equation 2.3 for h_S will first be restated as follows,

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

A slight change has been made to change x and x_i to $\mathbf{x}$ and $\mathbf{x}_i$ respectively. However, the meaning of the function is the same in that it evaluates to 1 if the input example matches an example from the training set and 0 otherwise.

The question is asking us to prove an if and only if statement. Therefore, it must be shown that if $h_S(\mathbf{x}) = 1$, then $p_S(\mathbf{x}) \geq 0$. Also, if $p_S(\mathbf{x}) \geq 0$, then $h_S(\mathbf{x}) = 1$. Starting with if $h_S(\mathbf{x}) = 1$, then $p_S(\mathbf{x}) \geq 0$, the function h_S evaluates to 1 in any situation where an input example exactly matches an example from the training set S. Therefore, if an input example exactly matches an example from the training set S, then it must be shown that $p_S(\mathbf{x}) \geq 0$. Let

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

In p_S, anytime an input $\mathbf{x}$ is given, then the function p_S calculates the product of $\|\mathbf{x} - \mathbf{x}_i\|^{2y_i}$ for all examples in the training set S. If the training set S contains examples where $y_i = 0$, then the $i'th$ product will automatically evaluate to 1, based on the power $2y_i$. In situations where $y_i = 1$, then it will calculate the L_2 norm of the difference between the input example and the $i'th$ training example. If the input example $\mathbf{x}$ matches one of those positive examples (i.e., where $y_i = 1$) from the training set, then that value will zero out. This leads to the entire product, p_S, evaluating to 0, or $p_S(\mathbf{x}) = 0$. However, if the input $\mathbf{x}$ does not exactly match any of the training examples, then the L_2 norm will be greater than 0 for all training examples where $y_i = 1$. Therefore, the resulting product p_S will be calculated with no 0's in the equation leading to a positive non-zero value which is multiplied by -1 as seen in the front of the product. The result is that when $\mathbf{x}$ does not match an example from the training set, then $p_S < 0$.

The other direction also holds true, that is, if $p_S(\mathbf{x}) \geq 0$, then $h_S(\mathbf{x}) = 1$. However, it will be first noted that the maximum value for p_S is 0, but the rule still holds. If an input example matches an example from the training set, then $p_S(\mathbf{x}) = 0$ as mentioned above. For those same input examples, because they match a training example, then $h_S(\mathbf{x}) = 1$. This is since h_S as described in Equation 2.3 evaluates to 1 anytime the input matches a training example and 0 otherwise. If the value of $p_S < 0$, then that indicates that the input example does not match an example from the training set and so $h_S = 0$. Therefore, it has been shown that $h_S(\mathbf{x}) = 1$ if and only if $p_S(\mathbf{x}) \geq 0$. ■

2. Let $\mathcal{H}$ be a class of binary classifiers over a domain $\mathcal{X}$. Let $\mathcal{D}$ be an unknown distribution over $\mathcal{X}$, and let f be the target hypothesis in $\mathcal{H}$. Fix some $h \in \mathcal{H}$. Show that the expected value of $L_S(h)$ over the choice of $S|_x$ equals $L_{(\mathcal{D},f)}(h)$, namely,

$$\mathbb{E}_{S|_x \sim \mathcal{D}^m}[L_S(h)] = L_{(\mathcal{D},f)}(h).$$

Ans:

Let the loss function for the sample S be defined as follows,

$$L_S(h) \stackrel{\text{def}}{=} \frac{|\{i \in [m]: h(x_i) \neq y_i\}|}{m} = \frac{\sum_{i=1}^{m} 1_{\{h(x_i) \neq y_i\}}}{m}, \tag{2.1}$$

where $[m] = \{1, \cdots, m\}$ and $1_{\{\cdot\}}$ is an indicator function that evaluates to 1 when the equation inside of the brackets, $\{\cdot\}$ is true and evaluates to 0 otherwise.

