Assignment 2

Applied Machine Learning

Fatemeh Saberi Khomami | 400422114 | fatima.saberi@gmail.com

Question 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 $S = \{(x_i, f(x_i))\}_{i=1}^{m} \subseteq (\mathbb{R}^d \times \{0,1\})^m$ there exists a polynomial p_s such that $h_s(x) = 1$ if and only if $p_s(x) \geq 0$, 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.

Answer: So we want to demonstrate one example of a threshold polynomial function which overfits the data, and no matter what, for data points outside of the training set, it fails to predict label 1.
The hypothesis h is given as follows:

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

So we only need to come up with an equation that is negative whenever $x_i \neq x$, and is positive or zero whenever $x_i = x$. We propose the below function:

$$p_s(x) = \sum_{i \in [m]\ s.t\ y_i \neq 0} -|x^4 - x_i^4|$$

As you can see, whenever $h_s(x) = 1$ and as a result $x_i = x$, $p_s(s) = 0$, and other times, $p_s(s) < 0$.
We successfully created a $p_s(x)$ which is overfitting on training set S.

Question 2

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

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

Answer:

$$\mathbb{E}_{S\sim D^m}[L_s(h)] = \mathbb{E}_{S\sim D^m}[\frac{1}{m}\sum_{i=1}^{m} 1_{h(x_i)} \neq f(x_i)] \quad \text{we just opened the formula of } L_s(h)$$

$$= \frac{1}{m}\mathbb{E}_{S\sim D^m}[\sum_{i=1}^{m} 1_{h(x_i)} \neq f(x_i)] \quad \text{constants can exit Expectation}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}_{S\sim D^m}[1_{h(x_i)} \neq f(x_i)] \quad \text{due to linearity of Expectation}$$

$$= m \cdot \frac{1}{m}\mathbb{E}_{S\sim D^m}[1_{h(x)\neq f(x)}] \quad \text{because instances are sampled i.i.d}$$

$$= \mathbb{E}_{x\sim D}[1_{h(x)\neq f(x)}] \quad \text{because now there is no dependency to training set S}$$

$$= L_{(D,f)}(h) \quad \text{due to definition}$$

Question 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

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

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

$$H^2_{rec} = \{h_{(a_1,b_1,a_2,b_2)}(x_1,x_2) : a_1 \leq b_1,\ and\ a_2 \leq b_2\}.$$

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

3.1

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

Answer: Note that our Hypothesis class H is rectangular. Since the realizability assumption holds true, we know that there is a rectangle in H in which $L_s(h_s) = 0$. This tells us that the binary class of $\{0,1\}$ are separable by rectangles.
According to sample set S, and learning algorithm A, what $A(S) = h_s$ returns:

1. is the smallest rectangle
2. includes **all** the positive samples. This means that the probability of label something outside of rectangle "positive" is zero. Furthermore, A returns the smallest rectangular region that consists of positive labels; so,everything outside of the rectangle will be correctly labeled negative. This means the error is zero, and we can conclude that $L_s(h_s) = 0$

We know that ERM return the hypothesis that $argmin_{h \in H} L_s(h)$; so, the given $A(S)$ which has $L_s(h_s) = 0$, has the least error; hence, it's an ERM.

3.2

Show that if A receives a training set of $size \geq \frac{4\log(\frac{4}{\delta})}{\epsilon}$, then with probability of at least $1-\delta$ it returns a hypothesis with error of at most ϵ.
Hint: Fix some distribution D over 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 D) of the rectangle $R_1 = R(a_1^*, a_1, a_2^*, b_2^*)$ is exactly $\frac{\epsilon}{4}$. 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 $\frac{\epsilon}{4}$. Let $R(S)$ be the rectangle returned by A.

- show that $R(s) \subseteq R^*$.

Answer: $R(s)$ is the result of $A(s)$ which returned the smallest rectangle. If $R^* = R(a_1^*, b_1^*, a_2^*, b_2^*)$ and $R(s) = R(a_1, b_1, a_2, b_2)$, with respect to what was given in the question, it is safe to say that:

$$\begin{cases} a_1 \geq a_1^* \\ b_1 \leq b_1^* \\ a_2 \geq a_2^* \\ b_2 \leq b_2^* \end{cases} \tag{1}$$

Because $a_1^* \leq a_1$ and $b_1^* \geq b_1$, we can see that the length of R^* which is $b_1^* - a_1^*$ has the longer range compare to $R(s)$. Also, $a_2^* \leq a_2$ and $b_2^* \geq b_2$. It is clear that R^* has the longer range in width as well. So, with longer width and length, R^* includes $R(s)$ and we can say that $R(s) \subseteq R^*$.

- Show that if S contains (positive) examples in all of the rectangles R_1, R_2, R_3, R_4, then the hypothesis returned by A has error of at most ϵ.

- For each $i \in \{1, \ldots, 4\}$, upper bound the probability that S does not contain an example from R_i .

- Use the union bound to conclude the argument.

