ECE 364: Assignment #2

1. (10 pts) A data analyst building a k-nearest neighbor model for a continous prediction (i.e., regression) problem is considering appropriate values for k.

   (a) (5 pts) Initially, the analyst uses a simple average of the target feature values for the k nearest neighbors in order to make a prediction. After experimenting with values in the 1-10 range, it occurs to the analyst that setting the value of k to the total number of data instances in the training set may yield very good results. Do you think this is true? Explain.

   (b) (5 pts) If the analyst was using a distance-weighted average rather than a simple average to make predictions, would it make the analyst's idea any more useful? Explain.

2. (10 pts)

   (a) (5 pts) Consider two descriptive features: x_1 and x_2, and two data instances, a and b, in this feature space, as shown in the figure below. What is the cosine similarity between a and b?

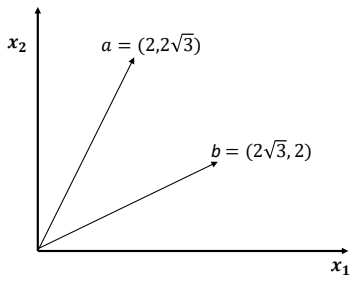

Figure 1: Feature space

   (b) (5 pts) What is the Minkowski distance, with parameter $p = 4$, between (3,5,7,9) and (7,9,3,5)?

3. (10 pts) Prove that the Russel-Rao similarity index is not a metric.

4. (10 pts) The following table shows the points a set of students received out of 10 on two modules, M1 and M2, and the final letter grade. The professor notices that there is a larger variance across students in the points for module M2 than there is for module M1. Therefore, she decides to use a 1-NN model based on Mahalanobis distance to determine the letter grade of a student who received 7 points on module M1 and 7 points on module M2. What is this letter grade if the covariance matrix of M1 and M2 is given by:

$$\begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix}$$

Student ID	M1	M2	Grade
1	6	9	A
2	5	9	B
3	6	7	C

5. (20 pts) **Coding project**

   In this project, you will train classifiers based on the k-nearest neighbors algorithm to predict whether or not a female patient has diabetes. The dataset consists of 767 patients with the following eight numerical descriptive features each:

   - Number of pregnancies
   - Glucose
   - Diastolic blood Pressure
   - Skin thickness
   - Insulin
   - BMI
   - Pedigree
   - Age

   The target label is a binary variable indicating whether or not a patient has diabetes.

   The data will be divided and preprocessed into a training set and a validation set. You will start with the default kNN classifier in the Scikit-learn library and evaluate the effect of varying the number of neighbors. You will also train kNN classifiers that use distance weighting and different power parameters for the Minkowski metric.

   Please see the Jupyter notebook for more details.

   GitHub repository for ECE364 coding projects: https://github.com/JHA-Lab/ece364_2024