Introduction to Machine Learning: Ex02

Theoretical Questions

1.

(12 points) PAC learnability of ℓ_2-balls around the origin. Given a real number $R > 0$ 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 $\mathcal{H}_{ball} = \{h_R \mid R > 0\}$. 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.

We will prover that $\mathcal{H}_{ball}$ is PA C learnable in the realizable case.
Plan:

- Design algorithm A that receives $(x, y) \in \mathbb{R}^d \times \{0,1\}$ and learns $\mathcal{H}_{ball}$ (return $h_r \in \mathcal{H}_{ball}$)
- Show that $\exists N(\varepsilon, \delta). \forall \varepsilon, \delta, P$ A returns h_r such that $e_p(h_r) \leq \varepsilon$ with probability $1 - \delta$.

Lets define positive vector as a vector labeled 1, and negative vector as a vector labeled 0.
First, given S a group of n labeled vectors, we define B_0 the smalless ball (with radius R_0) that is consistent with all the positive-labeled vectors. B_0 can have only false-positive mistakes, meaning negative vectors with distance (from the origin) smaller the R_0. We will define R_p the real distance defined by P so the algorithm error is the probability to have positive vectors with $r \in (R_0, R_p]$ which can be translated to:

$$e_p(h_r) = P[R_p/R_0] \leq P[R_p]$$

Lets assume $P[R_p] > \varepsilon$ since otherwise we get $e_p(h_r) \leq \varepsilon$ immediately.
we can define R_ε which is set to be the radius which has probability of ε to have positive vectors in the area of the ring formed by both R_ε and R_p (lets call this ring C). If there is a positive training vector C we get $R_\varepsilon \leq R_0 < R_p$. In that case the probability to have a positive sample with $R_0 < r$ is bounded by the probability to have a sample with $R_\varepsilon < r$, hence bounded by ε by definition of R_ε.
Therefore in case of error bigger the ε we can understand that there is a sample in C which wasn't part of the learning set or we had no training vector in the ring: so we have $P[e_p(h_r) > \varepsilon] \leq P[\forall i. x_i \notin C]$.
since $P[X_i \in C] = \varepsilon$ we know $P[x_i \notin C] = 1 - \varepsilon$ so $P[\forall i. x_i \notin C] = (1 - \varepsilon)^n \leq e^{-\varepsilon n}$.
All we have to do is make sure this probability $\leq \delta$:

$$e^{-\varepsilon n} \leq \delta \to n \geq \frac{-\ln(\delta)}{\varepsilon} \to N(\varepsilon, \delta) = \frac{-\ln(\delta)}{\varepsilon}$$

2.

(12 points) PAC in Expectation. Consider learning in the realizable case. We say an hypothesis class $\mathcal{H}$ is ***PAC learnable in expectation*** if there exists a learning algorithm A and a function $N(a) : (0,1) \to \mathbb{N}$ such that $\forall a \in (0,1)$ and for any distribution P, 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: On one direction, use the law of total expectation. On the other direction, use Markov's inequality).

we will show that $\mathcal{H}\ is\ PAC\ Learnavble \rightarrow\ \mathcal{H}\ is\ PAC\ Learnable\ in\ Expectation$

Calculating the total expectation and finding upper bounds:

$$\mathbb{E}[e_p(A(S))] = \mathbb{E}\left[e_p(A(S))|e_p(A(S)) \leq \frac{a}{2}\right]\mathbb{P}\left[e_p(A(S)) \leq \frac{a}{2}\right] + \left[e_p(A(S))|e_p(A(S)) > \frac{a}{2}\right]\mathbb{P}\left[e_p(A(S)) > \frac{a}{2}\right]$$

It is obvious (and self-implied) that $\mathbb{E}\left[e_p(A(S))|e_p(A(S)) \leq \frac{a}{2}\right] \leq \frac{a}{2}$, and $\mathbb{P}\left[e_p(A(S)) \leq \frac{a}{2}\right] < 1$ so

$$\mathbb{E}\left[e_p(A(S))|e_p(A(S)) \leq \frac{a}{2}\right]\mathbb{P}\left[e_p(A(S)) \leq \frac{a}{2}\right] < \frac{a}{2}$$

Since $e_p(h) = \mathbb{P}[h(x) \neq Y]$ (as shown in class) we can learn that $e_p(A(S)) = \mathbb{P}[A(S)(x) \neq Y] \leq 1$.

