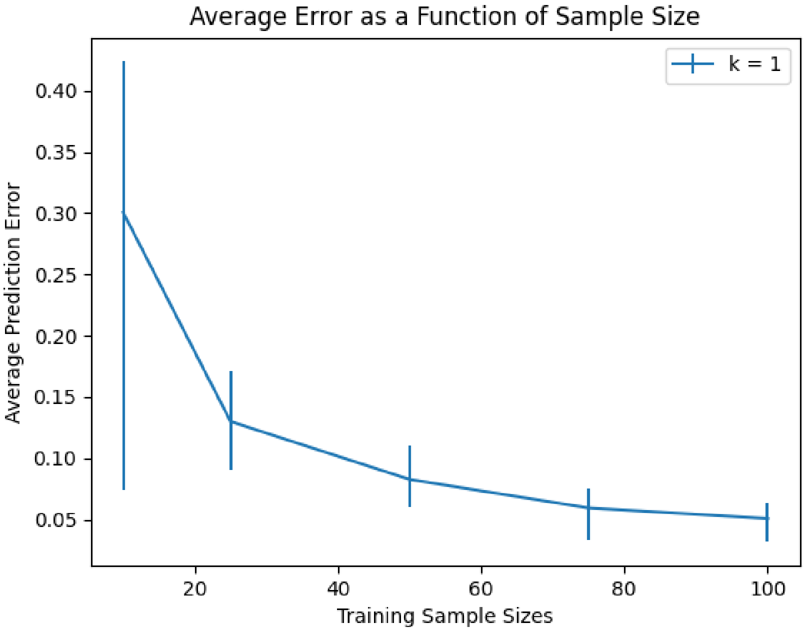
Exercise 1

Introduction to learning and analysis of big data

Nadav Shaked Kim Teldan

Question 2

a. For k = 1 and for the following sample sizes: 10, 25, 50, 75, 100, we received the graph:

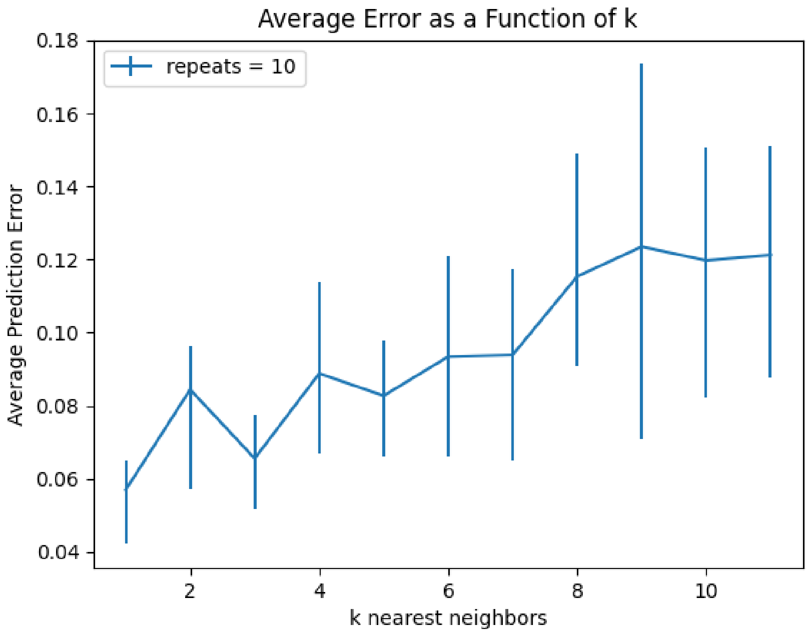


b. We observed that as the sample set size increases - the average error decreases. This trend matches our expectations - as the sample set size increases, the probability to find a closer neighbor in the sample set increases. hence the probability to find sample that resembles the tests samples, so the classifier prediction is more accurate and the average error decreases.

