Yoav Shoshan

Theory Questions (hw2)

(1) $$\Delta(y,\hat{y}) = \begin{cases} 0 & \hat{y}=y \\ \lambda & \hat{y}=L+1 \\ 1 & \text{otherwise} \end{cases}$$

(a) $$L(h) = \sum_{x,y} P(Y=y \wedge X=x)\,\Delta(y,h(x)) =$$

for a given $\hat{x}$ $$= \sum_{y} P(Y=y|X=\hat{x})\,P(X=\hat{x})\,\Delta(y,h(\hat{x}))$$

$$= P(X=\hat{x}) \cdot \sum_{i} P(Y=i|X=x)\,\Delta(i,h(\hat{x}))$$

$$= P(X=\hat{x}) \cdot \begin{cases} \lambda \cdot \sum_{i} P(Y=i|X=\hat{x}) & h(\hat{x})=L+1 \\ \sum_{j\neq i} P(Y=j|X=\hat{x}) & h(\hat{x})=i \end{cases}$$

$\forall i \in [L]$ s.t. $i = \arg\max_{j\in[L]} P(Y=j|X=\hat{x})$

and $P(Y=i|X=\hat{x}) \geq 1-\lambda$

$$\sum_{j\neq i} P(Y=j|X=\hat{x}) = 1 - P(Y=i|X=\hat{x}) \leq 1-(1-\lambda) = \lambda \;//$$

$$\lambda \underbrace{\sum_{i} P(Y=i|X=\hat{x})}_{=1} = \lambda$$

$$\Longrightarrow \sum_{j\neq i} P(Y=j|X=\hat{x}) \leq \lambda \cdot \sum_{i} P(Y=i|X=\hat{x})$$

Also, for $i = \arg\max_{j=[L]} P(Y=j \mid X=\hat{x})$

$$\sum_{j \neq i} P(Y=j \mid X=\hat{x}) = \sum_{j} P(Y=j \mid X=\hat{x}) - P(Y=i \mid X=\hat{x})$$

is minimal

if $P(Y=i \mid X=\hat{x}) < 1-\lambda$

$$\Rightarrow \sum_{j \neq i} P(Y=j \mid X=\hat{x}) = 1 - P(Y=i \mid X=\hat{x}) > 1-(1-\lambda)$$

$$> \lambda //$$

$$\Rightarrow \lambda = \lambda \sum_{j} P(Y=j \mid X=\hat{x}) < \sum_{j \neq i} P(Y=j \mid X=\hat{x})$$

$\Rightarrow$ we Alwayse Reject ! //

(b) we note that for $\lambda = 0$ $L(h)$ is minimal only for $h(\hat{x}) = L+1$ $\forall \hat{x}$ and for $\lambda = 1$:

$$\sum_{j \neq i} P(Y=j \mid X=\hat{x}) \leq \lambda \sum_{j} P(Y=j \mid X=\hat{x})$$

$\Rightarrow$ we never Reject !

λ is a Parameter that controls the probability to reject. when it increases from 0 to 1 so dose the chance that h will reject.

(2) (a) a good Algorithm for ERM shall be:
given a sample S of pairs (x_i, y_i).

- if there exist a pair $(x_j, 1)$ $ERM(S) = h_{x_j}$. This is because in the realizable case, given a pair like this, we know the true distribution P. for this case $\epsilon_p(ERM(S)) = 0$.
- if $y_i = 0 \quad \forall (x_i, y_i) \in S$, then $ERM(S) = h^-$. meaning we say that $\forall x \in \mathcal{X} \quad h^-(x) = 0$.
this is of cours consistent with $h(x) = \arg\max_{h \in H} P(Y=1 | X=x)$:

Say we choose $h_{x'}$ for a random $x' \notin S$ instead h

$$\lim_{\Delta \to 0} (h) = \begin{cases} \sum_{x \neq x'} P(Y=1|X=x)P(x) + P(Y=0|X=x')P(x') & h = h_{x'} \\ \sum_{x} P(Y=1|X=x)P(x) & h = h^- \end{cases}$$

