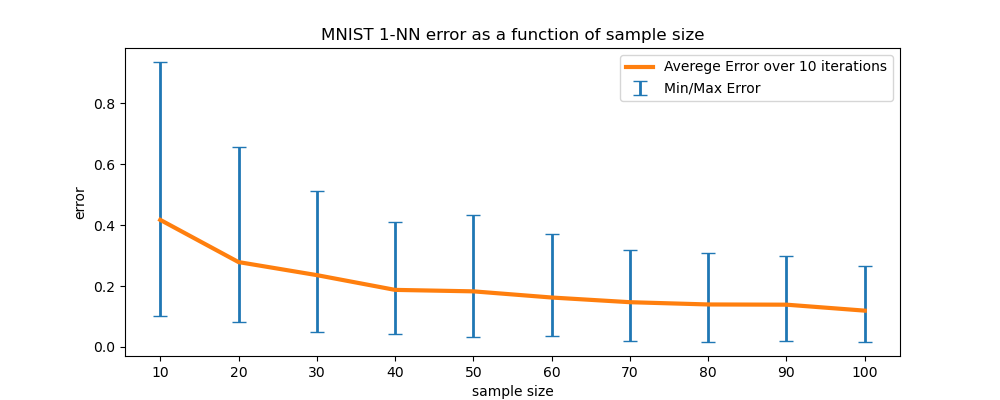
**Question 2:**

**a.**



**b.**

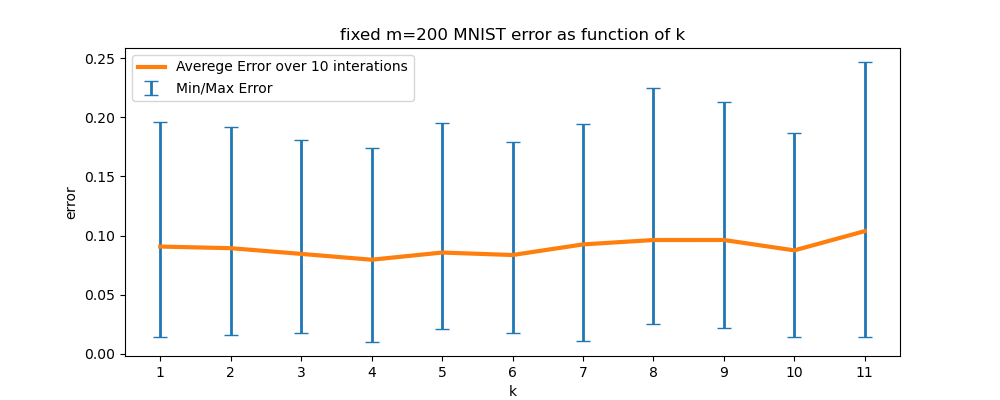
yes. The average error over 10 random sample decreases as the sample size increase This trend stems from the fact the when supplied with more examples and data, NN algorithm can make more precise generalization for the distribution based on the sample and thus generate a better rule.

**c.**

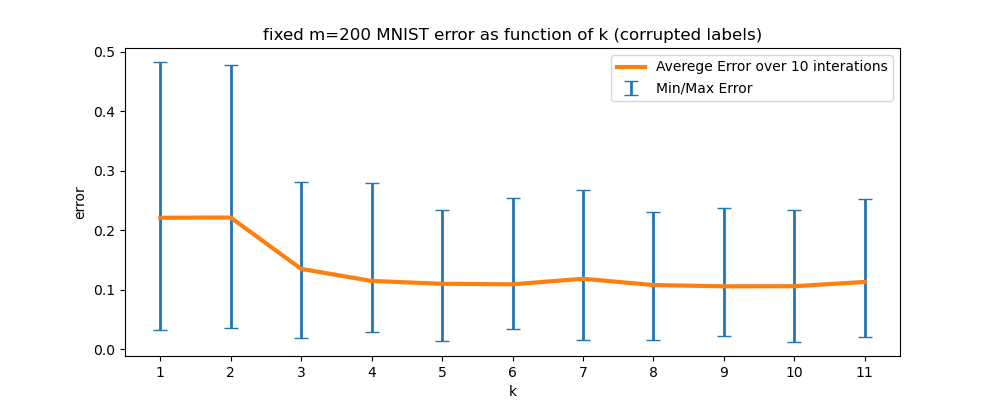
yes. we can see the difference between maximum and minimum error for each fixed sample size. this is because the algorithm learns and then tested each time on different samples, so the prediction rule it generates changes.

**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.**

value of k for each experiment? Is there a difference between the two experiments? How do you explain it?

The optimal k value for the first experiment is 4, and for the second is 5.

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, where in the **second experiment** it is getting smaller as k grows to 3-6. it can be explained by the fact that when k gets larger the label is decided based on irrelevant neighbors.

the same phenomenon is causing the second experiment the opposite effect, because considering more neighbors can compensate for the corrupted labels

**Question 3:**

**a.**

proof:

with the assumption the D has a Bayes-error of zero we know that there’s that for every input the output of will be the right label for .

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

therefore, by definition

Lemma:

, WLOG and D has Bayes-error of zero and therefore D has deterministic labels

main proof:

**b.**

Suppose for some there’s such as and

.

)

We know that covers the space of points in in balls with radius of such that for every ball there is some pair which satisfies .

so

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

So and oppose to c-Lipschitz that

The support of D does not include two points with different  
labels that are less than -far.

**Question 4:**

**a.**

we can represent each rabbit as vector in 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 .

we would like to predict if rabbit is black or white so we can let .

**b.**

**c.**

**d.**

let be hypothesis such that:

Note that since there are only possible:

if then

if then

if then

**e.**

**f.**

=

g.

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

**b.**

suppose

. also, we know that thus

then:

**c.**

suppose we run some ERM algorithm and the output classifier has returned,

let such that we know is the threshold and for that

Lemma:

*proof:* assume in contradictionthenis not the prediction rule thatminimizesbecause we know thatoppose to thatis theoutput of the ERM algorithm

so there’s some such that

also, .

hence and from previous section we conclude

**d.**

First, the probability that there does not exist a such that

proof:

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

the distribution there does not exist a such that is

.

note that we only used the length of the interval hence the proof applies for as well.

let

from the sub-question we know that

In section (c) we proved that if there exists some such that then

**e.**

let

for we get

**f.**

Lemma 1:

proof:

assume .

for some .

we will define an equivalent linear predicator

(We will use the labels instead of shown in class)

let and

we will show

as defined in class:

Lemma 2:

proof:

let be the largest set of examples that can be labeled in all possible label combinations using hypothesis from (i.e. ) ;

main proof:

based on lemma1 and lemma2 we get that