To find an upper boundary for $\mathbb{P}\left[e_p(A(S)) > \frac{a}{2}\right]$ we will use the fact that $\mathcal{H} is\ PAC$ and PAC definition:

Since $\mathcal{H}\ is\ PAC$ there is an algorithm A that satisfies

$$\forall \varepsilon, \delta > 0, \exists N(\varepsilon, \delta)\ s.t\ \forall\ n \geq N(\varepsilon, \delta): \forall P.\ \mathbb{P}\left[e_p(A(S_n)) \leq \varepsilon\right] \geq 1 - \delta$$

Let a be in (0,1). We know that for $\varepsilon = \frac{a}{2}, \delta = \frac{a}{2}$ (according to PAC definition) there is A such that:

$$\forall\ n \geq N\left(\frac{a}{2}, \frac{a}{2}\right): \forall P.\ \mathbb{P}\left[e_p(A(S_n)) \leq \frac{a}{2}\right] \geq 1 - \frac{a}{2}$$

$$\forall\ n \geq N\left(\frac{a}{2}, \frac{a}{2}\right): \forall P.\ \mathbb{P}\left[e_p(A(S_n)) \leq \frac{a}{2}\right] \leq \frac{a}{2}$$

So we learn that $\forall n \geq \widetilde{N}\left(\frac{a}{2}\right), \forall P: \mathbb{E}[e_p(A(S))] \leq 1 \cdot \frac{a}{2} + 1 \cdot \frac{a}{2}$ meaning $\forall n \geq \widetilde{N}\left(\frac{a}{2}\right), \forall P: \mathbb{E}[e_p(A(S))] \leq a$

we found an $N(a)$: (0,1) → ℕ such that for any $a \in (0,1)$ and $\mathbb{P}$, given a sample S such that $|S| \geq N(a)$ satisfies:

$$\mathbb{E}[e_p(A(S))] \leq a$$

we will show that $\mathcal{H}\ is\ PAC\ Learnable\ in\ Expectation \rightarrow \mathcal{H}\ is\ PAC\ Learnavble$

since $\mathcal{H}\ is\ PAC\ in\ Expectation$ there is an algorithm A and function $N(a)$ such that for any a ∈ (0,1) and any $\mathbb{P}$: for a given example S s.t $N(A) < |S| : E[e_p(A(S))] \leq a$.

For any δ, ε lets choose $a = \min(\delta\varepsilon, 1)$ so we get $E[e_p(A(S))] \leq \min(\delta\varepsilon, 1)$ and according to Markov bound:

$$\mathbb{P}[e_p(A(S_n)) > \varepsilon] \leq \frac{\mathbb{E}[e_p(A(S_n))]}{\varepsilon} \leq \frac{\min(\delta\varepsilon, 1)}{\varepsilon}$$

$$\begin{cases} \delta\varepsilon \leq 1 & \mathbb{P}[e_p(A(S_n)) > \varepsilon] \leq \frac{\delta\varepsilon}{\varepsilon} \leq \delta \\ \delta\varepsilon > 1 & \mathbb{P}[e_p(A(S_n)) > \varepsilon] \leq \frac{1}{\varepsilon} \leq \delta \cdot \frac{1}{\delta\varepsilon} \leq \delta \end{cases} \rightarrow \mathbb{P}[e_p(A(S_n)) > \varepsilon] \leq \delta$$

Since a is a function of (δ, ε) we get $N(a) = N(a(\delta, \varepsilon)) = N(\delta, \varepsilon)$

We proved that there exists an algorithm A and function $N(\delta, \varepsilon)$ such that for given training data S that satisfies $|S| \geq N(\delta, \varepsilon)$ and for every $\delta, \varepsilon > 0$ and every $\mathbb{P}$: $\mathbb{P}[e_p(A(S_n)) > \varepsilon] \leq \delta$ so $\mathcal{H}\ is\ PAC$ by definition

We prove $\mathcal{H}\ is\ PAC\ Learnable\ in\ Expectation \leftrightarrow \mathcal{H}\ is\ PAC\ Learnavble$

3.