$\Rightarrow$ Choosing $h = h^-$ if $P(Y=0|X=x') \geq P(Y=1|X=x'$

but for $|\mathcal{X}| > |S|$ choosing x' for wich $y' = 0$,

$$P(x=x', y=1) = \frac{1}{|\mathcal{X}| - |S|}; \quad P(x=x', y=0) \geq \frac{|\mathcal{X}| - |S| - 1}{|\mathcal{X}| - |S|}$$

$$\Rightarrow P(Y=1|X=x') = \frac{1}{P(x')(|\mathcal{X}| - |S|)} \leq \frac{|\mathcal{X}| - |S| - 1}{P(x')(|\mathcal{X}| - |S|)} \leq P(Y=0|x=x')$$

(2)(b) * if we had some $(x_i, 1)$ in our sample - Then we know the distribution and $e_p(h) = 0$.
otherwise, according to our ERM we choose $h = h^-$, $h^-(x) = 0 \ \forall x$.

Assume there exist $x' \notin S$ for which $y(x') = 1$ and $p(x') \geq \varepsilon$.
(if $p(x') < \varepsilon$ then $e_p(h^-) = p(Y=1 | X=x') p(x') < \varepsilon$ and we are done.)

Then:

$$p(e_p(h) > \varepsilon) = \overbrace{p(e_p(h) > \varepsilon | x' \in S)}^{=0 \ *} \cdot p(x' \in S) + \underbrace{p(e_p(h) > \varepsilon | x' \notin S)}_{\leq 1} p(x' \notin S)$$

$$\Rightarrow p(e_p(h) > \varepsilon) \leq p(x' \notin S) = p\left(\bigcup_{x_i \in S} x_i \neq x'\right)$$

$$= \prod_i p(x_i \neq x') = \prod_i (1 - p(x')) \leq \prod_i (1 - \varepsilon)$$

$$= (1-\varepsilon)^n \leq e^{-\varepsilon n} \overset{!}{\leq} \delta$$

$$\Rightarrow n \geq \frac{1}{\varepsilon} \ln \frac{1}{\delta}$$

PAC learnable!

(3) PAC learnable in expectation

def: for Algorithm A and $N(a):(0,1)\to\mathbb{N}$ s.t. $\forall a\in(0,1)$, $\forall p$ and sample S so that $|S|\geq N(a)$

$\Rightarrow E[\ell_p(A(S))]\leq a$

H is PAC learnable $\Leftrightarrow$ H is PAC learnable with expectation

Proof

H is PAC learnable $\Rightarrow$

$\forall n > N(\varepsilon,\sigma)$ and sample S from p

Denote $\ell_p(A(S)) \equiv \ell_p'$

$$E[\ell_p'] = \underbrace{E[\ell_p' \mid \ell_p'\geq\varepsilon]}_{\leq 1}\underbrace{P(\ell_p'\geq\varepsilon)}_{\leq\sigma} + \underbrace{E(\ell_p'\mid \ell_p'<\varepsilon)}_{\leq\varepsilon}\underbrace{P(\ell_p'<\varepsilon)}_{\leq 1}$$

total expectation

H is PAC learnable

$\Rightarrow E(\ell_p')\leq \sigma+\varepsilon$

Specifically there exist $\varepsilon=\sigma=\frac{a}{2}$. $a\in(0,1)$

$\Rightarrow E(\ell_p)\leq a \Rightarrow H$ is PAC learnable in expectation

Conversely, if H is PAC learnable in expectation

then: $P(\ell_p'\geq\varepsilon)\leq \frac{E(\ell_p')}{\varepsilon} \leq \frac{a}{\varepsilon} = \sigma$

Markov $(0\leq\ell_p'\leq 1)$

$a\equiv\sigma\cdot\varepsilon$

$\sigma,\varepsilon\in(0,1)$

$\Rightarrow H$ is PAC learnable.

(4) $H_k = \{h_I \mid I = \{(l_1, u_1) \dots (l_k, u_k)\}$ (disjoint intervals)

First, we prove that there exist a sample of size $2k$ that can be shattered meaning $VC\dim(H_k) \geq 2k$

Proof: for a set of unique x_i (sorted by the index i) and corresponding $y_i \in \{0,1\}$ we can choose $h(x_i)$ like this:

- for a sequence of $j = m, m+1 \dots m+l$ such that $y_j = 1 \ \forall j$, we define an interval $[x_m, x_{m+l}]$ in wich $h(x_j) = 1$
- we take l, m such that $y_{m-1} = y_{m+l+1} = 0$ meaning, we consider the longest possible sequence of j's and not subsequences of it.

