MALIS
Group Exercise

October 4 2022

Group Name:	
Group Members:	

Introduction and Setup

1. Give an example of a real world problem that can be solved using machine learning in each of the setups listed below. Be specific about the input data and the expected output. Do not use any example that may have been given during the lecture.

   a. Regression problem

   b. Classification problem

   c. Clustering

2. Suppose you have a file of data where the examples are classified into two possible classes, 0 and 1. In the file, the first half of examples belong to class 0 and the last half of examples belong to class 1. Before applying your learning algorithm, you split the data so that the first 70% of examples from the file correspond to your training data and the last 30% of examples from the file correspond to your test data. Why might this be problematic?

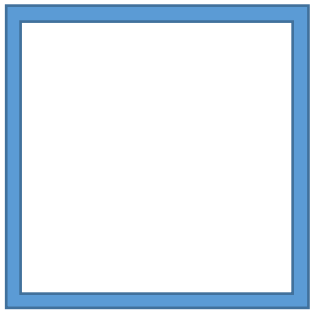
k-Nearest Neighbors

3. What is the training accuracy (i.e., accuracy on the training data) of a k nearest neighbor classifier when k=1?

4. Suppose when using a k nearest neighbor classifier on N training examples, you set k equal to N. What will be true about the classifier's predictions on the test data?

5. Recall the Minkowski distance. As p $\rightarrow\infty$, what does the distance represent?

6. Suppose a hyper cube in a D-dimensional space, with each edge of length 1, i.e. $[0,1]^D$. What is the volume of the 5% outermost part of the hypercube, when D= 2, 10, 1000? (See the illustration in 2D below) You need to show how you arrive to your answer.



Estimate the volume (area in this case) of the shaded zone for D=2, 10, 1000.

How do you see the observed phenomenon may affect the performance of the kNN algorithm?