Introduction to Machine Learning (0368-3235) - Assignment 2
Lior Erenreich

Homework 2:

Theory Questions:

1. **PAC learnability of ℓ_2-balls around the origin**:

(15 points) PAC learnability of ℓ_2-balls around the origin. Given a real number $R \geq 0$ define the hypothesis $h_R : \mathbb{R}^d \to \{0, 1\}$ by,

$$h_R(\mathbf{x}) = \begin{cases} 1 & \|\mathbf{x}\|_2 \leq R \\ 0 & otherwise. \end{cases}$$

Consider the hypothesis class $\mathcal{H}_{ball} = \{h_R \mid R \geq 0\}$. Prove directly (without using the Fundamental Theorem of PAC Learning) that $\mathcal{H}_{ball}$ is PAC learnable in the realizable case (assume for simplicity that the marginal distribution of X is continuous). How does the sample complexity depend on the dimension d? Explain.

To prove that the hypothesis class H_{ball} is PAC learnable in the realizable case, I need to show that given any target function f that can be represented as h_R for some R, and any ε and $\delta > 0$, there exists an algorithm A that with probability at least $1 - \delta$ outputs a hypothesis h with error ε, using a sample size of at most polynomial in $\frac{1}{\varepsilon}, \frac{1}{\delta}$, and d.
Let's consider an algorithm as follows:
Algorithm A:
Given a set of n data points $S = \{x_i\}_{i=1}^n$, outputs $R_S = \max_i \left|\left|x_i\right|\right|_2$.
Proof (H_{ball} is PAC learnable with algorithm A):

First of all, let's note that this question discusses the realizable case, meaning there exists R_X which gets $e_P(h_{R_X}) = 0$. So, we know for every data point x_i for which $h_R(x_i) = 1$, then $\max_i \left|\left|x_i\right|\right|_2 \leq R_X$. Let's note that S is a subset of the X, so $\max_i \left|\left|x_i\right|\right|_2 = R_S \leq R_X$. So a possible error in R_A can be caused only because of a data point in X which gets $h_R(x_i) = 1$ but according to A's output R_S, gets 0, meaning $h_{R_S}(x_i) = 0$. This happens if $\left|\left|x_i\right|\right|_2$ is greater than any data point's in S which was mapped to 1. Assuming (as described in the question) that the marginal distribution of X is continuous:

$$e_P = P\left[R_S \leq \left|\left|x_i\right|\right|_2 \leq R_X\right] = P\left[\left|\left|x_i\right|\right|_2 \leq R_X\right] - P\left[\left|\left|x_i\right|\right|_2 \leq R_S\right]$$

If $P\left[\left|\left|x_i\right|\right|_2 \leq R_X\right] < \epsilon$, we get that $e_P < \epsilon$ because $P\left[\left|\left|x_i\right|\right|_2 \leq R_S\right] \geq 0$ so $e_P \leq P\left[\left|\left|x_i\right|\right|_2 \leq R_X\right]$.

Else, let R_ϵ be the real number for which $P\left[R_\epsilon < \left|\left|x_i\right|\right|_2 \leq R_X\right] = \epsilon$, we get:

$$e_P(h_R) \leq P[R_X \backslash R_S] \leq P[R_\epsilon] = \epsilon$$
$$P[e_P(h_R) > \epsilon] \leq P[\exists x_i \in R_\epsilon]$$
$$P[e_P(h_R) > \epsilon] \leq (1-\epsilon)^n \leq e^{-n\epsilon} = \delta \ \ for\ n \leq \frac{1}{\epsilon}\ln\left(\frac{1}{\delta}\right)$$

Proved that there exists $N(\epsilon, \delta)$ that is not dependent on the dimension d. ■

Introduction to Machine Learning (0368-3235) - Assignment 2
Lior Erenreich

2. PAC in Expectation:

> 2. **(15 points) PAC in Expectation**. Consider learning in the realizable case. We say a hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** using algorithm A if there exists a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P (realizable by $\mathcal{H}$), given a sample set S such that $|S| \geq N(a)$, it holds that,
>
> $$\mathbb{E}\left[e_P(A(S))\right] \leq a.$$
>
> Show that $\mathcal{H}$ is PAC learnable *if and only if* $\mathcal{H}$ is PAC learnable in expectation (Hint: For one direction, use the law of total expectation. For the other direction, use Markov's inequality).

Proof:

First direction (H is PAC learnable in expectation -> H is PAC learnable):

H is PAC learnable in expectation, so there exists an algorithm A for any ϵ and δ such that $P\left[e_P\left(A(s)\right) > \epsilon\right] \leq \delta$ (while $\epsilon > 0$ and $e_P\left(A(s)\right) > 0$).