We are expected to use at most k intervals this is for the case of $y_1 = 0$, $y_i = |y_{i-1} - 1|$ meaning all labels are different from the labels of the two nearest neighbors and we have k sequences of j's, each with the length 1.

Also, if we use less than k intervals to cover all sequences of j's, we can always create two different intervals from one. $[x_m, x_{m+l}] \to [x_m, x_p] \cup [x_p, x_{m+l}]$ until we have k intervals to describe h such that $h \in H_k$.

Now we shall proove that $VCdim(H_k) \leq 2k$

Consider $2k+1$ points x_i and the set of coresponding labels $y_i = i \% 2$.
$(1, 0, 1 \ldots 1)$. (x_i dose not have to be unique points)

for this set of labeles we have to define $k+1$ disjoint intervals, each include exactly 1 point x_i for which $y_i = 1$. but we only have k intervals to use, so it to adjucent interval ar streched $[x_m, x_{m+1}), (x_{m+1}, x_{m+2}]$

included because $y_m = 1$ — not included $y_{m+1} = 0$ — included

$\longrightarrow [x_m, x_{m+2}]$ we now have k intervals, but for one of the points, x_{m+1} for which $y_{m+1} = 0$ we get $h(x_{m+1}) = 1$

$\Longrightarrow$ the given labels sequence could not be achived using any $h \in H_k$

$\Longrightarrow$ $VCdim \leq 2m$

$\Longrightarrow$ $VCdim = 2k$ //

using first part as well

(5)

$\sum_{h\in H} w(h) \le 1 \quad H = \bigcup_{i\in N} \{h_i\}$

$w(h) \in [0,1]$, $\sum_{h\in H} w(h) \le 1$

we start with the fact that

$P(|e_p(h) - e_s(h)| > \varepsilon) \le 2|H| e^{-2n\varepsilon^2} \le \delta$

$\Rightarrow \varepsilon \le \sqrt{\frac{1}{2n} \ln \frac{2|H|}{\delta}}$

which means that

$P\left(e_p > e_s + \sqrt{\frac{1}{2n} \ln \frac{2|H|}{\delta}}\right) \le \delta$

also, applying uniform convergence for each h_i seperatly with $\delta_i = \delta \cdot w(h_i)$

$P\left(e_p(h_i) > e_s(h_i) + \sqrt{\frac{1}{2n} \ln \frac{2|H|_i}{w(h_i)\delta}}\right) \le \delta_i$

define A_{h_i} as the event $e_p(h_i) > e_s(h_i) + \sqrt{\frac{1}{2n} \ln \frac{2|H|}{w(h_i)\delta}}$, then

$P(\cup A_i) \le \sum_i P(A_i) \le \sum_i \delta \cdot w(h_i) \le \delta$

meaning, with confidence $1-\delta$, $\forall h \in H$

$e_p(h) \le e_s(h) + \sqrt{\frac{1}{2n} \ln \frac{2|H|}{\delta w(h)}}$

Programming Assignment

(b) in this case, we know the true prob. and we would like to choos that is consistent with $h = \arg\max_{y \in \{0,1\}} P(Y=y \mid X=x)$

because we have intervals I for which the prob for $y=1$ is > 0.5 for $x \in I$ and < 0.5 for $x \notin I$ we can choose

$$h^*(x) = \begin{cases} 1 & x \in I \\ 0 & x \notin I \end{cases} \qquad I = [0, 0.2] \cup [0.4, 0.6] \cup [0.8, 1]$$

the true error for such h is

$$e_p(h^*) = P(Y=0 \wedge h^*(x)=1) + P(Y=1 \wedge h^*(x)=0)$$

$$= P(Y=0 \mid h^*(x)=1) P(h^*(x)=1) + P(Y=1 \mid h^*(x)=0) P(h^*(x)=0)$$

$$= \begin{cases} x \sim U(0,1) \text{ and } P(h(x)=1) = P(x \in I) \\ \qquad\qquad\qquad\quad P(h(x)=0) = P(x \notin I) \\ \text{so putting into the eq the numbers from P} \end{cases}$$