(10 points) Union Of Intervals. Determine the VC-dimension of the subsets of the real line formed by the union of k intervals (see question 1 of the programming assignment for a formal definition of $\mathcal{H}$).

As offered in class, upon encounter with such question we will use the number of parameters as first guess. Each interval is defined by two parameters hence we will prove $\mathcal{H}$ shutters a set of $2k$ data points and does not shutter set of $2k+1$ data points.

Example for $s \in S_k$: $\mathcal{H}\ shutters\ s \in S_{2k}$

we will define positive point if the label is 1 and negative point if label is 0.

For s set of different $2k$ points, each sub-set of sequential positive points will be in the same interval. Accordingly, each subset of sequential negative points will be between two intervals, meaning that it won't be included in any interval.

We will set each positive point to its interval. If we used $n \leq k$ intervals, we would use $n-k$ empty intervals so we get to a total number of k intervals. the reason n is bounded by k is because the fact that in order to use the maximum number of subsets, we have to minimize no number of elements in each sequential subset: meaning creating $\frac{2k}{2}$ intervals, to match all the single positive point.

assuming any we a sequence longer then one of any type of points (positive/negative) we would use $\mathrm{n} < \mathrm{k}$ intervals.

Proving $\mathcal{H}\ doesn't\ shutter\ S_{2k+1}$:

Any set with $2k+1$ can have a labeling L_a that labels points in alternate labeling (starting with positive).

L_a can't be achieved using k intervals. we mark $\forall i \in (0,1,\cdots 2k_1)\ p_i$ the point with the i^{th} value (sorted) and we define L as:

$$L_a(p_i) = \begin{cases} negative & i \in \mathbb{N}_{odd} \\ positive & else \end{cases}$$

L_a generates $k+1$ subsets of sequential positive points $positive(p_i) = \{\{\ p_i\ \} : i \in \mathbb{N}_{even}\} = \{\{0\},\{2\},\cdots\{2k\}\}$.

Assuming L_a can be achieved using k intervals

covering all positive points requires $\mathcal{H}$ to use an interval containing two positive points. If those points are p_j, p_{j+2} so p_{j+1} (which is negative) will be in the interval as well: hence L_a can't be achieved with k intervals.

4.

(10 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,

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

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 $1 \leq i \leq n$.

We will show that $VC_d(\mathcal{H}_{poly}) = \infty$:
According to the fact mentioned in the question: for any set $S = \{(x_i, v_i)\}$ that satisfies $|S| = n$ there is a polynomial P_{n-1} that can be constructed using n parameters that generates $P(x_i) = v_i$ as we desire:

$$P_{n-1} = \sum_{i=0}^{n-1} a_i x^i$$

So, since we have no limitation of values for z_i we can use a polynomial $p \in P_{n-1}$ for each possible set $S' = \{(x_i . v_i)|' f \forall i, j . h(x_i, x_j) \; is \; as \; we \; desire\}$ meaning the hypothesis can have 2^n possible outcomes. So we can learn that $\forall n \geq 2 . \exists P_n$ such that $h_{P_n}(x_1, x_2)$ shutters $\mathcal{H}_{poly_n} = \{h_p | \; h_p \in \mathcal{H}_{poly} \wedge P = \Theta(x^n\}$. This means $VC_d\left(\mathcal{H}_{poly_n}\right) = n$. Rewriting $\mathcal{H}_{poly}$ as

$$\mathcal{H}_{poly} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_{poly_n} \to \lim_{n \to \infty} \left(VC_d(\mathcal{H}_{poly})\right) = \lim_{n \to \infty} (n) = \infty$$

5.

