Introduction to Machine Learning

Section 2

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

Given a real number $R \geq 0$ define the hypothesis $h_R : R^d \rightarrow \{0,1\}$ and we will proof that hypothesis class $H_{ball} = \{h_R | R \geq 0\}$ is PAC learn-able in the realizable case.
lets design an algorithm A_{balls} that learns H_{ball}.

- Given a sample of size $N = \{u_1, ..., u_N\}$ lets find the smallest ball B which is consistent with the sample
  i.e $B_R : u_k = MAX\{u_1, ..., u_N\} \wedge ||u_k||_2 \leq R$

  mistake only by labeling positive points as negative.

- The error of the algorithm is $e_P(h_R) = P[B_0 \setminus B_R]$

We assume that $P(B_R) > \epsilon$. otherwise, we stand with the property and finished. now lets define T to be the real boundary of B_0, that "extend" to the direction $\rightarrow$ (0,0).such that for all $\epsilon, P(T) = \epsilon$. since any sample is in the form of $||u_i||_2 = (x_1^2 + x_2^2 ... + x_d^2)^{1/2} \geq 0$ for any $u \in T$, we get

$$e_P(h_R) = P[B_0 \setminus B_R] \leq P(T) = \epsilon \Rightarrow$$

since $e_P(h_R) \leq \epsilon$ exists j such that for all $1 \leq i \leq N, u_i \notin T$

$$P[e_P(h_R) > \epsilon] \leq P[\exists j \forall i : u(i) \in T]$$

we can notice that

$$P[e_P(h_R) > \epsilon] \leq (1-\epsilon)^n \leq e^{-n\epsilon} = \delta \Leftrightarrow n \leq \frac{1}{\epsilon} ln \frac{1}{\delta}$$

now lets set $N(\epsilon, \delta) = \frac{1}{\epsilon} ln \frac{1}{\delta}$
we proved that there exists $N(\epsilon, \delta) = \frac{1}{\epsilon} ln \frac{1}{\delta}$, such that for every ϵ, δ and every realizable distribution P over R^d with labeling function $B_0 \in H_{ball}$, when running A_{ball} on $n \geq N(\epsilon, \delta)$ training examples drawn i.i.d. from P, it returns a hypothesis $h_R \in H_{ball}$ that hold the property above. moreover we can notice that the complexity is not depend on the dimension d

2. PAC in Expectation.

Theorem. *hypothesis class* H *is PAC learnable if and only if* H *is PAC learnable in expectation*

Proof. $\Rightarrow$ by definition exist A for any δ, ϵ such that $P[e_P(A(s)) > \epsilon] \leq \delta$. and $\epsilon, e_P(A(s) > 0$ now by Markov's inequality we get

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

hence for $n \geq N(a)$ lets define $\hat{N}(\epsilon\delta) : (0,1) \to N | \forall a \in (0,1)$ and the following will hold

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

witch stand with the PAC in Expectation definition, with the same A.

$\Leftarrow$ we sow before that the same algorithm A work with both. now using the law of total expectation.

$$E[e_P(A(s))]$$

$$= \underbrace{E[e_P(A(s))|e_P(A(s)) \leq \epsilon]P[e_P(A(s)) \leq \epsilon}_{\leq 1\epsilon} +$$

$$\underbrace{E[e_P(A(s))|e_P(A(s)) > \epsilon]P[e_P(A(s)) > \epsilon]}_{\leq \delta 1}$$

$$\leq \epsilon + \delta$$

and in general its hold for any $\epsilon = 1 - \delta$ hence for $n \geq N(\epsilon, \delta)$ we get the equivalence $\square$

3 Union Of Intervals.

we can notice that any 2k distinct points on the real line can be shattered using k intervals. it suffices to shatter each of the k pairs of consecutive points with an interval. now lets look at set of 2k+1 points assume they sorted $x_1 < x_2 < \ldots x_{2k+1}$, now lets label any x_i with $(-1)^{i+1}$, hence we need 2k+1 intervals to shatter the set because no interval can contain two consecutive points. and the VC dimension is 2k

4 Prediction by polynomials.

The VC dimension of H is the size of the largest set of examples that can be shattered by H$\Rightarrow$ The VC dimension is infinite if for all m, there is a set of m examples shattered by H.
for all $m \in R$ lets say we have sample set size $m = (y_1, y_2 \dots y_m)$ now using the hint we know there for given n distinct values $x_1, ..., x_n \in R$ there exists a polynomial P of degree n - 1 such that $P(x_i) = y_i$. now we can set out some $h_p \in H_{poly}$ and reduce each ϵ from each sample set i.e 2^m times. and each time using the hint above we can label 0-1 all the element for all m.
Hence the VC dimension of H_{poly} is ∞.