$$= 0.2 \cdot \frac{\text{length}(I)}{1} + 0.9 \cdot \frac{\text{length}(I^c)}{1} =$$

$$0.2 \cdot 0.6 + 0.1 \cdot 0.4 = 0.16 //$$

this is of course the minimal true error

(c) from the results we can see that the empirical error increases with m - it is easy to find 3 intervals for correct classification of several points, but with an increasing sample size we expect some points that will not be consistent with our hope to classify many neighboring points with the same label using an interval. (see plot of section a). still, in any interval, P can generate either 0 or 1 labels. The true error however decreases with m. the more samples we have, the more likely we are to choose h that classifies with condition similar (Although not identical) to h^* (the best hypothesis)

(d) As we expect, the empirical error drops with k - more degrees of freedom in choosing intervals that will classify more sample points correctly.
The true error however, increases with k. This is of course consistent with the fact that it should increase with $|H|$ as we seen in class.
$k^* = k$ with the best empirical error.
we got $k^* = 10$ but it is of course not the best k, but merely the best for describing the sample. (with $k = 0 \dots 10$)

if we had ran the algorithm with largest k's, we would get larger k^* (expectedly, the largest k) because we had more degrees of freedom in the sample classification. it would not, however decrease the true error that after $k=3$ will increas do to poorer representation of the true distribution.

(e) Qn. 2 in theoretical part:
we got VCdim = 2k.
Also we have

$$P\left[\sup_{h\in H} |e_s(h) - e_p(h)| \geq \varepsilon\right] \leq 4\, \Pi_H(2n)\, e^{-\frac{n\varepsilon^2}{8}}$$

where $\Pi_H(n) \leq \left(\frac{en}{d}\right)^d$ $(d \equiv VCdim)$

$$\Rightarrow P\left[\sup |e_s(h) - e_p(h)| \geq \varepsilon\right] \leq 4 \cdot \left(\frac{2en}{2k}\right)^{2k} e^{-\frac{n\varepsilon^2}{8}}$$

$$= 4\left(\frac{n}{k}\right)^{2k} e^{2k - n\varepsilon^2/8} \overset{!}{\leq} \delta$$

$$e^{2k - n\varepsilon^2/8} \leq \frac{\delta}{4}\left(\frac{k}{n}\right)^{2k}$$

$$2k - n\varepsilon^2/8 \leq \ln\left(\frac{\delta}{4}\right) + 2k \ln\left(\frac{k}{n}\right)$$

$$\varepsilon^2 \geq \frac{16k}{n} + \frac{8}{n}\ln\left(\frac{4}{\delta}\right) + \frac{16k}{n}\ln\left(\frac{n}{k}\right)$$

$$\varepsilon \geq \sqrt{\frac{8}{n}\left(2k + 2k\ln\left(\frac{n}{k}\right) + \ln\left(\frac{4}{\delta}\right)\right)}$$