(16 points) Structural Risk Minimization. Let $\mathcal{H}_1, ..., \mathcal{H}_k$ be k finite hypothesis classes such that $|\mathcal{H}_1| \leq ... \leq |\mathcal{H}_k|$, and let $\mathcal{H} = \cup_{i=1}^{k} \mathcal{H}_i$.

(a) Show that if S is a set of training samples chosen i.i.d from the data generating distribution, then with probability $1 - \delta$, for every $1 \leq i \leq k$ and $h \in \mathcal{H}_i$,

$$|e_P(h) - e_S(h)| \leq \sqrt{\frac{1}{2|S|} \ln \frac{2k|\mathcal{H}_i|}{\delta}}.$$

(b) Let $h^* = \arg\min_{h \in \mathcal{H}} e_P(h)$ and $i^* = \min\{1 \leq i \leq k \mid h^* \in \mathcal{H}_i\}$. Show that if $|S| \geq \frac{2}{\epsilon^2} \ln \frac{2k|\mathcal{H}_{i^*}|}{\delta}$ then with probability of $1 - \delta$,

$$e_P(SRM(S)) \leq e_P(h^*) + \epsilon.$$

Remark This implies that if h^* is "simpler" (i.e. belong to small class) it will require fewer samples to learn.

a.

We are about to prove that for a training set S chosen from IID data generating distribution:

$$\mathbb{P}\left[\forall h, \forall 1 \leq i \leq k. |e_P(h) - e_S(h)| \leq \sqrt{\frac{1}{2|S|} \ln\left(\frac{2k|\mathcal{H}_i|}{\delta}\right)}\right] \geq 1 - \delta$$

Using negation, we will prove the equivalent form:

$$\mathbb{P}\left[\forall \exists h, i. |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} \ln\left(\frac{2k|\mathcal{H}_i|}{\delta}\right)}\right] < \delta$$

Let's mark $\frac{2k|\mathcal{H}_i|}{\delta}$ as p_i For a specific i we know that

$$\mathbb{P}\left[\exists h: |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} \ln(p_i)}\right] \leq \mathbb{P}\left[\sup_{h \in \mathcal{H}} |e_S(h) - e_P(h)| > \sqrt{\frac{1}{2|S|} \ln(p_i)}\right] \leq 2|\mathcal{H}_i| e^{-2|S| \cdot \frac{1}{2|S|} \ln(p_i)} \leq \frac{2|\mathcal{H}_i|}{p_i} \leq 2\frac{\delta}{k}$$

$$\mathbb{P}\left[\exists h: |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} \ln(p_i)}\right] \leq \frac{\delta}{k}$$

When looking on all possible i values: taking into consideration union:

$$\mathbb{P}\left[\exists h, \exists i: |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} ln\left(\frac{2k|\mathcal{H}_i|}{\delta}\right)}\right] \leq \sum_{i=1}^{k} \mathbb{P}\left[\exists h: |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} ln\left(\frac{2k|\mathcal{H}_i|}{\delta}\right)}\right] \leq \frac{\delta}{k} \cdot k \leq \delta$$

$$\mathbb{P}\left[\forall \exists h, i. |e_P(h) - e_S(h)| > \sqrt{\frac{1}{2|S|} \ln\left(\frac{2k|\mathcal{H}_i|}{\delta}\right)}\right] < \delta$$

b.

Assume SRM returned $h_{srm} \in \mathcal{H}_j$ (j: the first index $h_{srm} \in \mathcal{H}_j$): $\mathbb{P}[e_P(h_{srm}) - e_P(h^*) \leq \epsilon] \geq 1 - \delta$

So, with certainty $1 - \delta$ we know : $e_p\big(ERM_i(S)\big) \leq e_s(ERM_i(S) + \sqrt{\frac{1}{2n}\ln\left(\frac{2|H_i|k}{\delta}\right)}$

All of the following computation is done with the same certainty $(1 - \delta)$

Therefore:

$$e_P(h_{srm}) - e_P(h^*) \leq e_S(\bar{h}) + \sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_j|k}{\delta}\right)} - e_P(h^*)$$

