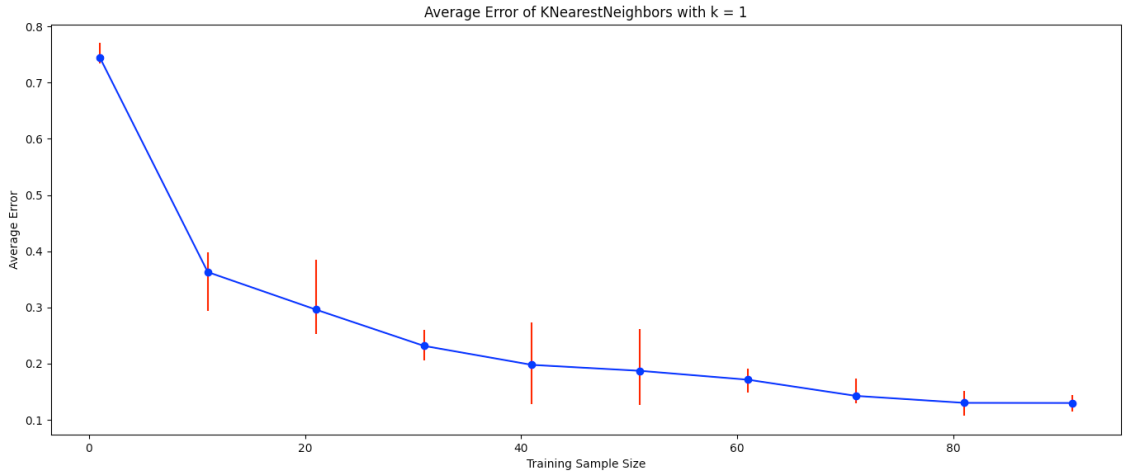
Exercise 1

Omri Bar Oz – 313325961, Ilana Pervoi – 318271640

Question 2:

a)

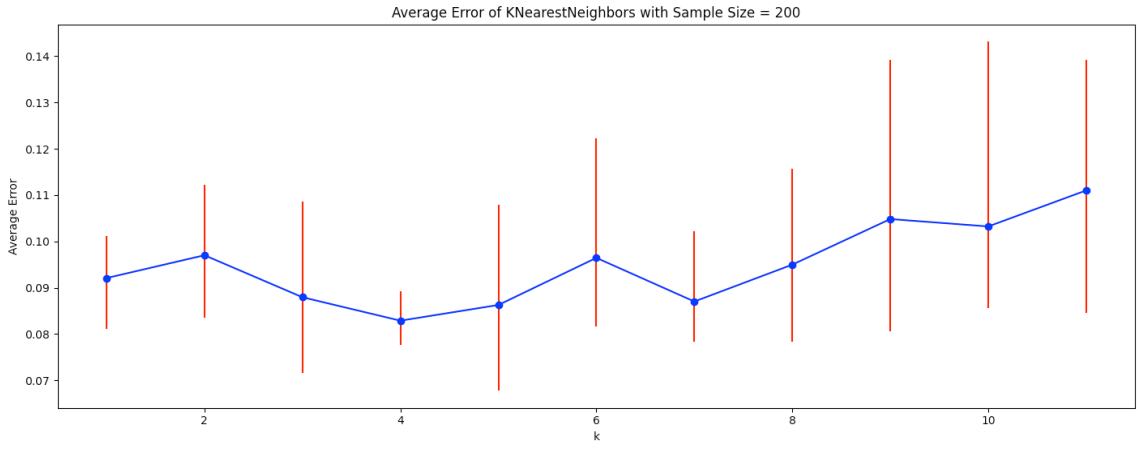


b) We can observe that when the training sample size increases, the average test error decreases. It happens because when we increase the sample size, our samples become a better representation of the distribution and each point in the test set has a higher probability to find a closer neighbor from the samples with the same label during the NN algorithm run.