$\delta = 0.1$ $\Rightarrow$

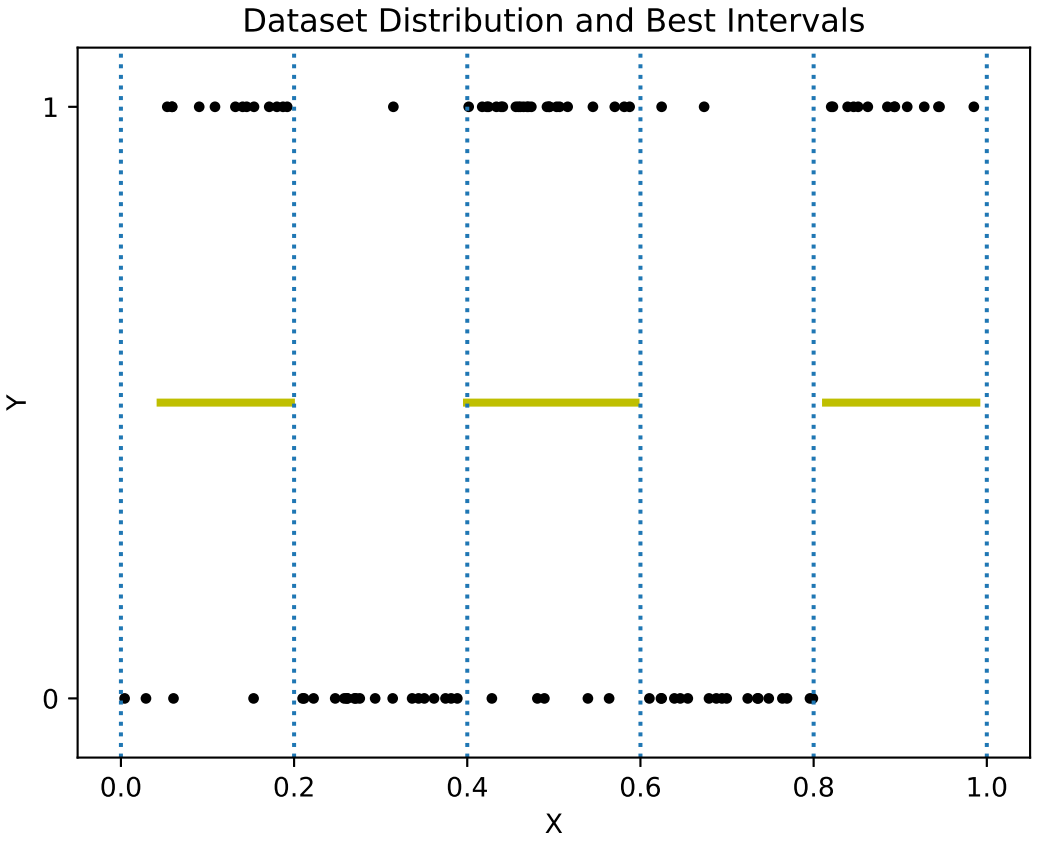
$$e_p(h) \leq e_s(h) + \sqrt{\frac{8}{n}\left(2k + 2k\ln\left(\frac{n}{k}\right) + \ln 40\right)}$$

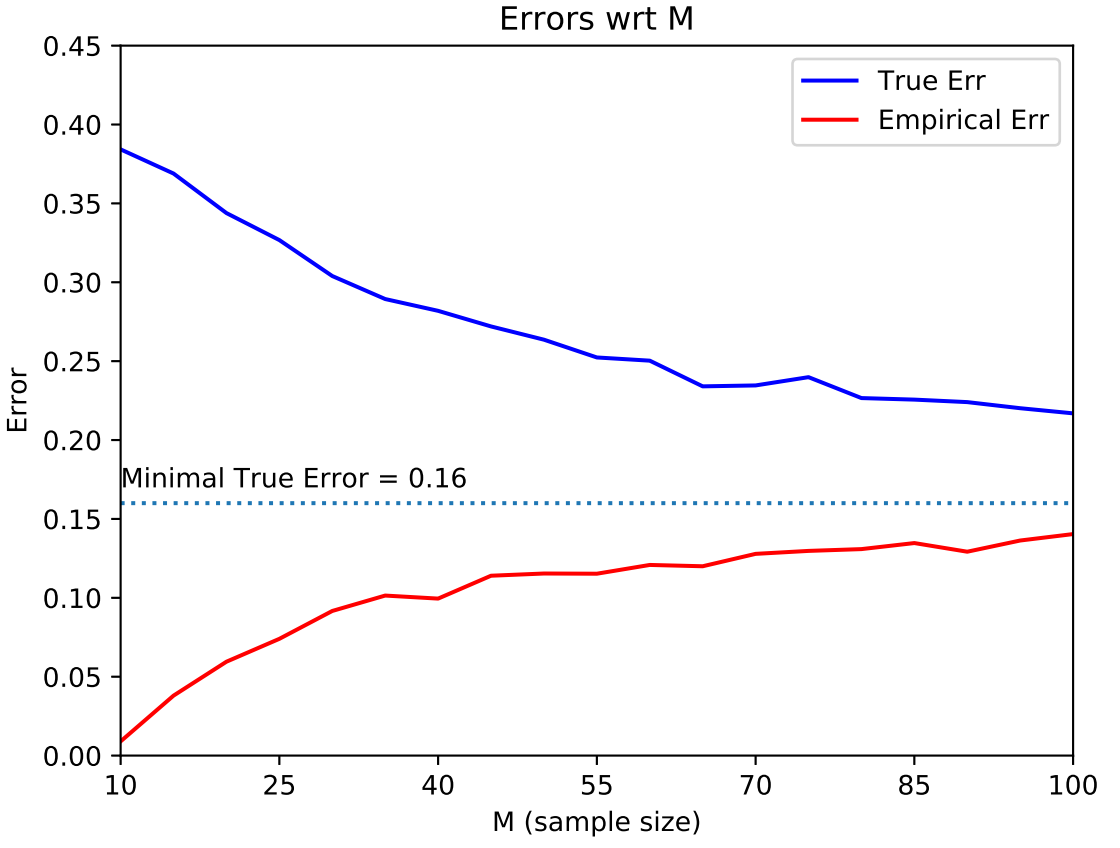
we can see that the penalty $\sim \sqrt{k - \ln k}$ which increases with k. smallest penalty for $k = 1$. the last inequality will hold in prob $\geq 1 - \delta$
in section d we had $k_{best} = 3$ of course, as explained in section b.