Since $\bar{h}$ was returned from SRM, we know that:

$$e_S(\bar{h}) + \sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_j|k}{\delta}\right)} \leq e_S(h^*) + \sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_{i^*}|k}{\delta}\right)}$$

Therefore:

$$e_P(h_{srm}) - e_P(h^*) \leq e_S(\bar{h}) + \sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_j|k}{\delta}\right)} - e_P(h^*) \leq e_S(h^*) - e_P(h^*) + \sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_{i^*}|k}{\delta}\right)}$$

Using what we learned from section a:

$$e_P(h_{srm}) - e_P(h^*) \leq 2\sqrt{\frac{1}{2|S|}\ln\left(\frac{2|\mathcal{H}_{i^*}|k}{\delta}\right)}$$

And finally, if $|S| \geq \frac{2}{\epsilon^2}\ln\left(\frac{2k|H_{i^*}|}{\delta}\right)$:

$$e_P(h_{srm}) - e_P(h^*) \leq 2\sqrt{\frac{\epsilon^2}{4}} = \epsilon$$

Programming Assignment

1.

a.

(8 points) Assume 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

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

and $P[y = 0|x] = 1 - P[y = 1|x]$. Since we know the true distribution P, we can calculate $e_P(h)$ precisely for any hypothesis $h \in \mathcal{H}_k$. What is the hypothesis in $\mathcal{H}_{10}$ with the smallest error (i.e., $\arg\min_{h \in \mathcal{H}_{10}} e_P(h)$)?

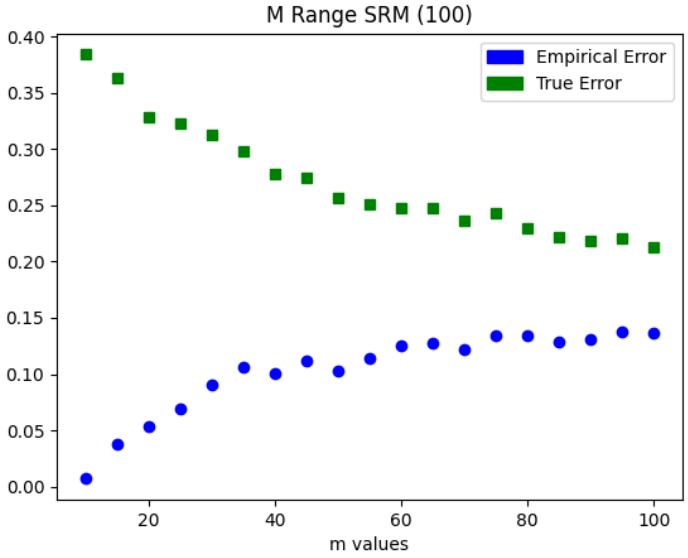
Using the optimal classifier $e_p(h) = E\left[l_{0-1}(Y, h(x))\right]$ Given an $x \in [0,0.2] \cup [0.4,0.6] \cup [0.8,1]$ we will classify 1, otherwise 0. So $P[h(x) = 0] = 0.4$ and $P[h(x) = 1] = 0.6$. we can also see that $P[Y = 1|h(x) = 0] = 0.1$ and $P[Y = 0|h(x) = 1] = 1 - 0.8$

$$e_p(h) = P[Y = 0, h(x) = 1] + P[Y = 0, h(x) = 0] = P[Y = 0|h(x) = 1]P[h(x) = 1] + P[Y = 1|h(x) = 0]P[h(x) = 0]$$

$$argmin_{h \in \mathcal{H}_{10}} \left(e_p(h)\right) = 0.2 \cdot 0.6 + 0.1 \cdot 0.4 = 0.16$$

b.

