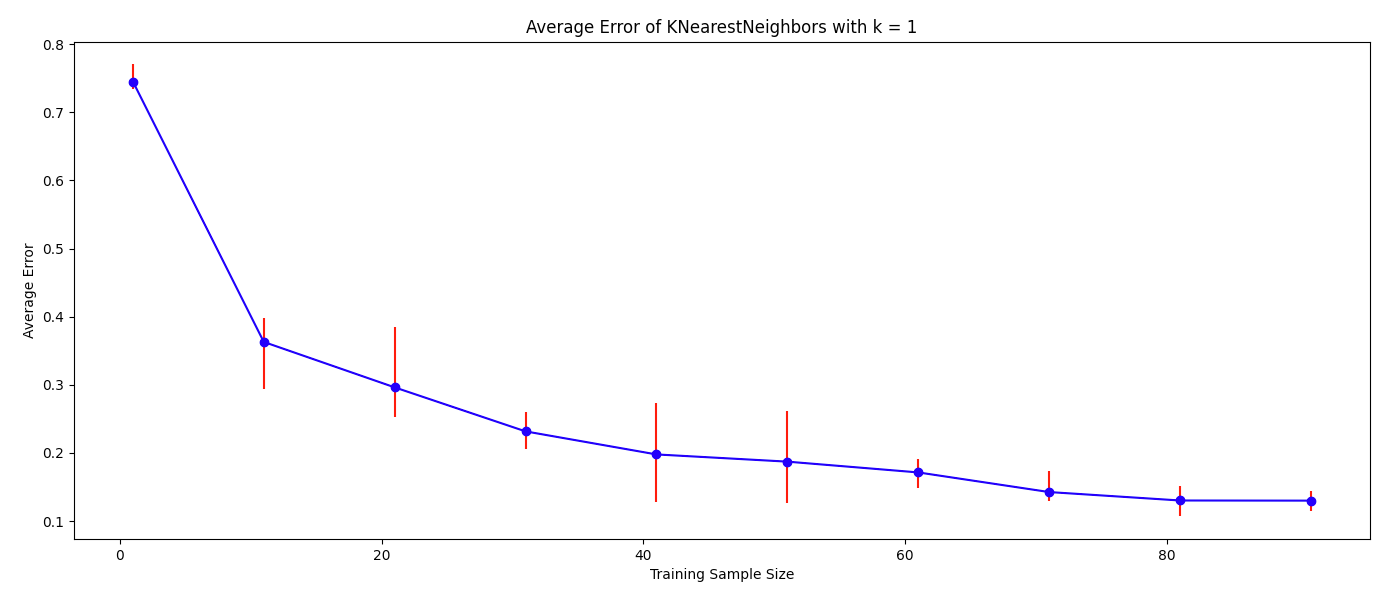
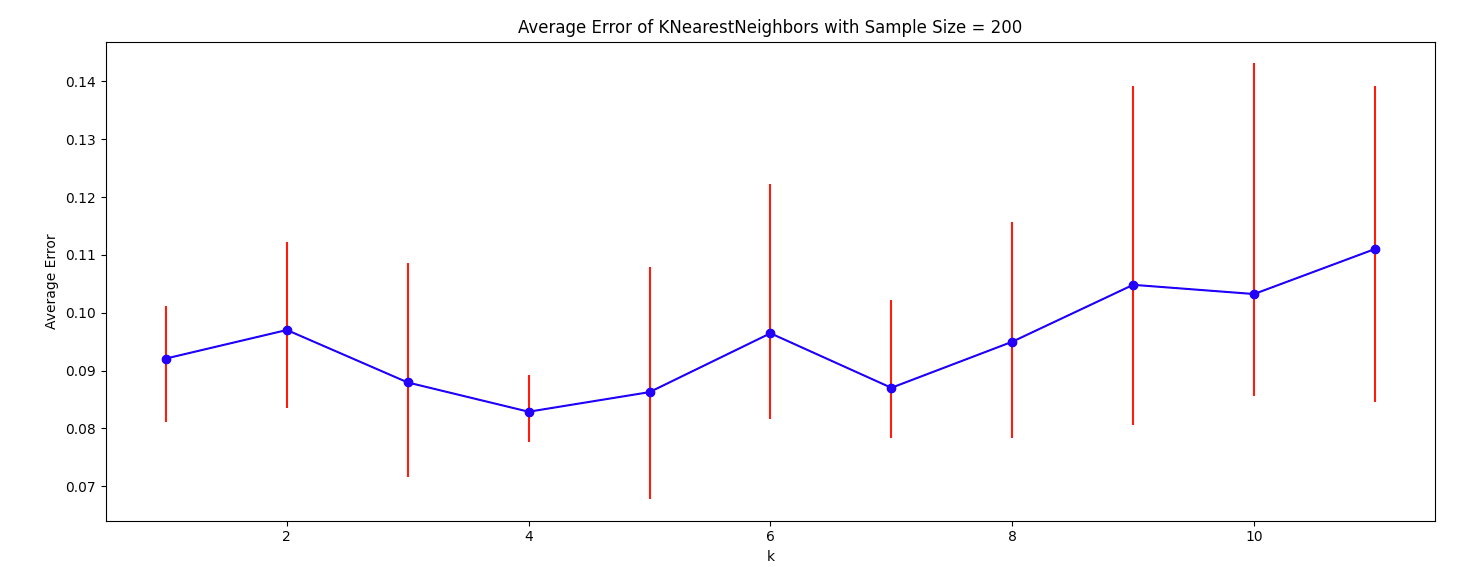
Exercise 1