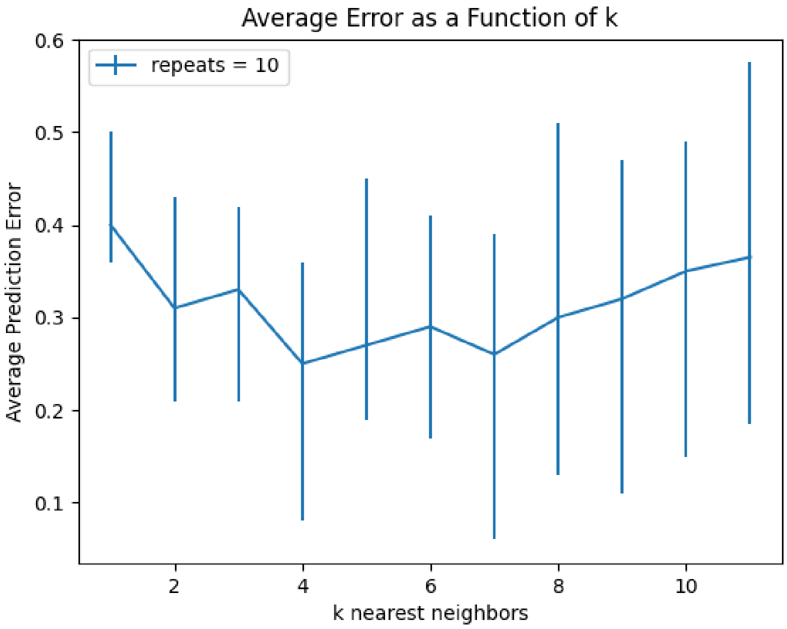
c. We received different results in different runs with the same sample size. as can be seen in the error bars, we received these results since we chose the sample tests randomly, so there are samples that are more reflective to the digits properties than others in the distribution, therefore we got different accuracies.

d. The size of the errors bar does change with the sample set size, the trend we have seen is that as the training sample set size increases, the range of the error bars decreases. As much as there are more samples, the probability to classify a label that reflective the test sample is higher, the test sample size effects the reflection of S on the distribution.

e. For k between 1 to 11 and sample size 100, we received the graph:



f. For k between 1 to 11, sample size 100 and corrupted samples by 20%, we received the graph:



g. For the first run, with the correct labels, the optimal value of k was 1. For the second run, with the corrupted labels, the optimal value of k was 4. The main difference that we observed was that the dependency of the average error in k was much more significant for the first test. For the first test we received an increasing trend. While, in the second test the graph is inconsistent, for low values of k the graph decreases and for higher values of k the graph increases.
   In the first graph, if we compare more neighbors, the Euclidean distance between some neighbors and our example is bigger, so the comparison is less accurate (since we see more neighbors we don't have to).
   In the second graph, fifth of the neighbors are misleading, so the error on smaller k’s is large. If we check more neighbors, there is smaller probability that most of them will be corrupted, Therefore we get the most decent errors around k=3, 5, 6, as these have relatively small Euclidean distance add have the expectation of only one corrupted neighbor. As k grows beyond 7, we see the same trend of increasing error, as we look at more irrelevant, distant neighbors.

Question 3

a. X should be a 2-dim vector, where the first parameter represents the student height in cm, and the second parameter represents the student age in years. Let's define X (assuming we allow only natural values for the cm and for the years as seen in the distribution table):

$$X = \{(h, a) \mid h \in \{0, 1, 2, \dots, 250\} \ \wedge \ a \in \{0, 1, 2, \dots 120\}\}$$

As for Y, it should be the student favorite movie genre

$$Y = \{Drama, Comedy\}$$

b. We will present the Bayes-optimal predictor for each $x \in X$:

$$h_{bayes} = \begin{cases} Drama & , x = (160, 20) \vee x = (180,25) \\ Comedy & , else \end{cases}$$

c. The Bayes-optimal error of D:

As seen in class, the prediction error is:

$$err\left(h_{bayes}, D\right) = \ \mathbb{P}_{(X,Y)\sim D}\left[h_{bayes}(X) \neq Y\right] \ = \sum_{x \in X} \mathbb{P}[X = x] * (1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right])$$

$$\mathbb{P}[X = (160, 20)] * (1 - \ \mathbb{P}[Y = Drama \mid X = (160,20)] = 0.13 * (1 - 1) = 0$$

$$\mathbb{P}[X = (160, 40)] * (1 - \ \mathbb{P}[Y = Comedy \mid X = (160,40)] = 0.5 * (1 - 0.6) = 0.2$$

$$\mathbb{P}[X = (180, 25)] * (1 - \ \mathbb{P}[Y = Drama \mid X = (180,25)] = 0.22 * \left(1 - \frac{17}{22}\right) = 0.05$$

$$\mathbb{P}[X = (180, 35)] * (1 - \ \mathbb{P}[Y = Comedy \mid X = (180,35)] = 0.15 * (1 - 1) = 0$$

$$\forall x \notin \{(160,20), (160,40), (180, 25), (180, 35)\}, \mathbb{P}[X = x] = 0 \implies \mathbb{P}[X = x] * \left(1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right]\right) \\ = 0$$

$$err\left(h_{bayes}, D\right) = \sum_{x \in X} \mathbb{P}[X = x] * (1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right]) = 0.25$$

d.

