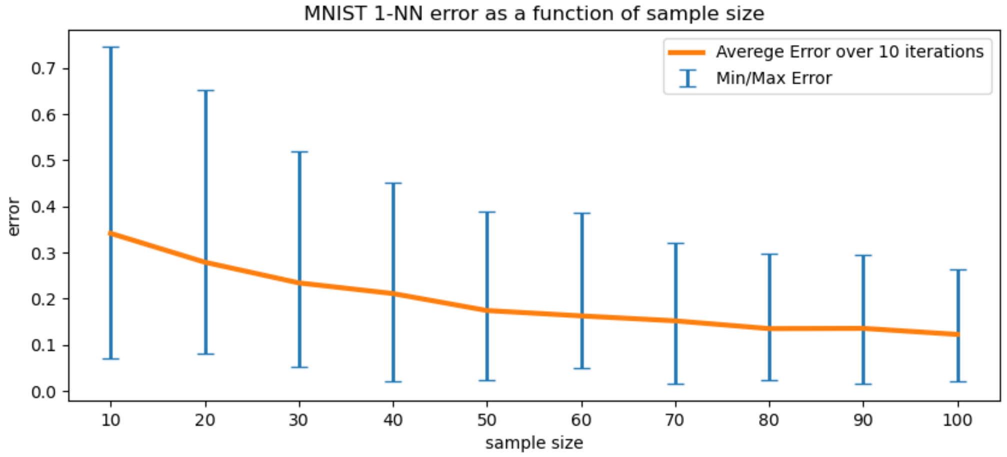
Question 2:

a.



b.

yes. The average error over 10 random sample decreases as the sample size increase This trend stems from the fact the when supplied with more examples and data, NN algorithm can make more precise generalization for the distribution based on the sample and thus generate a better rule.

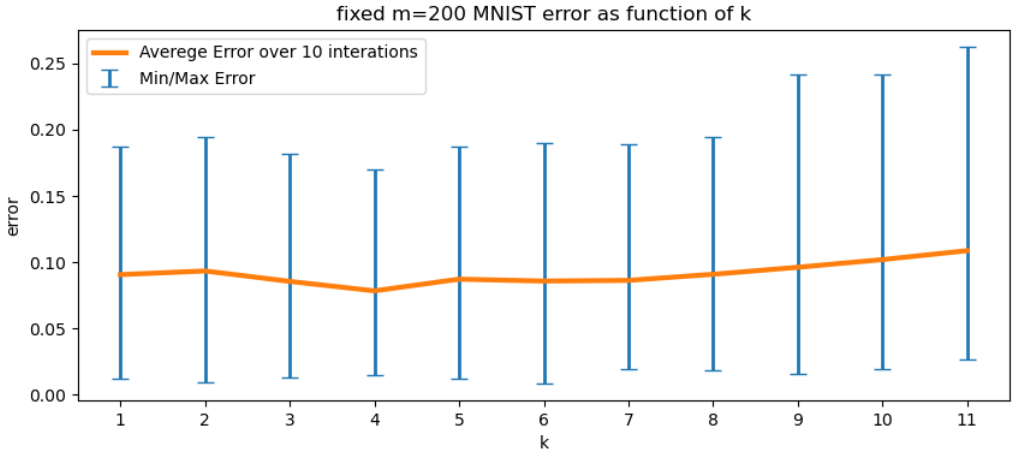
c.

yes. we can see the difference between maximum and minimum error for each fixed sample size. this is because the algorithm learns and then tested each time on different samples, so the prediction rule it generates changes.

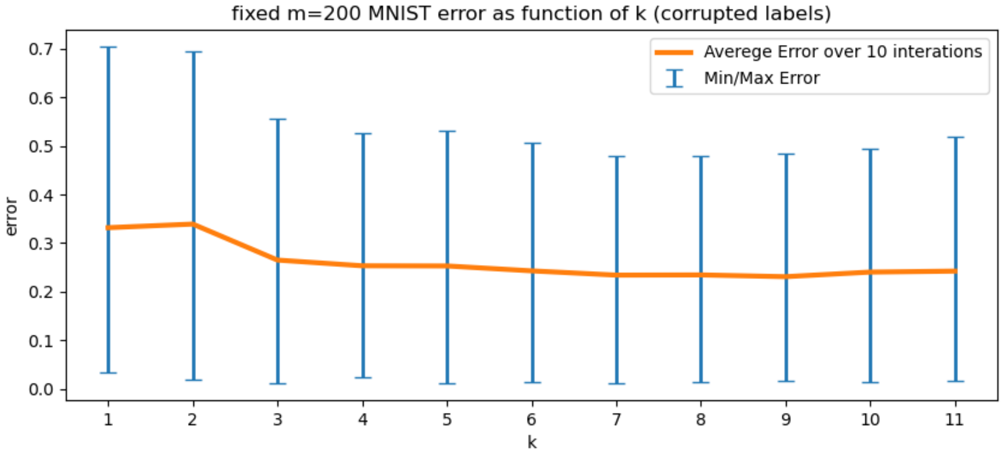
d.