Let the loss function for the population be defined as follows,

$$L_{(\mathcal{D},f)}(h) = \frac{|\{i \in [N]: h(x_i) \neq y_i\}|}{N} = \frac{\sum_{i=1}^{N} 1_{\{h(x_i) \neq y_i\}}}{N}, \tag{2.2}$$

where N is the size of the population of observations within domain $\mathcal{X}$, and $1_{\{\cdot\}}$ is the same as described previously.

$$\mathbb{E}_{S|_x \sim \mathcal{D}^m}[L_S(h)] = \frac{E\left[\sum_{i=1}^{m} 1_{\{h(x_i) \neq y_i\}}\right]}{m} \tag{2.3}$$

In equation 2.3, the expectation is applied the formula for $L_S(h)$ described in equation 2.1. For simplicity, also let $E[\cdot] = \mathbb{E}_{S|_x \sim \mathcal{D}^m}[\cdot]$.

$$= \frac{\sum_{i=1}^{m} E\left[1_{\{h(x_i) \neq y_i\}}\right]}{m} \tag{2.4}$$

In equation 2.4, the expectation is moved into the summation.

$$= \frac{\sum_{i=1}^{m} \{[P(h(x_i) \neq y_i)(1)] + [P(h(x_i) = y_i)(0)]\}}{m} = \frac{\sum_{i=1}^{m} P(h(x_i) \neq y_i)}{m} \tag{2.5}$$

In equation 2.5, the expectation of the indicator is expanded to show that it results in the probability of $h(x_i) \neq y_i$, or $P(h(x_i) \neq y_i)$. Looking closer at the formula, we can see the following,

$$P(h(x_i) \neq y_i) = \frac{\sum_{j=1}^{N} 1_{\{h(x_j) \neq y_j\}}}{N}. \tag{2.6}$$

The reason is that the probability of the classifier h being incorrect for some random observation is the same as the sum of the incorrect classifications over the population of N observations. The subscript j is used for clarity in the next step.

$$\cdots = \frac{\sum_{i=1}^{m} \left[\frac{\sum_{j=1}^{N} 1_{\{h(x_j) \neq y_j\}}}{N}\right]}{m} = \frac{m\left[\frac{\sum_{i=1}^{N} 1_{\{h(x_i) \neq y_i\}}}{N}\right]}{m} = \frac{\sum_{i=1}^{N} 1_{\{h(x_i) \neq y_i\}}}{N} = L_{(\mathcal{D},f)}(h) \tag{2.7}$$

In equation (2.7), the result from equation (2.6) is plugged back into equation (2.5). The summation turns into m, since equation (2.6) does not require the subscript i. The m in the numerator and denominator cancel out, leaving just the formula for $L_{(\mathcal{D},f)}(h)$. Therefore, it can be said that $\mathbb{E}_{S|_x \sim \mathcal{D}^m}[L_S(h)] = L_{(\mathcal{D},f)}(h)$. ■

3. **Axis aligned rectangles:** An axis aligned rectangle classifier in the plane is a classifier that assigns the value 1 to a point if and only if it is inside a certain rectangle. Formally, given real numbers $a_1 \leq b_1$, $a_2 \leq b_2$, define the classifier $h_{(a_1,b_1,a_2,b_2)}$ by

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

The class of all axis aligned rectangles in the plane is defined as

$$\mathcal{H}^2_{rec} = \left\{h_{(a_1,b_1,a_2,b_2)} : a_1 \leq b_1, \text{and } a_2 \leq b_2\right\}.$$

Note that this is an infinite size hypothesis class. Throughout this exercise we rely on the realizability assumption.

1. Let A be the algorithm that returns the smallest rectangle enclosing all positive examples in the training set. Show that A is an ERM.

Ans: (used reference [2])

Let S be the training set. The error of the classifier h over the training sample is:

$$L_S(h) \overset{\text{def}}{=} \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m},$$

where $[m] = \{1, \cdots, m\}$. The Empirical Risk Minimization (ERM) is the learning paradigm that comes up with a predictor h that minimizes $L_S(h)$. In $\mathcal{H}^2_{rec}$, all the classifiers are rectangles. Within $\mathcal{H}^2_{rec}$, there are infinite rectangle-shaped classifiers that can fit the sample data in S. If A is an algorithm that returns the smallest rectangle enclosing all positive examples in the training set, then the classifier h_S produced by algorithm A would have a corresponding $L_S(h_S) = 0$, since it would enclose all positive examples of S (by the realizability assumption). This is the minimum possible value for the training error. Therefore, A minimizes $L_S(h)$ and so it is an ERM. ■