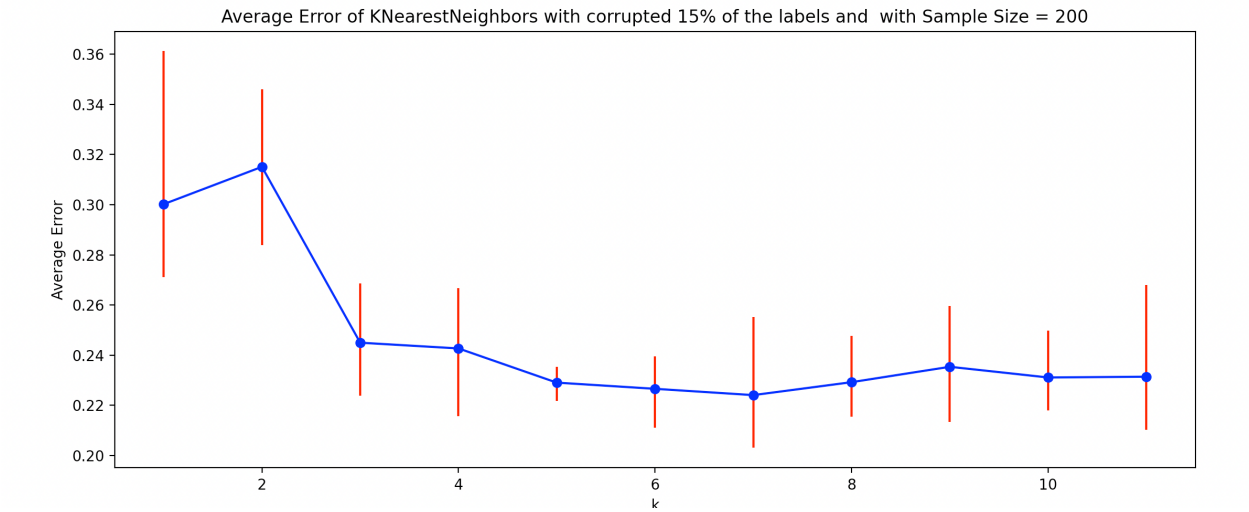
c) Yes. We get different results in different runs with the same sample size, since in each run the samples are chosen randomly, meaning they are not the same as in previous runs and therefore we might get a sample set with only 1 or 2 labels, which will cause us more errors than samples that are diverse and more representative of our distribution.

d) Yes, generally the error bars tend to decrease in size when the sample size increases. However, as we can observe there are cases in which the sample size increases, and yet the error bars increase, because as stated in section c, the accuracy of the model depends on how well the samples represent the distribution.

e)



f)



g) Without corruption the optimal value of k is 4 and with corruption it is 7. The difference between the two experiments is that in the first one, the error started low and decreased until we got the optimal k, and after the optimal k it began to increase and even got higher than the error with k=1. In the second experiment, the error starts high in comparison to the optimal k (around 15% more error) and decreases significantly at the start until it stabilizes and gets to the optimal k.
Our explanation is that in the first experiment, as we look at more neighbors, because the sample size stays the same, we look at more "bad" neighbors, which means neighbors with different labels than the real label. This explains why the error is increasing after the optimal k. In the second experiment, because we corrupted 15% of the labels, when k is low we might be more affected by the corrupted labels, while when the k rises we look at more neighbors. Hence the number of corrupted neighbors is less significant to decide the real label.

Question 3:

a) Need to prove:
for any two pairs $(x_1, y_1), (x_2, y_2) \in S$ if $y_1 \neq y_2$ then $\|x_1 - x_2\| \geq \frac{1}{c}$

Proof:
$y_1 \neq y_2$ and $y_1, y_2 \in \mathcal{Y} = \{0,1\}$ so WLOG we can decide that $y_1 = 0$ and $y_2 = 1$.
We obtain:

$$\eta(x_1) = \eta_1(x_1) = P(Y = 1|X = x_1) = 0$$
$$\eta(x_2) = \eta_1(x_2) = P(Y = 1|X = x_2) = 1$$

Recall that $x_1, x_2 \in Supp(D)$ and η of D is c-Lipschitz with respect to the Euclidean distance, hence from a property we learned in class:

$$|\eta(x_1) - \eta(x_2)| \leq c * \rho(x_1, x_2) = c \cdot \|x_1 - x_2\|$$

$$\rightarrow |0 - 1| \leq \ c \cdot \|x_1 - x_2\|$$

$$\rightarrow \frac{1}{c} \leq \|x_1 - x_2\| \ \blacksquare$$

b) Need to prove:
under the given assumptions, $err(f_S^{nn}, D) = 0$

Proof:

$$err(f_S^{nn}, D) = \sum_{x \in \chi} P(X = x) \cdot (1 - \eta_{f_S^{nn}(x)}(x))$$

So, we will show that $\forall x \in \mathcal{X},\ \eta_{f_S^{nn}(x)}(x) = 1$.