Using Markov's inequality:

$$P[e_P(h_R) > \epsilon] \leq P[e_P(h_R) \geq \epsilon] \leq \frac{E[e_P(h_R)]}{\epsilon}$$

So for $n \geq N(a)$ let's define N_a such that:

$$\frac{E[e_P(h_R)]}{\epsilon} \leq \frac{N_a}{\epsilon} = \frac{\epsilon\delta}{\epsilon} = \delta$$

We got PAC learnability.

Second direction (H is PAC learnable -> H is PAC learnable in expectation):

H is PAC learnable, so there exist an algorithm A and a function $N(\epsilon, \delta)$ such that for any S which holds $|S| > N(\epsilon, \delta)$ occurs $P\left[e_P\left(A(s)\right) > \epsilon\right] < \delta$, (while $\epsilon > 0$ and $e_P\left(A(s)\right) > 0$). Using the law of total expectation:

$$\begin{aligned} E\left[e_P\left(A(s)\right)\right] &= E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) > \epsilon\right] P\left[e_P\left(A(s)\right) > \epsilon\right] \\ &\quad + E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) \leq \epsilon\right] P\left[e_P\left(A(s)\right) \leq \epsilon\right] \\ &= E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) > \epsilon\right]\delta + E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) \leq \epsilon\right](1-\delta) \end{aligned}$$

Let's note that the error e_P is defined to be the expectation of the l_{o-1} errors, the maximal value of such error is 1:

$$E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) > \epsilon\right] \leq 1$$

Also let's note that $E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) \leq \epsilon\right] \leq \epsilon$

We get:

$$\begin{aligned} E\left[e_P\left(A(s)\right)\right] &= E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) > \epsilon\right]\delta + E\left[e_P\left(A(s)\right) \middle| e_P\left(A(s)\right) \leq \epsilon\right](1-\delta) \\ &\leq \epsilon + \delta \end{aligned}$$

H is PAC learnable in expectation.■

Introduction to Machine Learning (0368-3235) - Assignment 2
Lior Erenreich

3. Union of Intervals:

> **(15 points) Union Of Intervals.** Determine the VC-dimension of $\mathcal{H}_k$ - the subsets of the real line formed by the union of k intervals (see the programming assignment for a formal definition of $\mathcal{H}$). Prove your answer.

Claim: The VC-dimension of H_k is $2k$.

Proof:

a. $VC - dim(H_k) \geq k$:
Base case: for $k = 1$, given 2 points $\{x_1, x_2\}$, WLOG $x_1 < x_2$. I will use the fact that between every two given rational numbers, there are infinite rational numbers, so there exist y_1, y_2 such that:

$$x_1 < y_1 < y_2 < x_2$$
$$h_I(x_1) = h_I(x_2) = 1:\ l_1 = x_1 - 1, \quad u_1 = x_2 + 1$$
$$h_I(x_1) = h_I(x_2) = 0:\ l_1 = y_1, \quad u_1 = y_2$$
$$h_I(x_1) = 1, \quad h_I(x_2) = 0:\ l_1 = x_1 - 1, \quad u_1 = y_2$$
$$h_I(x_1) = 0, \quad h_I(x_2) = 1:\ l_1 = y_1, \quad u_1 = x_2 + 1$$

Induction step: from the induction assumption we get that with k intervals we can get all possible mappings of $2k$ points. So now I need to show that for $k + 1$ intervals, I can get all possible mappings of $2k + 2$ points: so I have 2 points left and one "free" interval, so induction step is equivalent to the base case.

b. $VC - dim(H_k) < 2k + 1$:
Base case: for $k = 1$, given 3 points $\{x_1, x_2, x_3\}$, WLOG $x_1 < x_2 < x_3$. There is no possible way to define l_1, u_1 to get:

$$h_I(x_1) = 1, h_I(x_2) = 0, h_I(x_3) = 1$$

Induction step: from the induction assumption we get that with k intervals we can't get all possible mappings of $2k + 1$ points. So now I need to show that for $k + 1$ intervals, I can't get all possible mappings of $2k + 3$ points: we know that for the "first" $2k + 2$ points I can get all possible mappings using $k + 1$ intervals from a (we can't use only k or less intervals for that from the induction assumption), let's look at this case in which $x_i < x_j, i < j$ and $x_i = 1 - x_{i-1}$:

$$X = \{x_1, x_2, x_3, \dots, x_{2k}, x_{2k+1}, x_{2k+2}\} = \{1,0,1, \dots ,0,1,0\}$$

Adding another point $x_{2k+3} > x_{2k+2}$ can't be mapped to 1 since all $k + 1$ intervals are used already and it can't be inside the interval of existing point (x_{2k+1} for example) since there are points mapped to zero between them (in this example – x_{2k+2}).