2. Show that if A receives a training set of size $\geq \frac{4\log(4/\delta)}{\epsilon}$ then, with probability of at least $1-\delta$ it returns a hypothesis with error of at most ϵ. *Hint:* Fix some distribution $\mathcal{D}$ over $\mathcal{X}$, let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels, and let f be the corresponding hypothesis. Let $a_1 \geq a_1^*$ be a number such that the probability mass (with respect to $\mathcal{D}$) of the rectangle $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\epsilon/4$. Similarly, let b_1, a_2, b_2 be numbers such that the probability masses of the rectangles $R_2 = R(b_1, b_1^*, a_2^*, b_2^*)$, $R_3 = R(a_1^*, b_1^*, a_2^*, a_2)$, $R_4 = R(a_1^*, b_1^*, b_2, b_2^*)$ are all exactly $\epsilon/4$. Let $R(S)$ be the rectangle returned by A. See illustration in Figure 2.2.

Ans: (used reference [3], [4], [5], [6], [8])

(Hint 1): Let $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ be the rectangle that generates the labels. Therefore, all points inside the rectangle generated by R^* are labeled 1 and all points outside are labeled 0. Let $R(S)$ be the rectangle returned by algorithm A. The rectangle of $R(S)$ is such that it is a rectangle R_i from $R_1, \cdots, R_4$ such that the $L_S(h_S)$ is the minimum from amongst the rectangles, where h_S is the hypothesis associated with the rectangle $R(S)$ that leads to a minimum error. Every rectangle R_i has boundaries that are within the area of the rectangle R^*, as defined in the question itself. So, for every possible $\mathbf{x} \in R(S)$, $\mathbf{x} \in R^*$. Therefore, $R(S) \subseteq R^*$.Then $R(S)$ produced by A must be a subset of R^*.

(Hint 2): Let the event $E = \{S|_x : S \cap R_i \neq \emptyset \text{ for } i = 1,2,3,4\}$. In other words, it is the event when for a sample S, each rectangle R_i is non-empty, or that the sample S contains at least one positive example from them.

If there are positive examples in all R_i for $i \in [1, \cdots, 4]$, then the $L_S(h_S) > 0$, where h_S is the hypothesis chosen by algorithm A implementing $ERM_{\mathcal{H}}$. The reason is that if all rectangles have at least one example in them, then no rectangle can have an error of 0, since none of the rectangles can cover all the space of the other rectangles. In such a situation, this implies that if the sample error is nonzero for all hypotheses, it is not an event associated with M, the set of misleading events. To be specific, let $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ and $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$. Furthermore, $\{S|_x : L_{(\mathcal{D},f)}(h) > \epsilon\} \subseteq M$, or in other words, the event of the error being greater than ϵ (a.k.a. a bad hypothesis) is a subset of M. Where M is the set of misleading events and $\mathcal{H}_B$ is the set of bad hypotheses.

Since event E implies that the sample S is not within M, then the current sample S (based on being part of event E) does not contain any bad hypotheses. Therefore, $L_{(\mathcal{D},f)}(h_S) \leq \epsilon$, when event E occurs.

(Hint 3): Let event $F_i = \{S|_x : S \cap R_i = \emptyset\}$. It is when in a sample S, there are no examples from rectangle R_i. The probability mass of each of the R_i for $i = 1, \cdots, 4$ is $\epsilon/4$. Then, the probability that S does not contain an example from R_i is $\left(1 - \frac{\epsilon}{4}\right)^m$. So, probability of event $F_i = \left(1 - \frac{\epsilon}{4}\right)^m$. This can be upper bounded using the following inequality: $1 - \epsilon \leq \exp(-\epsilon)$. Therefore, we can upper bound the probability of F_i with $\exp(-\frac{\epsilon}{4}m)$. To restate the result, $P(F_i) \leq \exp(-\frac{\epsilon}{4}m)$.

(Hint 4): Then, using the union bound, an upper bound can again be created with the following: $P(\bigcup_{i=1}^4 F_i) \leq \sum_{i=1}^4 P(F_i) = 4\exp(-\frac{\epsilon}{4}m)$.

(Conclusion): In the event of F_i, there is no longer the event E. Therefore, there is no longer the requirement that sample S is not an element of the set M of misleading samples. So, when event F_i occurs, it contains the possibility for a misleading sample from within M. Given that the sample S is an element of the set M, we also have the possibility of a hypothesis from $\mathcal{H}_B$. From this, we also have the following:

$$\mathcal{D}^m\left(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}\right) \leq \mathcal{D}^m(M) \leq P(F).$$

From this it follows that $\mathcal{D}^m\left(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}\right) \leq 4\exp(-\frac{\epsilon}{4}m)$.