Let $x' \in \mathcal{X}$ and $f_S^{nn}(x') = y_i$ such that $(x_i, y_i) \in S$.
Since x' is in at least one ball in the set of balls of radius $\frac{1}{3c}$ that covers the space of points χ and there is an assumption that S has a point in this ball as well, we obtain

$$\|x' - x_i\| \leq 2 \cdot r = \frac{2}{3c} < \frac{1}{c}$$

Now, from the assumptions on D we know that x' and x_i has the same label

$$\rightarrow \eta_{f_S^{nn}(x')}(x') = P(Y = f_S^{nn}(x') | X = x') = P(Y = y_i | X = x') = 1 \ \blacksquare$$

Question 4:

a)

$$\mathcal{Y} = \{black, white\}$$

$$\mathcal{X} = \{(x_1, x_2) \in R^2 : 0 \leq x_1 \leq 48, \quad 0 \leq x_2 \leq 4\}$$

b)

$$h_{bayes}\big((5,2)\big) = white$$

$$h_{bayes}\big((12,1)\big) = black$$

$$h_{bayes}\big((12,2)\big) = white$$

c) $err\big(h_{bayes}, D\big) = P_{(X,Y)\sim D}\big[h_{bayes}(X) \neq Y\big] = \sum_{(x,y)\in X\times Y\,, h_{bayes}(x)\neq y} P[X = x, Y = y]$
$= P[X = (5,2)\,, Y = black] + P[X = (12,1)\,, Y = white] = 0.08 + 0.04 = 0.12$

d) Let $H: \mathcal{X} \to \mathcal{Y}$ hypothesis class of constant functions. $\mathcal{Y} = \{black, white\}$ therefore $H = \{h_{black}, h_{white}\}$ where $\forall x \in \mathcal{X}, h_{black}(x) = black, h_{white}(x) = white$. By definition: $err_{app} = \inf_{h\in H} err(h, D) = min\{err(h_{black}, D), err(h_{white}, D)\}$.

$$err(h_{black}, D) = P\big((5,2), white\big) + P\big((12,1), white\big) + P\big((12,2), white\big)$$
$$= 0.47 + 0.04 + 0.21 = 0.72$$
$$err(h_{white}, D) = P\big((5,2), black\big) + P\big((12,1), black\big) = 0.08 + 0.2 = 0.28$$

Hence: $err_{app} = min\{err(h_{black}, D), err(h_{white}, D)\} = \min\{0.72, 0.28\} = 0.28$

e) As we say in class, every example in the $\mathcal{X}$ will be labeled by h_{bayes} with the label with the highest probability. In this case, if a rabbit is older than 25 months old, it has a probability of 50% or more to be black, so h_{bayes} will label it as black. More formally:

$$\forall x \in \mathcal{X}, h_{bayes}(x) = \begin{cases} black, & x.age \geq 25\ (monts) \\ white, & else \end{cases}$$

f) We cannot calculate the error of the predictor we gave in e), since we don't know how the rabbits ages are distributed, and to calculate the error we need to know $P(x.age \geq 25)$ where x is a rabbit from $\mathcal{X}$.

g) As we saw in class, where $k = |\mathcal{Y}|$, $p_x = \mathrm{P}_{(\mathrm{X,Y})\sim \mathrm{D}}[\mathrm{X} = \mathrm{x}]$ and m is the sample size:

$$E_{S\sim D^m}\left[err\left(\hat{h}_s, D\right)\right] = \frac{k-1}{k} \cdot \sum_{x\in\mathcal{X}} p_x(1-p_x)^m$$

in our case: $k = 2, m = 5, p_{(5,2)} = 0.06, p_{(5,3)} = 0.12, p_{(7,1)} = 0.53, p_{(9,4)} = 0.29$ therefore:

$$E_{S\sim D^m}\left[err\left(\hat{h}_s, D\right)\right]$$
$$= \frac{2-1}{2}$$
$$\cdot (0.06 \cdot (1-0.06)^5 + 0.12 \cdot (1-0.12)^5 + 0.53 \cdot (1-0.53)^5 + 0.29$$
$$\cdot (1-0.29)^5) = 0.08592$$

We are allowed to use this formula for D'' because D'' has a deterministic label condition on the example, whereas D does no as it has 2 different labels for the same example.