Introduction to Machine Learning (0368-3235) - Assignment 2
Lior Erenreich

4. Prediction by Polynomials:

(15 points) Prediction by polynomials. Given a polynomial $P : \mathbb{R} \to \mathbb{R}$ define the hypothesis $h_P : \mathbb{R}^2 \to \{0, 1\}$ by,

$$h_P(x_1, x_2) = \begin{cases} 1 & P(x_1) \geq x_2 \\ 0 & otherwise. \end{cases}$$

Determine the VC-dimension of $\mathcal{H}_{poly} = \{h_P \mid P \text{ is a polynomial}\}$. You can use the fact that given n distinct values $x_1, ..., x_n \in \mathbb{R}$ and $z_1, ..., z_n \in \mathbb{R}$ there exists a polynomial P of degree $n - 1$ such that $P(x_i) = z_i$ for every $1 \leq i \leq n$.

Claim: $VC - \dim(H_{poly}) = \infty$

Proof:

Let's notice that in order to get for any point (x_1, x_2) occurs $P(x_1) \geq x_2$, polynomial P has to pass through a point (x_1, y_1) such that $y_1 > x_2$. So for any $n \in \mathbb{N}$, we can get all possible mappings of $n + 1$ points $\{x_1, \dots, x_n, x_{n+1}\}$, for every point x_i if we want to get a mapping of 1, we will determine $y_i > x_{i2}$, else $y_i < x_{i2}$.
Using the fact that given $n + 1$ distinct values $x_1, \dots x_{n+1} \in \mathbb{R}$ and $y_1, \dots y_{n+1} \in \mathbb{R}$ there exists a polynomial P of degree n such that $P(x_i) = y_i$ for every $1 \leq i \leq n + 1$.

Introduction to Machine Learning (0368-3235) - Assignment 2
Lior Erenreich

Programming Assignment:

a. I'm given that the true distribution

$$P[x,y] = P[y|x] \cdot P[x]$$

Is as follows:

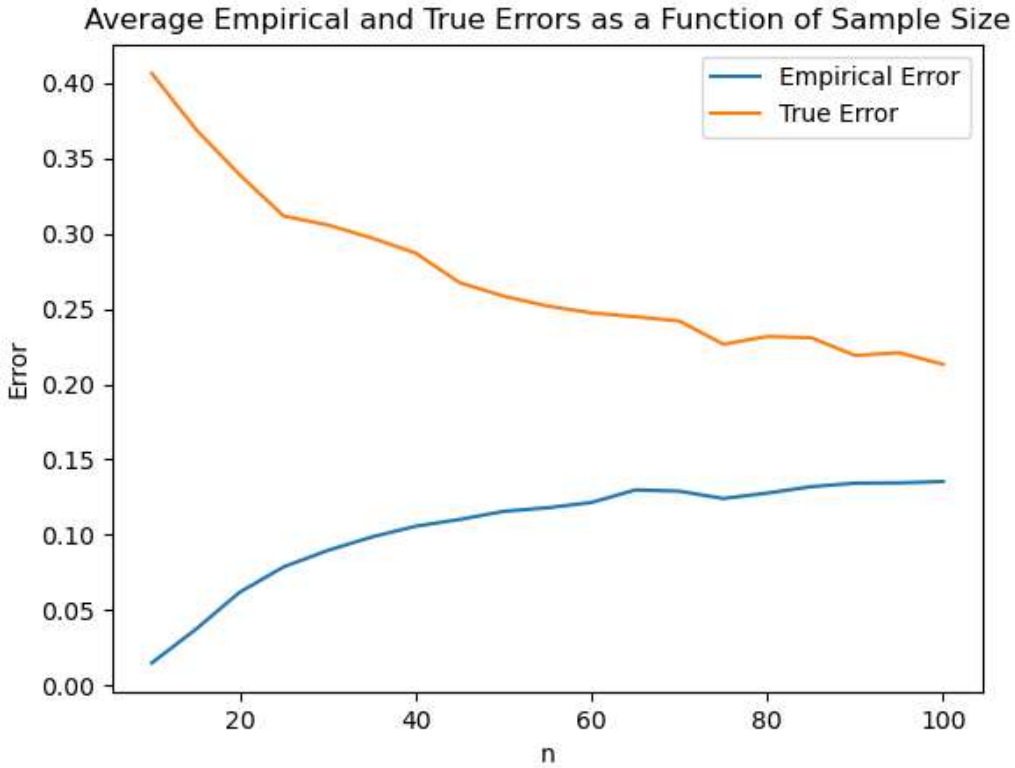
x is distributed uniformly on the interval $[0,1]$, and

$$P[y=1|x] = \begin{cases} 0.8, & if\ x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1] \\ 0.1, & if\ x \in (0.2,0.4) \cup (0.6,0.8) \end{cases}$$