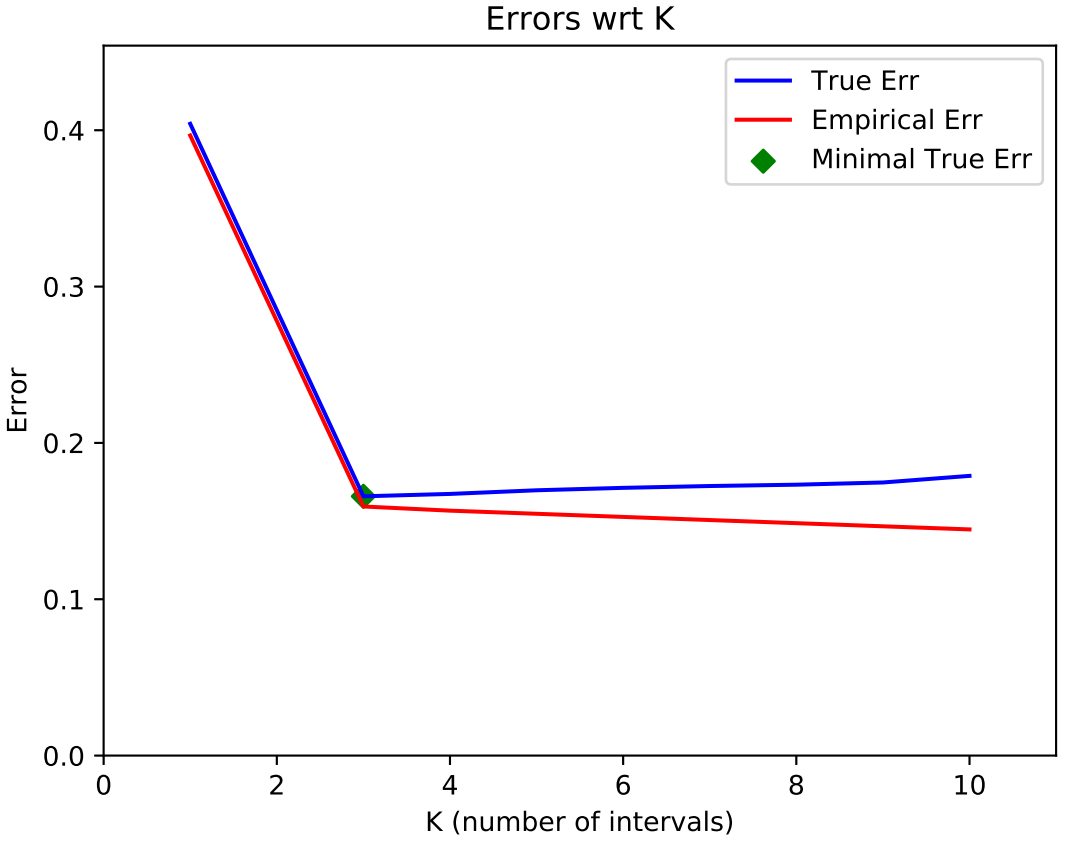
see plot for all errors wrt k

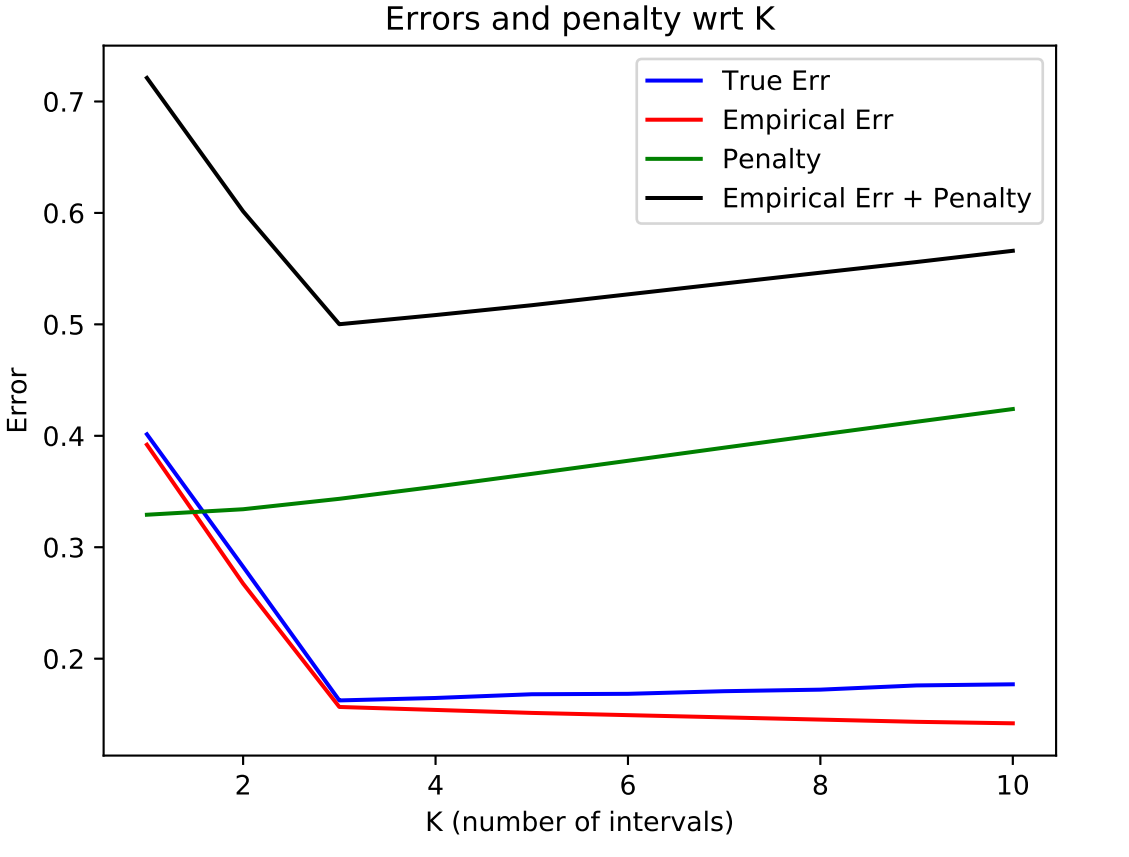
(f) Among 3 experiments, each for different values of k and a different subset of m points as a holdout set, we got that the best k (in most exp.) is $k = 3$ (sometimes one of the exp. yields best $k = 4$). This is not surprising of course as we expect that the holdout set would simulate the true distribution and therefore the error on it shall be close to the true error given the k intervals returned by the ERM on the train data.
As discussed in section 2e, the best hypothesis shall be 3 disjoint intervals identical to those with the higher prob. for $y = 1$ in the true distribution, in which all points are classified as $h(x) = 1$. The holdout set helps us find a hypothesis close to it because it represent the true distribution in nature. (see stdout for the intervals of each k)