5 Structural Risk Minimization.

Lets $\hat{H} = \cup_i^k H_i$ be k finite hypothesis such that $|H_1| \leq \cdots \leq |H_k|$, using the relating empirical and true errors property for any $h_j \in H_i$

$$P[sup_{h\in H}|e_s(h) - e_p(h)|] \leq 2|H|e^{-2n\epsilon^2}$$

now using the union bound we will get

$$\bigcup_{i=1}^{k} P[|e_s(h) - e_p(h)| > \sqrt{\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta}}] \leq$$

$$\sum_{i=1}^{k} P[sup_{h\in H}|e_s(h) - e_p(h)| > \sqrt{\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta}}]] \leq$$

$$kP[|e_s(h) - e_p(h)| > \sqrt{\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta}}]] \leq 2k|H|e^{-2n\epsilon^2}$$

for $|S| = n$ and $\epsilon = \sqrt{\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta}}$

$$\leq 2k|H_i|exp(-2|S|\sqrt{(\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta})^2}) = 2k|H_i|(\frac{2k|H_i|}{\delta})^{-1} = \delta$$

$$\Leftrightarrow \forall i \in \hat{H} \Rightarrow P[|e_s(h) - e_p(h)| > \sqrt{\frac{1}{2|S|}ln\frac{2k|H_i|}{\delta}}]] < 1 - \delta$$

(b)

Lets $\hat{i}$ be the hypothesis s.t SRM return $\text{ERM}_{\hat{i}}$. and lets i^* be index of $h*$

$$e_p(SRM) \leq e_s(\text{ERM}_{\hat{i}}) + \sqrt{\frac{1}{2n} ln \frac{2k|H_{\hat{i}}|}{\delta}} \leq \underbrace{e_s(\text{ERM}_{i^*}) + \sqrt{\frac{1}{2n} ln \frac{2k|H_{i^*}|}{\delta}}}_{\text{result from section a}}$$

now lets reduce h^* from the following ,and using the ERM property we get

$$e_s(\text{ERM}_{i^*}(S)) - e_p(h^*) + \sqrt{\frac{1}{2n} ln \frac{2k|H_{i^*}|}{\delta}} \leq e_s(h^*) - e_p(h^*) + \sqrt{\frac{1}{2n} ln \frac{2k|H_{i^*}|}{\delta}} \leq$$

$$2\sqrt{\frac{1}{2n} ln \frac{2k|H_{i^*}|}{\delta}} \leq \epsilon \Rightarrow n \geq \frac{2}{\epsilon^2} ln \frac{2k|H_{i^*}|}{\delta}$$

hence for $n \geq \frac{2}{\epsilon^2} ln \frac{2k|H_{i^*}|}{\delta}$ we will get the 1-δ probability

6. Programming Assignment Union Of Intervals

(a)

Lets the true distribution $\Pr[x, y] = \Pr[y|x] \Pr[x]$is as x is distributed uniformly on the interval [0, 1], and

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

hence we looking for $h = (\hat{x}) \arg\max_{y \in 0,1} \Pr[Y = y, X = \hat{x}]$
when $x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$,

$$\Pr[Y = 1|X = \hat{x}] = 0.8 > Pr[Y = 0|X = \hat{x}] = 0.2$$

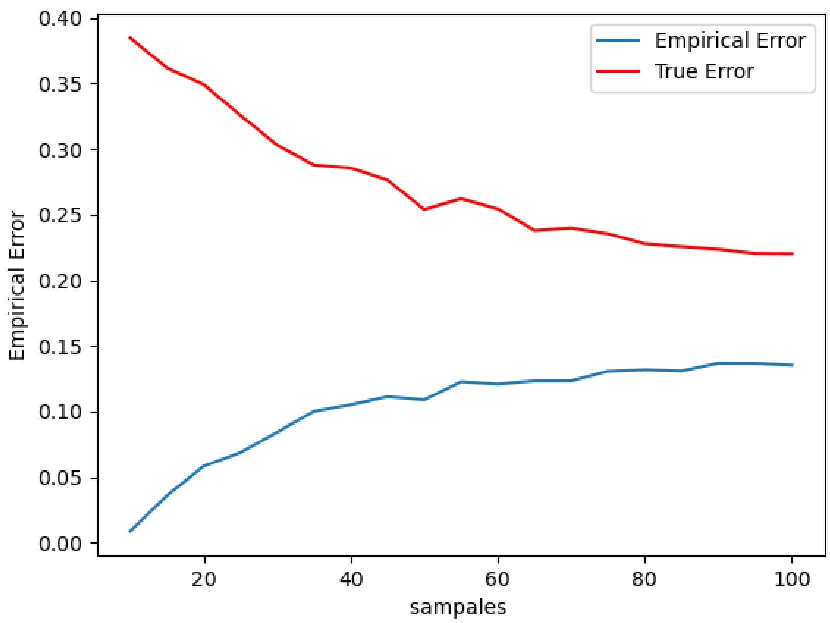
and $x \notin [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$

$$\Pr[Y = 0|X = \hat{x}] = 0.9 > Pr[Y = 1|X = \hat{x}] = 0.1$$

and x is distributed uniformly on the interval [0, 1], and the optimal hypothesis for H_{10}