The hypothesis in H_{10} with the smallest error:

$$h(x) = \arg\min_{h \in H_{10}} e_P(h) = \arg\min_{h \in H_{10}} P[Y = y|X = x]$$
$$= \begin{cases} 1, & x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1] \\ 0, & otherwise \end{cases}$$

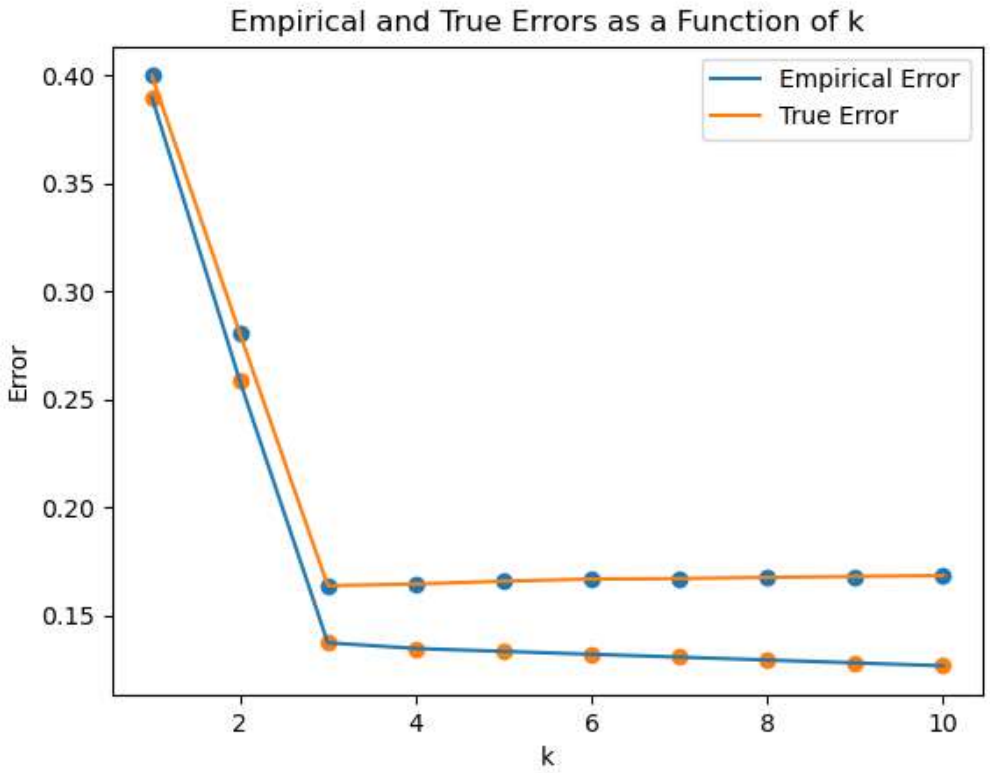
b. The plot:



Looking at the plot, we observe that as n increases, the true error decreases while the empirical error increases. This is because as the sample size increases, the ERM algorithm is more likely to select a hypothesis that fits the sample well but may not generalize well to the true distribution. However, the true error decreases because a larger sample provides a better representation of the true distribution.
Furthermore, we can see that the empirical error approaches a value which seems to be around 0.15. This makes sense since it is the middle value between the false positive and true negative error rates. This suggests that the ERM algorithm is able to learn the

underlying distribution reasonably well, but still has some errors due to the limited sample size.

c. The method call of $ass.experiment\_k\_range\_erm(1500, 1, 10, 1)$ results with 9 as the best value of k according to the ERM algorithm and with the plot:



Looking at the plot, we observe a sharp decrease in both errors until $k = 3$, after which the empirical error continues to decrease gradually while the true error starts increasing gradually. This indicates that for values of k beyond 3, we are overfitting the data and the model is becoming less accurate on new unseen data.
Therefore, the best value for k is 3, since it achieves a good balance between fitting the data and generalizing to new data. The smallest empirical error is achieved with $k = 10$, but choosing k solely based on the empirical error can lead to overfitting and poor performance on new data. Instead, we should aim to minimize the difference between the empirical and true errors, which is achieved by choosing $k = 3$.

d. The method call of $ass.cross\_validation(1500)$ results with 3 as the best value of k according to the cross validation algorithm. As discussed in the previous question, this makes sense since the ERM algorithm found that the best value for k is 9, which minimizes the empirical error on the given data set. However, this does not necessarily mean that the true error of the hypothesis with k=9 is also the lowest. On the other hand, cross-validation estimates the true error of each hypothesis by validating it on a different set of data. In this case, cross-validation found that the best k value is 3, which provides the lowest empirical error on the validation set.