Let $m \geq \frac{4\log(4/\delta)}{\epsilon}$, where m is the size and the right side of the inequality is as mentioned in the question itself. Then,

$$4\exp(-\frac{\epsilon}{4}m) \leq 4\exp(-\frac{\epsilon}{4} \cdot \frac{4\log\left(\frac{4}{\delta}\right)}{\epsilon}) = \delta.$$

So, it has been shown that $\mathcal{D}^m\left(\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\}\right) \geq \delta$. Therefore, it also follows that $\mathcal{D}^m\left(\{S|_x : L_{(\mathcal{D},f)}(h_S) \leq \epsilon\}\right) \leq 1 - \delta$.■

3. Repeat the previous question for the class of axis aligned rectangles in $\mathbb{R}^d$.

Ans: (used reference [4])

Previously, the question showed that:

$$h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) = \begin{cases} 1 & \text{if } a_1 \le x_1 \le b_1 \text{ and } a_2 \le x_2 \le b_2 \\ 0 & \text{otherwise} \end{cases}$$

with

$$\mathcal{H}^2_{rec} = \left\{h_{(a_1,b_1,a_2,b_2)} : a_1 \le b_1, \text{and } a_2 \le b_2\right\}.$$

Changing the problem to $\mathbb{R}^d$, it becomes:

$$h_{(a_1,b_1,\cdots,a_d,b_d)}(x_1,\cdots,x_d) = \begin{cases} 1 & \text{if } a_i \le x_i \le b_i \text{ for all } i \in [1,\cdots,d] \\ 0 & \text{otherwise} \end{cases}$$

with

$$\mathcal{H}^d_{rec} = \left\{h_{(a_1,b_1,\cdots,a_d,b_d)} : a_i \le b_i \text{ for all } i \in [1,\cdots,d]\right\}.$$

Also, change the training set to be of size $\ge \frac{2d\log(2d/\delta)}{\epsilon}$, where d corresponds with the number of dimensions. The rectangle R^* would change to $R^* = R(a_1^*, b_1^*, \cdots, a_d^*, b_d^*)$. The rectangles R_i, for $i = 1,\cdots,2d$, would have a similar design as in question 3.2, except that they would include the higher dimensions. The probability mass for each of the R_i would be exactly $\frac{\epsilon}{2d}$.

(Hint 1): Let $R^* = R(a_1^*, b_1^*, \cdots, a_d^*, b_d^*)$ be the rectangle that generates the labels. Therefore, all points inside the rectangle generated by R^* are labeled 1 and all points outside are labeled 0. Let $R(S)$ be the rectangle returned by algorithm A. The rectangle chosen by $R(S)$ is such that it is a rectangle R_i from $R_1,\cdots,R_{2d}$ such that the $L_S(h_S)$ is the minimum from amongst the rectangles, where h_S is the hypothesis associated with the rectangle $R(S)$ that leads to a minimum error. Every rectangle R_i has boundaries that are within the area of the rectangle R^*, as defined in the question itself. So, for every possible $\mathbf{x} \in R(S)$, $\mathbf{x} \in R^*$. Therefore, $R(S) \subseteq R^*$.Then $R(S)$ produced by A must be a subset of R^*.

(Hint 2): Let the event $E = \{S|_x : S \cap R_i \ne \emptyset \text{ for } i = 1,\cdots,2d\}$. In other words, it is the event when for a sample S, each rectangle R_i is non-empty, or that the sample S contains at least one positive example from them.

If there are positive examples in all R_i for $i \in [1,\cdots,2d]$, then the $L_S(h_S) > 0$, where h_S is the hypothesis chosen by algorithm A implementing $ERM_{\mathcal{H}}$. The reason is that if all rectangles have at least one example in them, then no rectangle can have an error of 0, since none of the rectangles can cover all the space of the other rectangles. In such a situation, this implies that if the sample error is nonzero for all hypotheses, it is not an event associated with M, the set of misleading events. To be specific, let $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\}$ and $\mathcal{H}_B = \left\{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\right\}$. Furthermore, $\left\{S|_x : L_{(\mathcal{D},f)}(h) > \epsilon\right\} \subseteq M$, or in other words, the event of the error being greater than ϵ (a.k.a. a bad hypothesis) is a subset of M. Where M is the set of misleading events and $\mathcal{H}_B$ is the set of bad hypotheses.