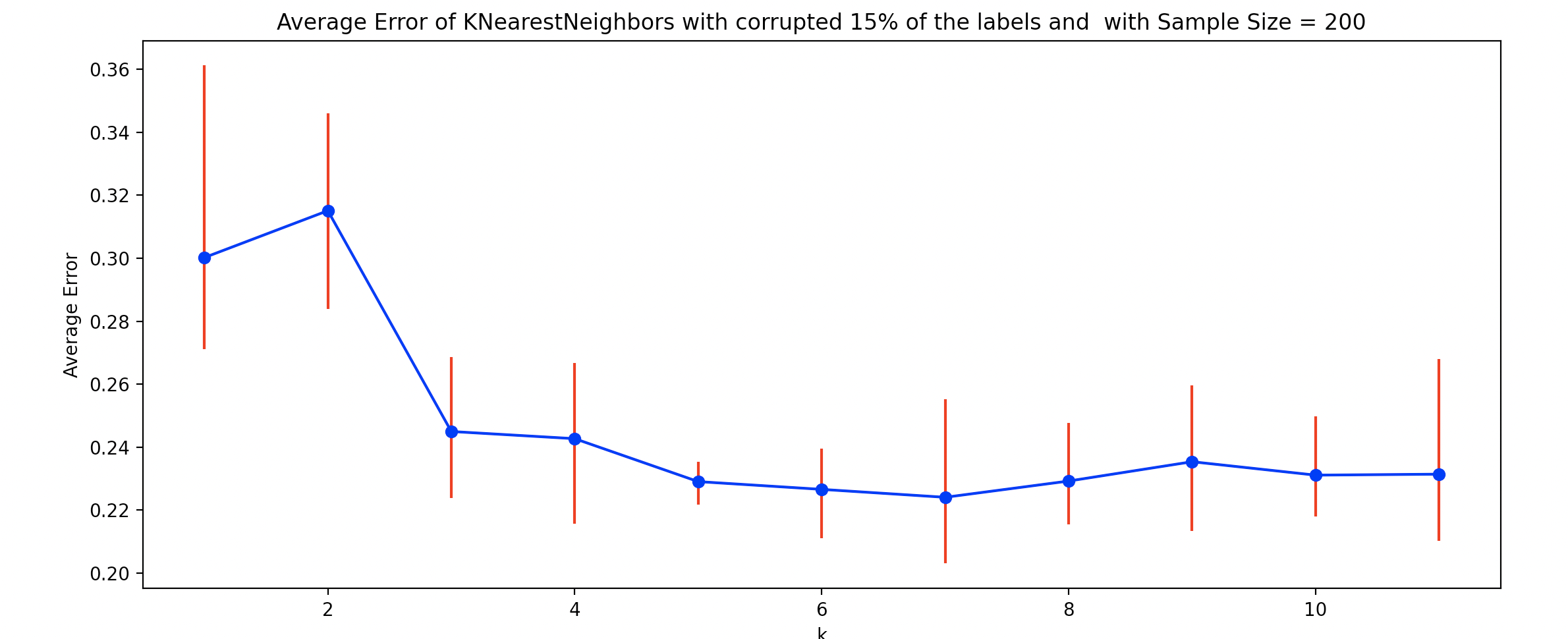
Omri Bar Oz – 313325961, Ilana Pervoi – 318271640

Question 2:



1. 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.
2. 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.
3. 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.





1. 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:

1. Need to prove:  
   for any two pairs if then

Proof:  
 and so WLOG we can decide that and .  
We obtain:

Recall that and of is c-Lipschitz with respect to the Euclidean distance, hence from a property we learned in class:

1. Need to prove:  
   under the given assumptions,

Proof:

So, we will show that .

Let and such that .  
Since is in at least one ball in the set of balls of radius that covers the space of points and there is an assumption that has a point in this ball as well, we obtain

Now, from the assumptions on we know that and has the same label

Question 4:

1. Let hypothesis class of constant functions. therefore where . By definition:  
   .  
   Hence:
2. As we say in class, every example in the will be labeled by 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 will label it as black. More formally:
3. 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 where x is a rabbit from .
4. As we saw in class, where , and is the sample size:  
   in our case: therefore:  
   We are allowed to use this formula for because has a deterministic label condition on the example, whereas does no as it has 2 different labels for the same example.

Question 5:

Let . Consider the hypothesis class of thresholds:  
Let be a distribution over and suppose that the marginal distribution of on is uniform on and that for , for any

1. In this section we will talk about N evenly-spaced threshold for an integer N:  
   We assume that is realizable by . We will use the PAC bounds for finite hypothesis classes we learned in class with . We can use it as is finite and , and is realizable by it and we get:
2. Let . Let .  
   by the definition, an example will get 1 only if it is higher or equal to and 0 only if it is lower than . Therefore:
3. Let with : . Let be the classifier returned from some ERM algorithm trained on S with . We will show that for by showing that and that for every .  
   Let Then . Notice that and equality happens only if , by the definition of .   
   So, for every and for every . Because and it holds that and it implies that is labeling correctly all the points in S which implies that .  
   Let . Wlog assume that .  
   It is easy to see that . While true label is 1, will label as 0, which implies that .  
   This concludes that an ERM algorithm will return a for .  
   Let be the hypothesis returned by an ERM algorithm. then by sub-question (b),
4. Let . Then:  
   this holds for as well because   
   Let be the hypothesis returned by a ERM algorithm with .
5. We will use . by the last sub-question:  
   we want it to be greater than then:
6. 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 and that we know for which . Now, we assume that there is a such that it suffices the assumptions for sub-question (a), but we cannot find this 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 that suffices the assumptions.  
   Moreover, even if we knew the 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 . Let’s check what the value of N should be, so m would less than 121.109:  
   Which means that we can’t use any hypothesis class to match the sample size of sub-question (e), then it is better to use the approach in sub-question (e).