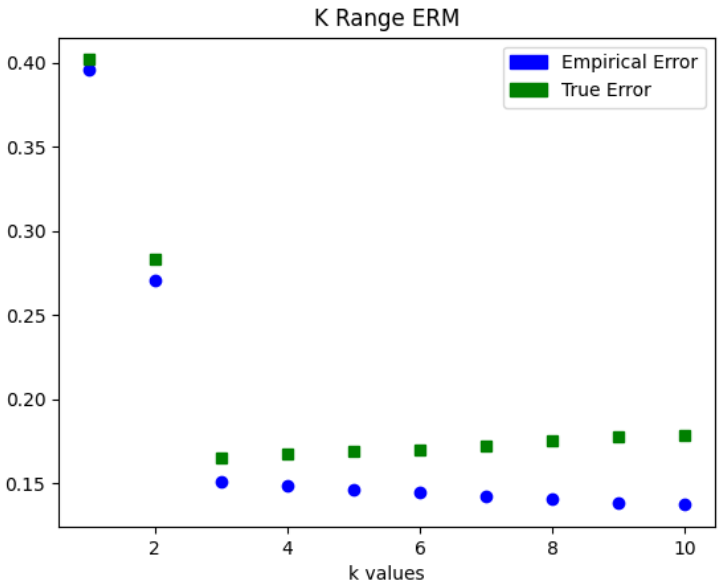
(8 points) Write a function that, given a list of intervals I, calculates the true error $e_P(h_I)$. Then, for $k = 3$, $n = 10, 15, 20, \ldots, 100$, perform the following experiment $T = 100$ times: (i) Draw a sample of size n and run the ERM algorithm on it; (ii) Calculate the empirical error for the returned hypothesis; (iii) Calculate the true error for the returned hypothesis. Plot the empirical and true errors, averaged across the T runs, as a function of n. Discuss the results. Do the empirical and true errors decrease or increase with n? Why?



As n grows the set of samples is more similar to the original distribution P.
as a result the ERM classifier is more accurate as can be understood from the decreasing of true error with increase of n.
Increasing n has a negative effect on the empirical error, since it becomes harder to find a classifier to match a bigger amount of samples.

c.

(8 points) Draw a sample of size $n = 1500$. Find the best ERM hypothesis for $k = 1, 2, \ldots, 10$, and plot the empirical and true errors as a function of k. How does the error behave? Define k^* to be the k with the smallest empirical error for ERM. Does this mean the hypothesis with k^* intervals is a good choice?



We clearly see from the graph that $k^* = 10$ (minimizes empirical error).

It makes sense that k^* is the highest value possible $(\max(k))$ since it provides more freedom for ERM to fit for training set.

But when considering TE (true error) in oppose to empirical error we can see that for $k \geq 3$ even though $\left|\frac{d}{dk}(TE)\right| \approx 0$, the true error increases as k increases $\left(\frac{d}{dk}(TE) < 0\right)$.

Using a $k > 3$ will have minor negative effect on the true error but more important will increase the complexity, so it is clear that the recommended k for minimizing true error is $k_T = 3$.

Using k^* intervals doesn't seem such a good idea, choosing k_T intervals instead seem a better decision.

d.

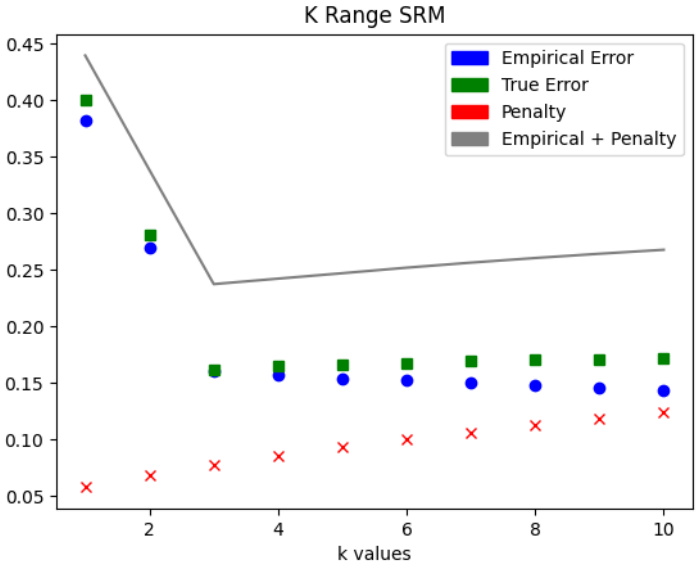
(8 points) Now we will use the principle of structural risk minimization (SRM), to search for a k that gives a good test error. Let $\delta = 0.1$:

- Use to following penalty function:

$$2\sqrt{\frac{\text{VCdim}(\mathcal{H}_k) + \ln\frac{2}{\delta}}{n}}$$

- Draw a data set of $n = 1500$ samples, run the experiment in (c) again, but now plot two additional lines as a function of k: 1) the penalty for the best ERM hypothesis and 2) the sum of penalty and empirical error.
- What is the best value for k in each case? is it better than the one you chose in (c)?