Height (cm)	**Favorite movie genre**	**Probability**
160	Drama	33%
160	Comedy	30%
180	Drama	17%
180	Comedy	20%

e. X is now a 1-dim vector, represent student height.

We will present the Bayes-optimal predictor for each $x \in X$:

$$h_{bayes} = \begin{cases} Drama & , x = (160, 20) \vee x = (180,25) \\ Comedy & , else \end{cases}$$

The Bayes-optimal error of D:

$$err\left(h_{bayes}, D_2\right) = \ \mathbb{P}_{(X,Y)\sim D_2}\left[h_{bayes}(X) \neq Y\right] \ = \sum_{x \in X} \mathbb{P}[X = x] * (1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right])$$

$$\mathbb{P}[X = 160] * (1 - \ \mathbb{P}[Y = Drama \mid X = 160] = 0.63 * \left(1 - \frac{11}{21}\right) = 0.3$$

$$\mathbb{P}[X = 180] * (1 - \ \mathbb{P}[Y = Comedy \mid X = 180] = 0.37 * \left(1 - \frac{20}{37}\right) = 0.17$$

$$\forall x \notin \{160, 180\}, \mathbb{P}[X = x] = 0 \implies \mathbb{P}[X = x] * \left(1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right]\right) = \ 0$$

$$err\left(h_{bayes}, D_2\right) = \sum_{x \in X} \mathbb{P}[X = x] * (1 - \ \mathbb{P}\left[Y = h_{bayes}(x) \mid X = x\right]) = 0.47$$

f. As we seen in class the expected error formula is:

$$\mathbb{E}_{S\sim G^m}\left[err\left(\hat{h}, G\right)\right] = \frac{k-1}{k}\sum_{x\in X} p_x * (1-p_x)^m$$

In this case we can use the formula for our problem since D has a deterministic label conditioned on that.

In our problem $m = 5$, $k = 2$, and G is:

Height (cm)	**Age (years)**	**Favorite movie genre**	**Probability**
160	20	Drama	20%
170	40	Comedy	30%
180	25	Drama	10%
180	35	Comedy	40%

So

$$\mathbb{E}_{S\sim G^5}\left[err\left(\hat{h}, G\right)\right] = \frac{1}{2}\sum_{x\in X} p_x * (1-p_x)^5 = \frac{1}{2} * [0.2 * 0.8^5 + 0.3 * 0.7^5 + 0.1 * 0.9^5 + 0.4 * 0.6^5]$$
$$= 0.103055$$

Question 4

a. We will prove that the 1-NN algorithm will return a function from the hypothesis class H1 for a non-conflicting sample set S.
Given S, a sample as described in the question, we will define M as follows:
$M = \{(x, y) | (x, y) \in S\}$
$M \subseteq S$ so $|S| \geq |M|$, and M is a non-conflicting set.
We will define $|M| = m$ and will sort the objects of M by the value of x for every (x, y), ascending:
$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)$
So:
$\forall i < j,\ x_i < x_j$
Now we define the following:
$$\forall 1 \leq i \leq m - 1, \qquad a_i = \frac{x_i + x_{i+1}}{2}$$
$$\forall 1 \leq i \leq m - 1, \qquad y_i = y_i$$
$$y_m = y_m$$
We will define n = m - 1 and the following function:
$$f_{n,a_1,\dots a_n,y_1,\dots y_{n+1}}(x) = \begin{cases} y_i & i\ is\ the\ smallest\ index\ such\ that\ x \leq a_i \\ y_{n+1} & x > a_n \end{cases}$$
Now we will show that the function above is the function returned by the 1-NN algorithm with distance function Δ.
Let $x \in X$. $h_S^{nn}(x) = y$. We have 2 cases:
1. if $x \geq a_{n-1}$ then $y = y_{n+1}$ and $\min_{a \in \{a_1,..,a_n\}} |x - a| = |x - a_n|$, so the nearest neighbor of x is a_n, so $h_S^{nn}(x) = y_{n+1}$.
2. else, define t to be the smallest index such that $x \leq a_i$, so $f(x) = a_t$. By the definition of a_i. We know that $\min_{a \in \{a_1,..,a_n\}} |x - a| = |x - a_t|$ so the nearest neighbor of x is a_t, so $h_S^{nn}(x) = y_t$.
Hence, $f = h_S^{nn}$ and it is what we wish to prove.
As seen in class, h_S^{nn} with distance function Δ, is an ERM algorithm for H1, and $err(h_S^{nn}, S) = 0$. And this is the minimal value we can get.

b. To prove there exists a non-conflicting sample on which the k-NN algorithm with distance function Δ does not behave like an ERM algorithm for H1, we will show such sample.
For the k-NN algorithm with k = 3, let set $S = \{(x_1, 0), (x_2, 0), (x_3, 1)\}$, h_S^{3-nn} will return 0 for every x, because the most common label is 0 for every x, and we will define $f'_{3,a_1,\dots,y_4}(x) = \begin{cases} 0 & , x \leq a_3 \\ 0 & , x > a_3 \end{cases}$
So $f' \in H_1$.
$ERM(f', S) > 0$, because $f'(x_3) = 0 \neq 1$, And as we know from the preview question, the minimal ERM value for f is 0.