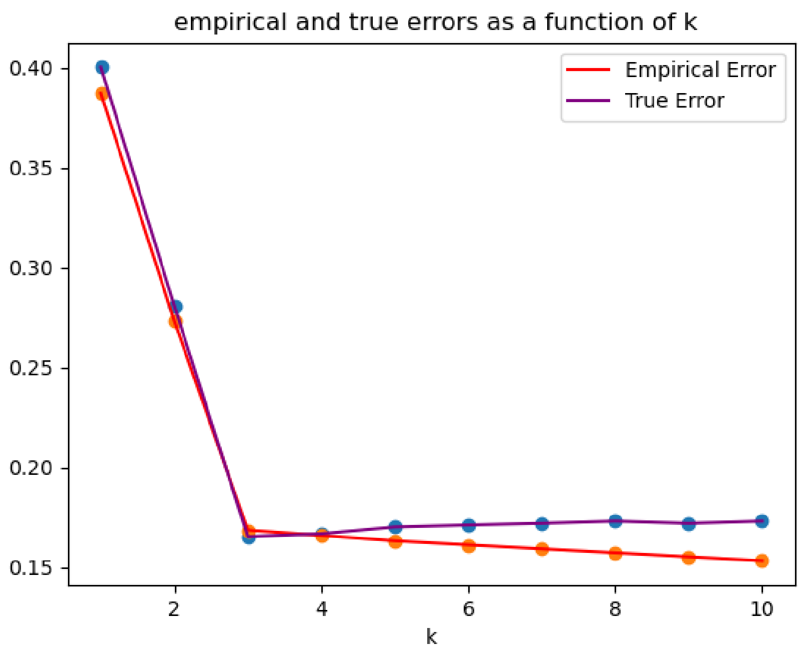
$$h(x) = \begin{cases} 1 & \text{if } x \in [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1] \\ 0 & \text{else} \end{cases}$$

(b)



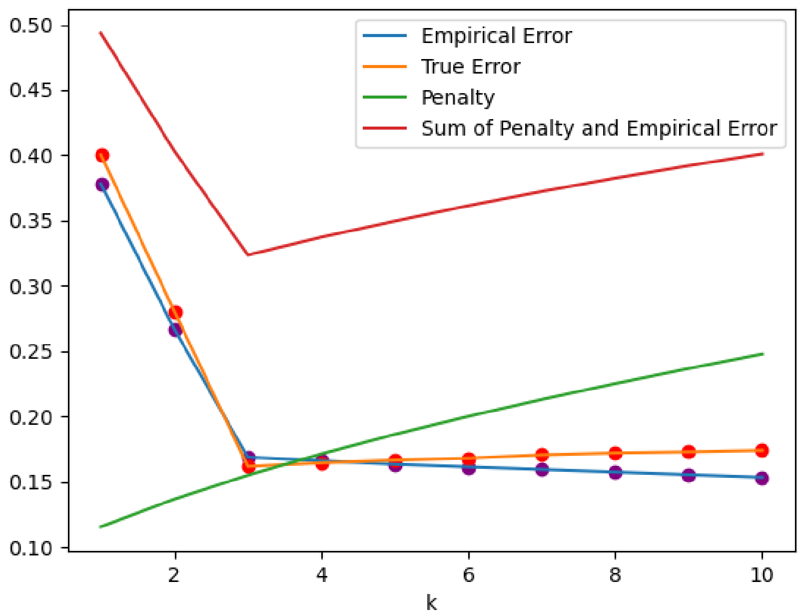
From the plot, we can notice that the empirical error increasing according to the amount of samples taken since the probability to see samples outside the 3 intervals is increasing, moreover we can see that the empirical error approach to 0.15 since its the middle of the false-positve and true-negtive error from section a . i.e $(0.2 + 0.1)/2$. the true error is decreeing since we test more samples since we getting closer to the real distribution P

(c)



The empirical risk decreeing while k increase since the ERM algorithm have more option of disjoint interval to choose for given data so its can cover more samples. on the other hand while $k > 3$ we can notice the true error increasing since the model over fitting to the sample set. and $k = 3$ is the one with the best behaviour

(d)



We can notice that when the when h come from H_3 the sum of the pendalty and the empirical error is minimizing.

(e)

best hypothesis found

```
********************************* The best hypothesis ********************************

[(3.6696865123420075e-05, 0.19918746561659967), (0.4003336336357707, 0.6006597985513338), (0.7988421265362857, 0.9997013095878844)]

***************************************************************************************
```

after drawing the data we can notice that the following stand with the hold out property for $1 - \delta$, witch is close to the optimal true error