yes. We can see the gap between the maximum and minimum error shrinks as the sample size grows. When the algorithm sees more examples, it can generate a better generalization on the distribution and even for special cases makes smaller errors.

e.



f.



g.

The optimal k value for the first experiment is 4, and for the second is 7.

the main differences between the experiments are: In the **first experiments** with correct labels, we can see that both average error and the difference between maximum and minimum error grows with k>4, where in the **second experiment** it is getting smaller as k grows to 3 and stables. it can be explained by the fact that when k gets larger the label is decided based on more neighbors, thus in the first experiment it causing the consideration of irrelevant data, but in the second experiment considering more data makes up for using corrupted labels.

Question 3:

a.

proof:

with the assumption the D has a Bayes-error of zero we know that there's $h^* \in \mathrm{X} \to \{0,1\}$ that for every input $x \in \mathrm{X}$ the output of h^* will be the right label for x.

we know η of D is c-Lipschitz with respect to the Euclidean distance.

therefore, by definition $|\eta(x_1 - \eta(x_2)| \leq c * ||x_1 - x_2||$

Lemma:

$y_1 \neq y_2$, WLOG $y_1 = 1, y_2 = 0$ and D has Bayes-error of zero and therefore D has deterministic labels $\Rightarrow (\eta(x_1) = 0 \; and \; \eta(x_2) = 1) \; or \; (\eta(x_1) = 1 \; and \; \eta(x_2) = 0) \Rightarrow |\eta(x_1) - \eta(x_2)| = 1$

main proof:

$$\left||x_1 - x_2|\right| = \rho(x_1, x_2)$$

$$\Rightarrow c * \left||x_1 - x_2|\right| = c * \rho(x_1, x_2) \underset{by\ c-Lipschitz\ defintion}{\geq} |\eta(x_1) - \eta(x_2)|$$

$$\Rightarrow \left||x_1 - x_2|\right| \geq \frac{|\eta(x_1) - \eta(x_2)|}{c} \underset{by\ Lemma}{=} \frac{1}{c}$$

b.

Suppose $\boldsymbol{err(f_S^{nn}, D) > 0} \Rightarrow$ for some $\boldsymbol{x} \in \mathbf{X}$ there's $(\boldsymbol{x'}, \boldsymbol{y}) \in \boldsymbol{S}$ such as $\boldsymbol{x'} \in \boldsymbol{\pi(x)}$ and

$\boldsymbol{f_S^{nn}(x) \neq f_S^{nn}(x')}$.

$(\boldsymbol{\pi(x)} = \{\boldsymbol{nearest\ neigbors\ of\ x}\})$

We know that $\mathbb{B}$ covers the space of points in $\mathbf{X}$ in balls with radius of $\frac{\mathbf{1}}{\mathbf{3c}}$ such that for every ball $\boldsymbol{B} \in \mathbb{B}$ there is some pair $(\boldsymbol{x'}, \boldsymbol{y}) \in \boldsymbol{S}$ which satisfies $\boldsymbol{x'} \in \boldsymbol{B}$.

$\boldsymbol{\pi(x) = x'}$ so $\boldsymbol{\rho(x, x') \leq \frac{1}{3c}}$

We also know that D is c-Lipschitz with respect to the Euclidean distance.

So $\boldsymbol{\rho(x, x') \leq \frac{1}{3c} \leq \frac{1}{c}}$ and $\boldsymbol{f_S^{nn}(x) \neq f_S^{nn}(x')}$ oppose to c-Lipschitz that

The support of D does not include two points with different
labels that are less than $\frac{1}{c}$-far.

