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

Open question. The output should be a continuous value. Ideally, some details about the input data would be appreciated.

    b. Classification problem

    Open question. The output should be a discrete value (numerical labels or the categories are OK). Ideally, some details about the input data would be appreciated.

    c. Clustering

Open question. The answer should describe the kind of groups that are expected to be found. Some description about the input features is appreciated.

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?

No validation set. The test set contains only points from class 1. The training set has an imbalanced representation of one class.

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?

100%.

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?

The majority class in the training data.

5. Recall the Minkowski distance. As p →∞, what does the distance represent?

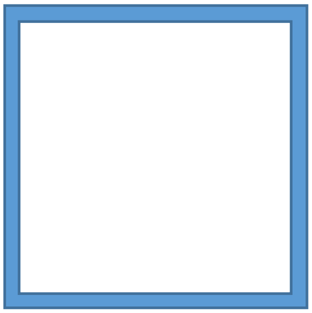
   The max of the differences among all the dimensions.

   Simple illustration. Suppose two 2D points (1000,1) and (1,0.99). The inner term in the Minlowski distance will be:

   $(1000 - 1)^{\infty} + (1 - 0.99)^{\infty}$.

   The second term will be negligible compared to the first one, so after doing 1/p the result will be very close to the largest of the two terms.

6. Suppose a hyper cube in a D-dimensional space, with each edge of length 1, i.e. $[0,1]^D$. Suppose you take the 5% outermost part of each edge. What is its volume 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.

   Volume = $1^D - (0.9)^D$

   Replace D, accordingly

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

   This means that in very large dimensions, points are very spread out (sparse). The concept of a nearest neighbor may not really hold as it is necessary to travel a long distance to reach it.