We saw in questions 3 (in this assignment) that $VCdim(\mathcal{H}_k) = 2k$ so $Penalty = \sqrt{\frac{2k+\ln\left(\frac{2}{\delta}\right)}{n}}$. In the script m is replacing n.

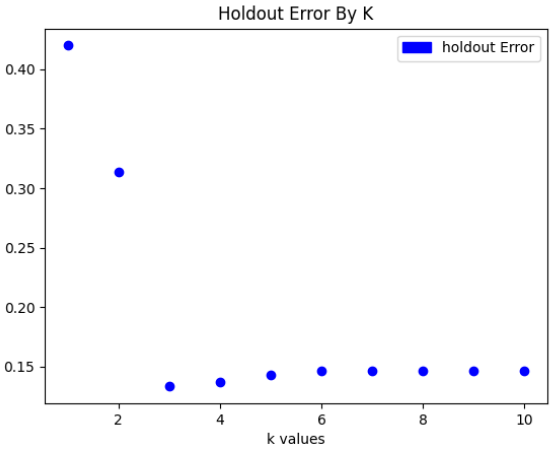


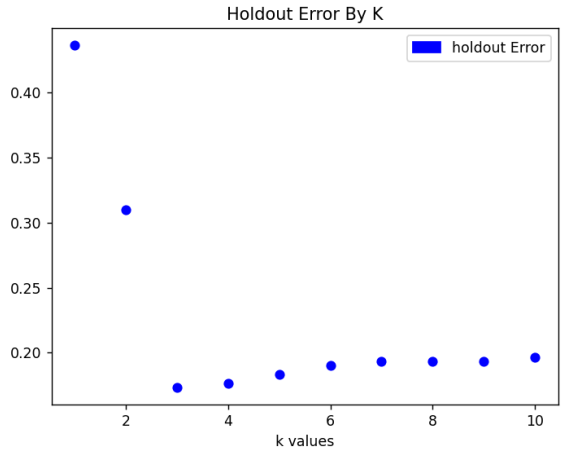
We now reassured that optimal $k = 3$ (since higher values result over-fitting)

e.

(8 points) Here we will use holdout-validation to search for a $k \in \{1, ..., 10\}$ that gives good test error. Draw a data set of $n = 1500$ samples and use 20% for a holdout-validation. Choose the best hypothesis and discuss how close this gets you to finding the hypothesis with optimal true error.

Running 2 times cross-validation function on 1500 samples with hold out ratio of 20% resulted the following:





We learn that the observation from section C, and reassured in section D, is valid.
The optimal k, which achieves the best performance is $k_{opt} = 3$. we can see that empirical error is not the best estimator.
after running the algorithm twice: we had the following results:

```
best k:3
with intervals:
[('0.0001', '0.2006'), ('0.4005', '0.6019'), ('0.8036', '0.9987')]
with intervals:
[(0.00014298043783189662, 0.20061739514033994), (0.40049358914500754, 0.6018803712585661), (0.8035631079747165, 0.998691
```

This print-out of the run shows that we managed to generate a decision rule approximately equal to true $\mathbb{P}$ of the data.
Another interesting observation is that
on the 1st run we had 52 errors on the hold out (out of 300), and 208 empirical errors (out of 1200): having 17.3% true error
on the 2nd run we had 40 errors on the hold out (out of 300), and 194 empirical errors (out of 1200): having 15.6% true error
The average is 16.45% which is pretty close to the true error calculated in section A (16%)