Question 4:

a.

we can represent each rabbit as vector in $\mathbb{R}^2$, since we only consider the parameters age and weight. it's also known rabbits are limited to live 48 month and weigh up to 4kg, hence $X \coloneqq [0,48] \times [0,4]$.

we would like to predict if rabbit is black or white so we can let $Y \coloneqq \{black, white\}$.

b.

$$h_{bayes} = \begin{cases} black, x = (12,1) \\ white, x \in \{(5,2), (12,2)\} \end{cases}$$

c.

$$err(h, D) = \mathbb{P}_{X,Y\sim D}[h(X) \neq Y] = \sum_{(x,y)\in X\times Y : h_{bayes}(x)\neq y} p(X = x \wedge Y = y) =$$

$$p(X = (5,2) \wedge Y = black) + p(X = (12,1) \wedge Y = white) = 0.08 + 0.04 = 0.12$$

d.

let $h_{black}: X \to Y$ and $h_{white}: X \to Y$ be hypothesis such that:

$h_{black}(x) = black\ \forall x \in X,\ h_{white}(x) = white\ \forall x \in X$

Note that since $|Y| = 2$, there are only 3 H possible: $\{h_{black}\}, \{h_{white}\}, \{h_{black}, h_{white}\}$

if $H = \{h_{black}\}$ then $\inf_{\mathrm{h}\in\mathrm{H}} err(h, D) = err(h_{black}, D) = p\big(y = black \big| x = (5{,}2)\big) + p\big(y = black \big| x = (12{,}2)\big) = 0.47 + 0.21 = 0.68$

if $H = \{h_{white}\}$ then $\inf_{\mathrm{h}\in\mathrm{H}} err(h, D) = err(h_{white}, D) = p\big(y = white \big| x = (12, 1)\big) = 0.20$

if $H = \{h_{white},\ h_{black}\}$ then $\inf_{\mathrm{h}\in\mathrm{H}} err(h, D) = \ err(h_{white}, D) = 0.20$

e.

$$h_{bayes} = \begin{cases} black & x = (x_1, x_2) \colon 25 < x_1 \leq 48 \\ white & \ \ x = (x_1, x_2) \colon 1 \leq x_1 \leq 25 \end{cases}$$

f.

$$err(h, D) = \sum_{((a,w),y)\in X\times Y : h_{bayes(x)} \neq y,\ a\leq 25} p(y = black | x = (a, w)) \ + \ \sum_{((a,w),y)\in X\times Y : h_{bayes(x)} \neq y,\ a>25} p(y = white | x = (a, w)) =$$

$$\sum_{i=1}^{25} 0.02 * i + \sum_{i=26}^{48} (1 - 0.02 * i) = 6.5 + 5.88 = 12.38$$

g. $\mathbb{E}_{S\sim D^m}\big[err\big(\hat{h}_s, D\big)\big] = \frac{k-1}{k} \sum_{\mathrm{x}\in\chi} p_x (1 - P_x)^m =$

$\frac{1}{2}(0.06 * 0.94^5 + \ 0.12 * 0.88^5 + 0.53 * 0.47^5 + 0.29 * 0.71^5) \approx 0.0883641$

The formula we learned in class assumed that D has a deterministic label conditioned on the example, and the distribution D is nondeterministic.

Question 5:

a.

sample complexity ($\epsilon = 0.03, \delta = 0.05$):

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

b.

suppose $\epsilon \in (0,1), a \in [\beta - \epsilon, \beta + \epsilon]$

$err(f_\alpha, D) = P_{(X,Y)\sim D}[f_a(x) \neq y]$,

That probability has 2 cases:

I. $f_a(x) = 1 \; and \; y = 0 \Rightarrow p(x \geq a \wedge x < \beta) = p(a \leq x < \beta)$:

$$p(a \leq x < \beta) \leq p(\beta - \epsilon \leq x < \beta) \overset{*}{=} \frac{\beta - \beta + \epsilon}{1 - 0} = \epsilon$$

II. $f_a(x) = 0 \; and \; y = 1 \Rightarrow p(x < a \wedge \; x \geq \beta) = p(\beta \leq x < a)$:

$$p(\beta \leq x < a) \leq p(\beta \leq x < \beta + \epsilon) \overset{*}{=} \frac{\beta + \epsilon - \beta}{1 - 0} = \epsilon$$

*- the marginal distribution of D on χ is uniform on [0,1]

We got that for both cases $err(f_a, D) \leq \epsilon$

c.

suppose we run some ERM algorithm and the output classifier $\hat{h}_s$ has returned,

let $(x_1, y_1), (x_2, y_2) \in S$ such that $x_1 \in [\beta - \epsilon, \beta] \; and \; x_2 \in [\beta, \beta + \epsilon]$ we know β is the threshold and for that $y_1 = 0, y_2 = 1$