Question 5:

Let $\mathcal{X} = [0,1], \mathcal{Y} = \{0,1\}$. Consider the hypothesis class of thresholds:

$$H_{th} = \{f_a \mid a \in [0,1], where\ f_a(x) \coloneqq \mathbb{I}[x \geq a]\}$$

Let D be a distribution over $\mathcal{X} \times \mathcal{Y}$ and suppose that the marginal distribution of D on $\mathcal{X}$ is uniform on $[0,1]$ and that for $(X,Y) \sim D$, for any $x \in \mathcal{X}, P(Y = 1 \mid X = x) = \mathbb{I}[x \geq \beta]$

a) In this section we will talk about N evenly-spaced threshold for an integer N:

$$H_{th}^N \coloneqq \left\{ f_a :\ a \in \left\{\frac{i}{N}\right\}_{i \in \{0,\dots,N\}}, where\ f_a(x) \coloneqq \mathbb{I}[x \geq a] \right\}$$

We assume that D is realizable by H_{th}^N. We will use the PAC bounds for finite hypothesis classes we learned in class with $\delta = 0.05, \epsilon = 0.03$. We can use it as H_{th}^N is finite and $|H| = N + 1$, and D is realizable by it and we get:

$$m \geq \frac{\log(|H|) + \log\left(\frac{1}{\delta}\right)}{\epsilon} = \frac{\log(N+1) + \log\left(\frac{1}{0.05}\right)}{0.03}$$

b) Let $\epsilon \in [0,1]$. Let $a \in [\beta - \epsilon, \beta + \epsilon]$.

$$err(f_a, D) = P_{(X,Y)\sim D}(f_a(X) \neq Y) = P(X \geq a \wedge Y = 0) \overset{\text{איחוד מאורעות זרים}}{+} P(X < a \wedge Y = 1)$$

by the definition, an example will get 1 only if it is higher or equal to β and 0 only if it is lower than β. Therefore:

$$= P(\beta > x \geq a\ \wedge\ Y = 0) + P(\beta \leq x < a\ \wedge\ Y = 1) =$$

$$\overset{wlog\ a > \beta}{=} 0 + P(\beta \leq x < a \wedge Y = 1) = P(\beta \leq x < a) =$$

$$\overset{X\ is\ uniform}{=} a - \beta \overset{a \leq \beta + \epsilon}{\leq} \beta + \epsilon - \beta = \epsilon\ \blacksquare$$

c) Let $S \sim D^m$ with $(x_1, y_1), (x_2, y_2) \in S : x_1 \in [\beta - \epsilon, \beta],\ x_2 \in [\beta, \beta + \epsilon],\ \nexists\, (x', y') \in S :\ x_1 < x' < x_2$. Let $\hat{h}_s \in H_{th}$ be the classifier returned from some ERM algorithm trained on S with H_{th}. We will show that $\hat{h}_s = f_a$ for $a \in [x_1, x_2] \subseteq [\beta - \epsilon, \beta + \epsilon]$ by showing that $err(f_a, S) = 0$ and that for every $a' \notin [\beta - \epsilon, \beta + \epsilon],\ err(f_{a'}, S) > 0$.

Let $a_1 \in [x_1, x_2]$. Then $f_{a_1} = \mathbb{I}[x \geq a_1]$. Notice that $\forall (x,y) \in S: x \geq x_2\ or\ x \leq x_1$ and equality happens only if $x = x_2\ or\ x = x_1$, by the definition of x_1, x_2.

So, for every $x \geq x_2 \geq a_1,\ f_{a_1}(x) = 1$ and for every $x \leq x_1 \leq a_1, f_{a_1}(x) = 0$. Because $x_2 \geq \beta$ and $x_1 < \beta$ it holds that $y_1 = 0,\ y_2 = 1$ and it implies that f_{a_1} is labeling correctly all the points in S which implies that $err(f_{a_1}, S) = 0$.

Let $a_2 \in [0, \beta - \epsilon) \cup (\beta + \epsilon, 1]$. Wlog assume that $a_2 \in (\beta + \epsilon, 1]$.

It is easy to see that $a_2 > x_2 \geq \beta$. While x_2 true label is 1, $f_{a_2} = \mathbb{I}[x \geq a_2]$ will label x_2 as 0, which implies that $err(f_{a_2}, S) > 0$.

This concludes that an ERM algorithm will return a f_a for $a \in [\beta - \epsilon, \beta + \epsilon]$.

