ELE 364: Assignment #2

1. (10 pts) Email spam filltering models often use a bag-of-words representation for emails. In this representation, the descriptive features that describe a document (in our case, an email) represent how many times a particular word occurs in the document. One descriptive feature is included for each word in a predefined dictionary. The dictionary is typically defined as the complete set of words that occur in the training dataset. The table below lists the bag-of-words representation for the following four emails and a target feature, SPAM, indicating whether they are spam emails or genuine emails: (i) free machine learning book, (ii) free, free gambling fun, (iii) fun, fun, fun machine learning, (iv) gambling fun with machine learning.

ID	Free	Machine	Learning	Book	Gambling	Fun	With	Spam
1	1	1	1	1	0	0	0	false
2	2	0	0	0	1	1	0	true
3	0	1	1	0	0	3	0	false
4	0	1	1	0	1	1	1	true

   (a) What target level would a weighted 4-NN model (using weight = reciprocal of the squared Euclidean distance between the neighbor and the query) return for the following query: free fun with machine learning?

   (b) There are a lot of zero entries in a spam bag-of-words dataset. This is indicative of sparse data and is typical for text analytics. Cosine similarity is often a good choice when dealing with sparse non-binary data. What target level would a 3-NN model using cosine similarity return for the above query?

2. (10 pts) You have been given the job of building a recommender system for a large online shop that has a stock of over 100,000 items. In this domain, the behavior of customers is captured in terms of what items they have bought or not bought. For example, the following table lists the behavior of two customers in this domain for a subset of the items that at least one of the customers has bought.

ID	Item 9	Item 91	Item 341	Item 855	Item 1729
1	true	true	true	false	false
2	true	false	false	true	true

   (a) The company has decided to use a similarity-based model to implement the recommender system. Which of the following three similarity indexes do you think the system should be based on: Russell-Rao, Sokal-Michener, Jaccard?

   (b) Which item will the system recommend to the following customer? Assume that the recommender system uses the similarity index you chose and is trained on the sample dataset given above.

	Item	Item	Item	Item	Item
ID	9	91	341	855	1729
Query	true	false	true	false	false

(c) (2 extra pts) If data were provided for a sixth item, what would be its item number assuming it follows the same rule as the one followed by the first five item numbers?

3. (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	68	1.7	true
2	55	1.6	false
3	65	1.6	true
4	100	1.9	true
5	65	1.5	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?

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. So she decides to use 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	4	3	F
3	5	9	B
4	6	7	C

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

For this project, you will use the nearest neighbor algorithm to train models to classify the type of a glass.

The dataset consists of numerical descriptive features, measuring the refractive index as well as the weight concentration of different elements, e.g., sodium, magnesium.

The target feature is a categorical variable indicating whether the glass is a float building window, non-float building window, float vehicle window, non-float vehicle window, container, tableware, or headlamp.

You will use the nearest neighbors algorithm to train classifiers and investigate the effect of different hyperparameters, including the number of neighbors and distance function choice.

See the Jupyter notebook for more details.