<u>Lemma:</u> $\hat{h}(x_1) = 0 \; and \; \hat{h}(x_2) = 1$

proof: assume in contradiction $\hat{h}(x_1) = 1$ or $\hat{h}(x_2) = 0$ then $\hat{h}$ is not the prediction rule that minimizes $err(h, S)$ because we know that $f_\beta(x_1) = 0$ and $f_\beta(x_2) = 1$ oppose to that $\hat{h}$ is the output of the ERM algorithm

$\hat{h}(x_1) = 0 \; and \; \hat{h}(x_2) = 1$ so there's some $a \in (x_1, x_2]$ such that $\hat{h} = f_a$

also, $\beta - \epsilon \leq x_1 \leq \beta \leq x_2 \leq \beta + \epsilon$.

hence $a \in [\beta - \epsilon, \beta + \epsilon]$ and from previous section we conclude $err\left(\hat{h}, D\right) = err(f_a, D) \leq \epsilon$

d.

First, the probability that there does not exist a $(x, y) \in S$ such that $x \in [\beta, \beta + \epsilon]$ is $(1 - \epsilon)^m$

proof:

as described above the marginal distribution of D on χ is uniform on [0,1] so

$\mathbb{P}[x_i \notin [\beta, \beta + \epsilon] = 1 - \mathbb{P}\big[x_i \in [\beta, \beta + \epsilon]\big] = 1 - \dfrac{\beta + \epsilon - \beta}{1 - 0} = 1 - \epsilon$

the distribution there does not exist a $(x, y) \in S$ such that $x \in [\beta, \beta + \epsilon]$ is

$\Pi_{i=1}^{m} \mathbb{P}\big[x_i \notin [\beta, \beta + \epsilon]\big] = (1 - \epsilon)^m.$

note that we only used the length of the interval hence the proof applies for $x \in [\beta - \epsilon, \beta]$ as well.

let $A := x_1, .., x_m \notin [\beta - \epsilon, \beta\ ], B \coloneqq x_1, .., x_m\ \notin [\beta, \beta + \epsilon]$

from the sub-question we know that $p(A) = p(B) = (1 - \epsilon\ )^m$.

In section (c) we proved that if there exists some $(x_1, y_1), (x_2, y_2) \in S$ such that $x_1 \in [\beta - \epsilon, \beta] \wedge x_2 \in [\beta, \beta + \epsilon]$ then $err\big(\hat{h}_s, D\big) \leq \epsilon \Rightarrow \bar{A} \cap \bar{B}$

$$\mathbb{P}_{S\sim D^m}\big[err\big(\hat{h}_s, D\ \big) \leq \epsilon\big] =$$

$$p(\bar{A} \cap \bar{B}) = 1 - p(A \cup B) \underset{union\ bound}{\geq} 1 - \big(p(A) + p(B)\big) = 1 - 2(1 - \epsilon)^m$$

e.

let $\epsilon \coloneqq 0.03\ and\ 1 - \ 2(1 - 0.03)^m = 0.95$

for $1 - 0.05 = \ 1 - \ 2(1 - 0.03)^m$ we get $m = 122$

$$\mathbb{P}_{S\sim D^m}\big[err\big(\hat{h}_s, D\ \big) \leq \epsilon\big] = \mathbb{P}_{S\sim D^m}\big[err\big(\hat{h}_s, D\ \big) \leq 0.03\big] \underset{section\ (d)}{\geq} 1 - 2(1 - \epsilon)^m$$

$= \ 1 - \ 2(1 - 0.03)^m \Rightarrow m = 122$

f.

The approach for bounding the sample we used in (e) is better.

In (a) the bottom bound is dependent on $|H|$ (as log(N+1) is part of the calculation) whereas in the approach in (e) there's no such dependency.

let us calculate for which N the approach in (e) is better:

$$\frac{\log(N+1) + \log\left(\frac{1}{0.05}\right)}{0.03} \geq 122$$

$$\log(N+1) + \log\left(\frac{1}{0.05}\right) \geq 122 * 0.03 = 3.66$$

$$\log(N+1) \geq 3.66 - \log\left(\frac{1}{0.05}\right)$$

$$\log(N+1) \geq 2.36$$

$$N + 1 \geq 10^{2.36}$$

$$N \geq 10^{2.36} - 1$$

Thus, for such N we get a better sample bounding by the approach in (e)