Let f_a be the hypothesis returned by an ERM algorithm. then by sub-question (b),

$$err(f_a, D) \leq \epsilon\ \blacksquare$$

d) Let $S \sim D^m$. Then:

$$P(\nexists (x,y) \in S : x \in [\beta, \beta + \epsilon]) = P(\forall (x,y) \in S : x \notin [\beta, \beta + \epsilon]) =$$

$$P(x_1 \notin [\beta, \beta+\epsilon], \dots, x_m \notin [\beta, \beta+\epsilon]) =$$
$$\overset{x_i \text{ are i.i.d}}{=} \prod_{i=1}^{m} P(x_i \notin [\beta, \beta+\epsilon]) =$$
$$\overset{uniform}{=} \prod_{i=1}^{m} (1 - (\beta + \epsilon - \beta) = \prod_{i=1}^{m} (1-\epsilon) = (1-\epsilon)^m$$

this holds for $x \in [\beta - \epsilon, \beta]$ as well because

$$P(x_i \notin [\beta-\epsilon, \beta]) = 1 - \big(\beta - (\beta - \epsilon)\big) = 1 - \epsilon = 1 - (\beta + \epsilon - \beta)$$
$$= P(x_i \notin [\beta, \beta+\epsilon]) \blacksquare$$

Let $\hat{h}_s$ be the hypothesis returned by a ERM algorithm with H_{th}.

$$P_{S\sim D^m}\left[err\left(\hat{h}_s, D\right) \le \epsilon\right] \overset{sub-question\ (c)}{=} P(\exists (x_1, y_1) \in S : x_1 \in [\beta-\epsilon, \beta] \wedge \exists (x_2, y_2) \in S : x_2 \in [\beta, \beta+\epsilon])$$
$$= 1 - P(\nexists (x_1, y_1) \in S : x_1 \in [\beta-\epsilon, \beta] \vee \nexists (x_2, y_2) \in S : x_2 \in [\beta, \beta+\epsilon])$$
$$\overset{union\ bound}{\ge} 1 - \left(\overbrace{P(\nexists (x_1, y_1) \in S : x_1 \in [\beta-\epsilon, \beta])}^{(1-\epsilon)^m} + \overbrace{P(\nexists (x_2, y_2) \in S : x_2 \in [\beta, \beta+\epsilon])}^{(1-\epsilon)^m}\right)$$
$$= 1 - 2 \cdot (1-\epsilon)^m \blacksquare$$

e) We will use $\delta = 0.05, \epsilon = 0.03$. by the last sub-question:

$$P_{S\sim D^m}\left[err\left(\hat{h}_s, D\right) \le \epsilon\right] \ge 1 - 2 \cdot (1-\epsilon)^m$$

we want it to be greater than $1 - \delta$ then:

$$P_{S\sim D^m}\left[err\left(\hat{h}_s, D\right) \le \epsilon\right] \ge 1 - 2 \cdot (1-\epsilon)^m \ge 1 - \delta$$
$$1 - 2 \cdot (1-\epsilon)^m \ge 1 - \delta$$
$$\rightarrow \frac{\delta}{2} \ge (1-\epsilon)^m$$
$$\rightarrow \log\left(\frac{\delta}{2}\right) \ge m log(1-\epsilon)$$
$$\overset{\log(1-\epsilon)<0}{\Rightarrow} \frac{\log\left(\frac{\delta}{2}\right)}{\log(1-\epsilon)} \le m$$
$$\rightarrow \frac{\log\left(\frac{0.03}{2}\right)}{\log(1-0.03)} \le m$$
$$\rightarrow 121.109 \le m$$

f) It's better that we use the approach we took in sub-question (e).
In sub-question (a) we assumed the distribution is realizable by H_{th}^N and that we know for which N. Now, we assume that there is a N such that it suffices the assumptions for sub-question (a), but we cannot find this N as we have only the samples to verify it and we don't know the distribution. Thus, we cannot use sub-question (a) bound as we don't know the N

that suffices the assumptions.
Moreover, even if we knew the N that suffices the assumptions in sub-question (a), it would still be better using sub-question's (e) bound.
In sub-question (e) we got that the sample size needs to be at least 121.109 to assure with 95% error of at most 3% on the distribution. With the same parameters in sub-question (a), we get that it should be at least $\frac{\log(N+1)+\log\left(\frac{1}{0.05}\right)}{0.03}$. Let's check what the value of N should be, so m would less than 121.109:

$$\frac{\log(N+1)+\log\left(\frac{1}{0.05}\right)}{0.03} \leq 121.109 \rightarrow$$
$$\rightarrow \log(N+1) \leq 3.63327 - \log\left(\frac{1}{0.05}\right) = -0.6886 \rightarrow$$
$$N+1 \leq 2^{-0.6886} = 0.62 \rightarrow$$
$$N \leq -0.38 < 0$$

Which means that we can't use any hypothesis class H_{th}^{N} to match the sample size of sub-question (e), then it is better to use the approach in sub-question (e).