Answer: To prove that the error is at most ϵ, we'll show that the probability that learner fails is less than δ, that is,

$$\mathbb{P}(L_{(D,f)}(h_s) > \epsilon) < \delta$$

The worst case of learner's failure can be interpreted as "not including R_1, R_2, R_3, R_4 in the learning process", which means their contribution in training set was 0. in other words:

$$\text{We define } B_i \; as : B_i = \{s|_x : \forall i \in [1,2,3,4] \, , \, S \cap R_i = \emptyset\}$$

$$\mathbb{P}^m\Big(\{s|_x : L_{(D,f)}(h_s) > \epsilon\}\Big) \leq \mathbb{P}^m\Big(\bigcup_{i=1}^{4} B_i\Big) \leq \sum_{i=1}^{4} \mathbb{P}^m\Big(B_i\Big)$$

we Know that the probability of falling into each of R_1, R_2, R_3, R_4 rectangles is $\frac{\epsilon}{4}$. So, the probability for one training example to not fall into one of the mentioned classes is $1 - \frac{\epsilon}{4}$. Because the sampling process was carried out i.i.d, the probability for m training examples would be:

$$\begin{aligned} \mathbb{P}^m\Big(B_i\Big) &= (1 - \frac{\epsilon}{4})^m \\ &\leq e^{\frac{-m\epsilon}{4}} \qquad \text{due to Taylor series} \end{aligned}$$

Let's jump back into our proof:

$$\mathbb{P}^m\Big(\{s|_x : L_{(D,f)}(h_s) > \epsilon\}\Big) \leq \sum_{i=1}^{4} e^{\frac{-m\epsilon}{4}} = 4e^{\frac{-m\epsilon}{4}}$$

Now we want the above statement to be less than δ.

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

We just proved that if $m \geq \frac{4\ln(\frac{\delta}{4})}{\epsilon}$, with probability of at least $1-\delta$, $L_{(D,f)}(h_s) < \epsilon$.

3.3

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

Answer: If instead of 2 dimensions, we have d dimensions, we can expand our solution as well. For example, if both R^* and $R(s)$ have $2d$ parameters instead of 4 (for 2-dimensional space), A will still return an area which is the smallest enclosing that comprises all the positive labels; so, $L_s(h_s) = 0$ and A would be ERM.
On the other hand, Instead of 4 R regions, in d dimension, we would have $2d$ regions of R. Hence, the probability mass of each region would be $\frac{\epsilon}{2d}$. We only need to prove $\mathbb{P}(L_{(D,f)}(h_s) > \epsilon) < \delta$.

$$B_i = \{s|_x \,:\, \forall i \in [1,2,\ldots,2d]\,,\, S \cap R_i = \emptyset\}$$

$$\mathbb{P}^m\Big(\{s|_x \,:\, L_{(D,f)}(h_s) > \epsilon\}\Big) \leq \mathbb{P}^m\Big(\bigcup_{i=1}^{2d} B_i\Big) \leq \sum_{i=1}^{2d} \mathbb{P}^m\Big(B_i\Big)$$

We also know that:

$$\begin{aligned} \mathbb{P}^m\Big(B_i\Big) &= (1 - \frac{\epsilon}{2d})^m \\ &\leq e^{\frac{-m\epsilon}{2d}} && \text{due to Taylor series} \end{aligned}$$

Back to proof:

$$\mathbb{P}^m\Big(\{s|_x \,:\, L_{(D,f)}(h_s) > \epsilon\}\Big) \leq \sum_{i=1}^{2d} e^{\frac{-m\epsilon}{2d}} = 2de^{\frac{-m\epsilon}{2d}}$$

We want the above statement to be less than δ.

$$
\begin{aligned}
2de^{\frac{-m\epsilon}{2d}} &\leq \delta \\
e^{\frac{-m\epsilon}{2d}} &\leq \frac{\delta}{2d} \\
-m\epsilon &\leq 2d \ln(\frac{\delta}{2d}) \\
m &\geq \frac{2d \ln(\frac{\delta}{2d})}{\epsilon}
\end{aligned}
$$

We proved that for the class of axis aligned rectangles in $\mathbb{R}^d$, if $m \geq \frac{2d \ln(\frac{\delta}{2d})}{\epsilon}$, with probability of at least $1-\delta$, $L_{(D,f)}(h_s) < \epsilon$.

3.4

Show that the run-time of applying the algorithm A mentioned earlier is polynomial in d, $\frac{1}{\epsilon}$ and in $\log(\frac{1}{\delta})$.

Answer: We showed that the least amount of training examples needed in dimension d to have error at most ϵ with probability of at least $1-\delta$, is $\frac{2d \ln(\frac{2d}{\delta})}{\epsilon}$.
The time complexity or run-time is m times d (dm), because each of our training examples have d dimension. In other words, the run-time function is :

$$
d \times m = d \times \frac{2d \ln(\frac{2d}{\delta})}{\epsilon} = \frac{2d^2 \ln(\frac{2d}{\delta})}{\epsilon}
$$

As you can see, time complexity is a polynomial function of $\frac{1}{\epsilon}$ and $\log(\frac{1}{\delta})$