Question 5

a. We saw in class that h_{bayes} function must satisfy:

$$h_{bayes}(x) \in argmax_{y \in Y} \eta_y(x)$$

Since $Y = \{0,1\}$ we can set:

$$h_{bayes}(x) = I\left[\eta(x) \geq \frac{1}{2}\right] = I\left[\alpha \geq \frac{1}{2}\right]$$

So the Bayes-optimal error is:

$$err\left(h_{bayes}, D\right) = P_{XxY\sim D}[h(x) \neq y] = \begin{cases} P_{XxY\sim D}[y = 0] & if\ \alpha \geq \frac{1}{2} \\ P_{XxY\sim D}[y = 1] & if\ \alpha < \frac{1}{2} \end{cases}$$

From the law of total probability, the integral of the whole distribution is 1, so:
Let $f(x)$ be the density function of D the $\int_0^1 f(x)dx\ = 1$.

$$P_{XxY\sim D}[y = 0] = \int_0^1 P_{XxY\sim D}[y = 0|X = x]f(x)dx = \int_0^1 (1-\alpha)f(x)dx\ =\ (1-\alpha)\int_0^1 f(x)dx$$
$$= 1 - \alpha$$

$$P_{XxY\sim D}[y = 1] = \int_0^1 P_{XxY\sim D}[y = 1|X = x]f(x)dx = \int_0^1 \alpha f(x)dx\ = \alpha \int_0^1 f(x)dx\ = \alpha$$

Therefore:

$$err\left(h_{bayes}, D\right) = \begin{cases} 1 - \alpha\ \ if\ \alpha \geq \frac{1}{2} \\ \alpha\ \ if\ \alpha < \frac{1}{2} \end{cases}\ =\ min\ (\alpha, 1-\alpha)$$

b. Notice that D is a uniform distribution the probability that $(x, y) \in S$ is 0. Let $h(x)$ be the nearest neighbor algorithm by S sample set. Notice that $f(\alpha)$ is the probability that y, the label of x is different for the label of $h(x)$, therefore:

$$\begin{aligned} f(\alpha) &= P[h(X) \neq Y|\ x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \\ &= P[(x,0),(h(x),1)|x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \\ &\quad + P[(x,1),(h(x),0)|x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \\ &= P[(h(x),1)|X = x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \\ &\qquad * P[(x,0)|\ X = x, the\ examples\ x_1, \dots, x_m\ in\ S] \\ &\qquad + P[(h(x),0)|X = x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \\ &\qquad * P[(x,1)|\ X = x\ and\ the\ examples\ x_1, \dots, x_m\ in\ S] \end{aligned}$$

$h(x)$ is independent on x and x is independent on S, thus:

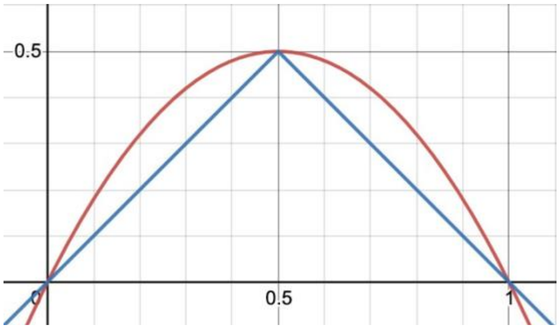
$$\begin{aligned} f(\alpha) = P[(h(x),1)|\ the\ examples\ x_1, \dots, x_m\ in\ S] * P[(x,0)\ |\ X = x] \\ +\ P[(h(x),0)|\ the\ examples\ x_1, \dots, x_m\ in\ S] * P[(x,1)\ |\ X = x] \end{aligned}$$

Since $S \sim D^m$:

$$\begin{aligned} f(\alpha) &= P[(h(x),1)|\ X = h(x)] * P[(x,0)\ |\ X = x] + P[(h(x),0)|\ X = h(x)] * P[(x,1)\ |\ X = x] \\ &= P[Y = 1|X = h(x)](1 - P[(Y = 1)|\ X = x]) \\ &\qquad + P[Y = 0|X = h(x)](1 - P[(Y = 0)|\ X = x]) \\ &= 2\alpha - 2\alpha^2 - 2\alpha^2 \end{aligned}$$

c.