Since event E implies that the sample S is not within M, then the current sample S (based on being part of event E) does not contain any bad hypotheses. Therefore, $L_{(\mathcal{D},f)}(h_S) \le \epsilon$ when event E occurs.

(Hint 3): Let event $F_i = \{S|_x : S \cap R_i = \emptyset\}$. It is when in a sample S, there are no examples from rectangle R_i. The probability mass of each of the R_i for $i = 1, \cdots, 2d$ is $\frac{\epsilon}{2d}$. Then, the probability that S does not contain an example from R_i is $\left(1 - \frac{\epsilon}{2d}\right)^m$. So, probability of event $F_i = \left(1 - \frac{\epsilon}{2d}\right)^m$. This can be upper bounded using the following inequality: $1 - \epsilon \leq \exp(-\epsilon)$. Therefore, we can upper bound the probability of F_i with $\exp(-\frac{\epsilon}{2d}m)$. To restate the result, $P(F_i) \leq \exp(-\frac{\epsilon}{2d}m)$.

(Hint 4): Then, using the union bound, an upper bound can again be created with the following: $P\left(\bigcup_{i=1}^{2d} F_i\right) \leq \sum_{i=1}^{2d} P(F_i) = 2d \exp(-\frac{\epsilon}{2d}m)$.

(Conclusion): In the event of F_i, there is no longer the event E. Therefore, there is no longer the requirement that sample S is not an element of the set M of misleading samples. So, when event F_i occurs, it contains the possibility for a misleading sample from within M. Given that the sample S is an element of the set M, we also have the possibility of a hypothesis from $\mathcal{H}_B$. From this, we also have the following:

$$\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \leq \mathcal{D}^m(M) \leq P(F).$$

From this it follows that $\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \leq 2d \exp(-\frac{\epsilon}{2d}m)$.

Let $m \geq \frac{2d \log(2d/\delta)}{\epsilon}$, where m is the size and the right side of the inequality is as mentioned in the question. Then,

$$2d \exp(-\frac{\epsilon}{2d}m) \leq 2d \exp(-\frac{\epsilon}{2d} \cdot \frac{2d \log\left(\frac{2d}{\delta}\right)}{\epsilon}) = \delta.$$

So, it has been shown that $\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) > \epsilon\right\}\right) \geq \delta$. Therefore, it also follows that $\mathcal{D}^m\left(\left\{S|_x : L_{(\mathcal{D},f)}(h_S) \leq \epsilon\right\}\right) \leq 1 - \delta$.■

4. Show that the runtime of applying the algorithm A mentioned earlier is polynomial in d, $1/\epsilon$, and in $\log(1/\delta)$.

Ans: (used reference [4])
The way that algorithm A works is that it needs to scan each of the d dimensions and analyze all the m points in these dimensions to find the correct a_i or b_i values to help determine the R_i such that it has a probability mass of exactly $\frac{\epsilon}{2d}$. Doing so implies that the runtime is $O(md)$, where m is the number of examples in the training set and d is the number of dimensions. Furthermore, the size of the training set m has been upper bounded by $\frac{2d \log(2d/\delta)}{\epsilon}$. Therefore, taking the product of $\frac{2d \log(2d/\delta)}{\epsilon}$ and d leads to a polynomial in d, $\frac{1}{\epsilon}$, and $\log 1/\delta$.

References

[1] http://www.math.ubc.ca/~elyse/220/2016/7Nonconditional.pdf
Used to double-check how to prove an if-and-only-if statement.
[2] http://www.people.vcu.edu/~rhammack/Math501/3ways.pdf
Used to double-check how to prove an if-then statement.

[3] https://www.sciencedirect.com/topics/computer-science/probability-mass
Used for understanding the context of the term “probability mass” in problem 3.
[4] https://www.cs.huji.ac.il/~shais/UnderstandingMachineLearning/MLbookSol.pdf
Found from Piazza and used as a guidance for problem 3.
[5] http://karlstratos.com/notes/pac.pdf
Used as a guidance for problem 3.
[6] https://piazza.com/class/kc0jkwru805u1?cid=38
Question on Piazza about probability mass.
[7] https://piazza.com/class/kc0jkwru805u1?cid=37
Question on Piazza about problem 1.
[8] https://piazza.com/class/kc0jkwru805u1?cid=17
Comment on Piazza about resources for the textbook.