ELE 364: Assignment #2

1. (10 pts) Suppose **a** and **b** are vectors of the same size. The triangular inequality states that $\|\mathbf{a}+\mathbf{b}\| \leq \|\mathbf{a}\| + \|\mathbf{b}\|$. Show that the following is also true: $\|\mathbf{a}+\mathbf{b}\| \geq \|\mathbf{a}\| - \|\mathbf{b}\|$.

2. (10 pts) A video streaming company uses its users' historical streaming records to predict whether the user will watch its newly released music video. Below are five historical records. Each row lists the number of times a user watches the relevant videos in other categories, and whether that user finally watches the target music video:

User ID	Romance	Thriller	Comedy	Music	News	Politics	TV Series	Target feature
1	1	1	0	0	0	1	1	False
2	1	0	1	3	3	1	0	True
3	0	0	1	1	1	0	0	True
4	0	2	1	1	1	0	1	False
5	0	0	0	0	0	0	3	False

A new user now watches videos from the Romance category twice, Thriller category twice, Comedy category once, and News category twice. He/she does not watch videos from the other categories.

Use a k-NN ($k = 3$) model to predict whether this user will watch the music video. Predict first using Euclidean distance, and then Manhattan distance. Are the predictions the same?

3. (10 pts) An online retail company develops a recommender system. The system recommends items based on the previous purchases of the users. The table below shows three users and their previous purchases (1 means the user bought the item).

User ID	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8
1	1	1	0	1	0	0	1	1
2	1	0	0	0	1	1	0	0
3	0	0	1	1	0	0	1	0

   (a) Calculate the similarity of users 1 and 2 using Russel-Rao, Sokal-Michener, and Jaccard similarity indexes. Which similarity index do you think the system should adopt?

   (b) Which items does the system recommend to user 3 based on the similarity index you chose? Note that the system would find the user who is more similar to user 3 and recommend to user 3 the items he/she has not purchased yet.

   (c) What if the data were obtained from a medical institution where 1 means that a patient has a particular symptom? Assume that patient 1 (user 1) is diagnosed with a disease, whereas patient 2 is not diagnosed. What similarity index would you choose to decide whether patient 3 has that disease? What would the algorithm predict for patient 3?

4. (10 pts) The following table describes a set of individuals in terms of their Weight in kilograms and Height in meters, and whether or not they have Diabetes.

ID	Weight	Height	Diabetes
1	73	1.6	true
2	55	1.7	false
3	64	1.5	true
4	90	1.9	false
5	70	1.8	false

Clinicians often use BMI as a combined measure of an individual's weight and height. BMI is defined as an individual's weight in kilograms divided by height in meters-squared. Assuming that the profiles of the five individuals in the system were updated so that the features Weight and Height were replaced by a single feature BMI and also that the doctor entered the patient's BMI into the system, what prediction would the system make for a patient with Weight = 65 kilograms and Height = 1.7 meters, based on a 1-NN classifier?

5. (20 pts) 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 refer to the Jupyter notebook for more details.