$f(\alpha)$
$err(h_{bayes}, D)$



d.

$f(\alpha) = 2\alpha - 2\alpha^2$
$err(h_{bayes}, D) = \min(\alpha, 1 - \alpha)$

We want to prove that:
$err(h_{bayes}, D) \leq f(\alpha) \leq 2err(h_{bayes}, D)$
Split to cases:
$\alpha \leq \frac{1}{2}$:

$$\alpha = \min(\alpha, 1 - \alpha) \leq 2\alpha - 2\alpha^2$$
$$0 \leq \alpha(1 - 2\alpha)$$

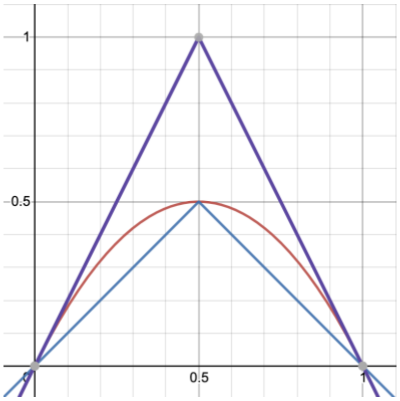
the equation is accurate for every $0 \leq \alpha \leq \frac{1}{2}$
For $\alpha \geq \frac{1}{2}$:

$$2\alpha - 2\alpha^2 \leq 2\min(\alpha, 1 - \alpha) = 2(1 - \alpha) = 2 - 2\alpha$$
$$0 \leq 2 - 2\alpha^2$$

the equation is accurate for every $\frac{1}{2} \leq \alpha \leq 1$

So

$$err(h_{bayes}, D) \leq f(\alpha) \leq 2err(h_{bayes}, D)$$



Question 6

a. For given H the problem is agnostic because there exists a distribution D as follows:
There is no $h \in H$ such that $err(h, D) = 0$. (realizable).
We will show a distribution as described above:
The deterministic function for the labeling defined as $f\big(g(x)\big) = \begin{cases} 1, & \{1\}^n \vee \{2\}^n \\ 0, & else \end{cases}$
We can notice there is more than one sequences of vertices which is labeled in 1.
In this case, by the definition of hypothesis class, there exists no hypothesis that can mark both x's in 1, since they are represented by two different sequences while n is classified in 1. Only one sequence is possible - so H is agnostic.

b. We start by calculate the size of class H.
There are n vertices, and it is known that each vertex has between 0-5 neighbors, so for every coordinate in $g(x) = v$ - the coordinate can have the values 0-5. Since there are n coordinates, We know that: $|H| = 6^n$.
As seen in class, for finite H, the agnostic PAC-learning upper bound for sample complexity is:

$$m \geq \frac{2\log(|H|) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = \frac{2\log(6^n) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2} = \frac{2\,\mathrm{n} * \log(6) + 2\log\left(\frac{2}{\delta}\right)}{\epsilon^2}$$

Also seen in class - for finite H the required sample size depends on $\log(|H|)$, therefore sample complexity is bounded by $O(n)$.

c. For give H, we will prove that VC(H) = 1 for n > 1 and for n = 1 VC(H) = 0,
For n > 1, let h such that $h_v \in H$, h will classify by 1 all the graphs that maintaining the g(x)=v, the number of neighbors that in the i-th coordinate, so there is only one combination,
an explanation for the claim:
If $v_1 = g(x_1) \neq g(x_2) = v_2$, there is no $h_v \in H$ such that $h_v(x_1) = h_v(x_2) = 1$, so the labeling must be different or 0 for both, else $g(x_1) = g(x_2)$ so for any $h_v \in H$, there is 0 possibility for x_1 labeled as 1 and x_2 labeled as 0 or the opposite way around.
So VC(H) = 1.
In case n = 1, X is set of vectors with one coordinate so |H|=1, therefore h(x)=1, so $h_v(x) = 0$ indicate that no h predicts x right thus VC(H)=0.
And as we saw in class, the sample complexity bound by:

$$m(n, \epsilon, \delta) = \Theta\left(\frac{\log\big(VC(H)\big) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right) = \Theta\left(\frac{\log(1) + \log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right) = \Theta\left(\frac{\log\left(\frac{1}{\delta}\right)}{\epsilon^2}\right)$$