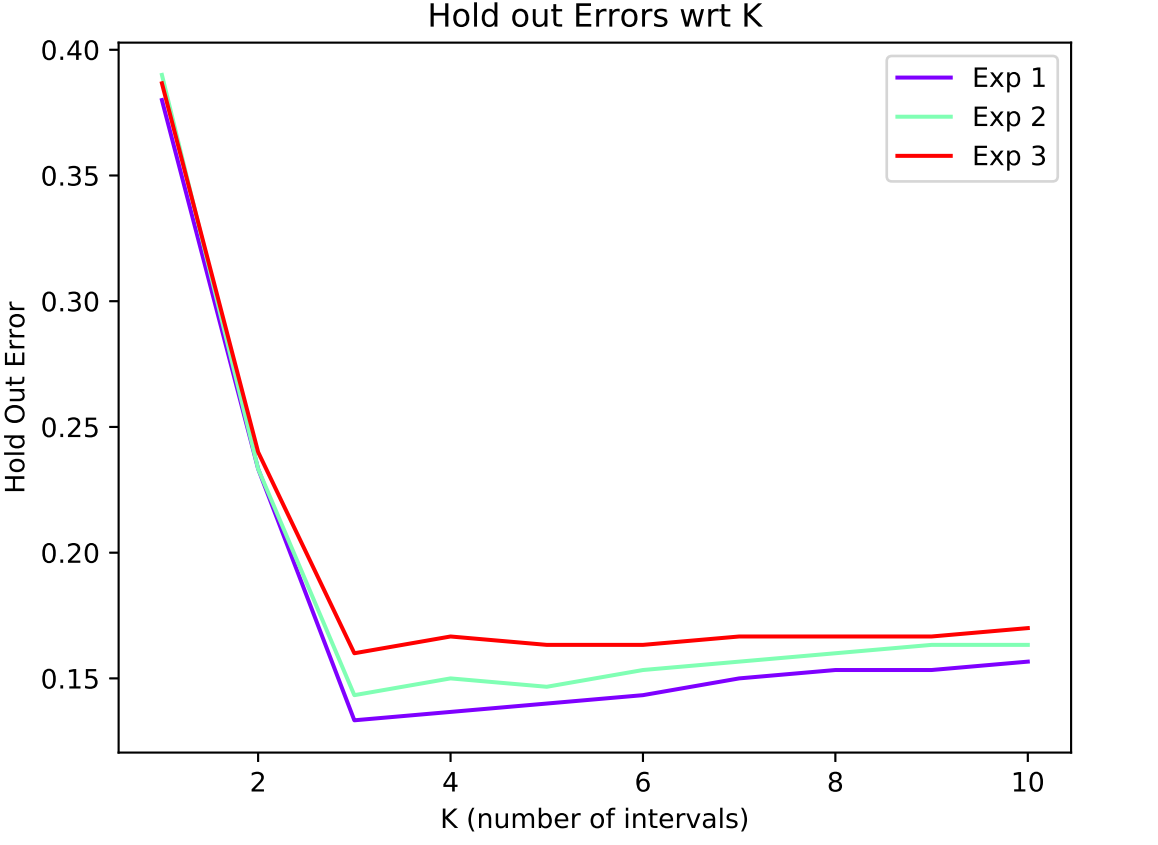
(see figure for 3 exp. results, holdout error wrt k)


Dataset Distribution and Best Intervals
1
0
Y
0.0
0.2
0.4
0.6
0.8
1.0
X


Errors wrt M
True Err
Empirical Err
Minimal True Error = 0.16
Error
M (sample size)
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
10
25
40
55
70
85
100


Errors wrt K
True Err
Empirical Err
Minimal True Err
Error
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
10
K (number of intervals)


Errors and penalty wrt K
True Err
Empirical Err
Penalty
Empirical Err + Penalty
Error
0.7
0.6
0.5
0.4
0.3
0.2
2
4
6
8
10
K (number of intervals)


Hold out Errors wrt K
Exp 1
Exp 2
Exp 3
Hold Out Error
0.40
0.35
0.30
0.25
0.20
0.15
2
4
6
8
10
K (number of intervals)