ELE 364: Assignment #2 Solutions

1. (10 pts)

   (a) The Euclidean distances between the query and the four data instances are 1.7321, 2.2361, 2.4495, and 1.4142, respectively. Thus, the reciprocals of the squared Euclidean distances are 0.3333, 0.2, 0.1667, and 0.5, respectively. Thus, the total weight for Spam = true target level is 0.7, and for Spam = false target level is 0.5. Thus, the prediction is Spam = true.

   (b) Cosine similarity needs the lengths of the vectors depicting the data instance and query as well as their dot product. The lengths of the four data instances and the query are 2.0, 2.4495, 3.3166, 2.2361, 2.2361, respectively.

   The cosine similarity is calculated with the data in the table below.

Pair	$q \times d[i]$	Dot prod.	Cosine similarity
$(q, d[1])$	1 1 1 0 0 0 0	3	0.6708
$(q, d[2])$	2 0 0 0 0 1 0	3	0.5477
$(q, d[3])$	0 1 1 0 0 3 0	5	0.6742
$(q, d[4])$	0 1 1 0 0 1 1	4	0.8

   Hence, the three closest data instances to the query are $d[4]$, $d[3]$, and $d[1]$. Thus, prediction is majority(true,false,false) = false.

2. (10 pts)

   (a) In a domain where there are hundreds of thousands of items, co-absences are not that meaningful. For example, you may be in a domain where there are so many items that most people have not seen, listened to, bought, or visited that the majority of features will be co-absences. The technical term to describe a dataset where most of the features have zero values is sparse data. In these situations, you should use a similarity index that ignores co-absences. For a scenario such as this one, where the features are binary, the Jaccard similarity index is ideal as it ignores co-absences.

   (b) Jaccard$(q, d[1]) = \frac{2}{2+1} = 0.6667$ and Jaccard$(q, d[2]) = \frac{1}{4} = 0.25$. Thus, the query customer is more similar to customer $d[1]$ than to customer $d[2]$. There is only one item that customer $d[1]$ has bought that the query customer has not bought: item 91. Hence, the system will recommend item 91 to the query customer.

   (c) (2 extra points) The item numbers follow the $i^3 + (i+1)^3$ rule, where i is equal to 1, 3, 5, 7, and 9, respectively. Hence, the next item number would be $11^3 + 12^3 = 3059$.

3. (10 pts) The modified table is given below.

   The BMI of the query patient is 22.49. The nearest neighbor is individual 2. Hence, the prediction is false.

ID	BMI	Diabetes
1	23.53	true
2	21.48	false
3	25.39	true
4	27.70	true
5	28.89	false

4. (10 pts) Mahalanobis distance requires the inverse of the covariance matrix. Since a matrix multiplied by its inverse leads to the identity matrix, we have

$$\begin{bmatrix} 1 & 2 \\ 2 & 8 \end{bmatrix} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This leads to four equations: $a + 2c = 1$, $b + 2d = 0$, $2a + 8c = 0$, and $2b + 8d = 1$. These four equations can be solved to obtain $a = 2$, $b = -0.5$, $c = -0.5$, $d = 0.25$. Since the query is (7,7), the square of its Mahalanobis distance from the four data instances in the dataset can be obtained by

$$\begin{bmatrix} 1 & -2 \end{bmatrix} \begin{bmatrix} 2 & -0.5 \\ -0.5 & .25 \end{bmatrix} \begin{bmatrix} 1 \\ -2 \end{bmatrix} = 5$$

$$\begin{bmatrix} 3 & 4 \end{bmatrix} \begin{bmatrix} 2 & -0.5 \\ -0.5 & .25 \end{bmatrix} \begin{bmatrix} 3 \\ 4 \end{bmatrix} = 10$$

$$\begin{bmatrix} 2 & -2 \end{bmatrix} \begin{bmatrix} 2 & -0.5 \\ -0.5 & .25 \end{bmatrix} \begin{bmatrix} 2 \\ -2 \end{bmatrix} = 13$$

$$\begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} 2 & -0.5 \\ -0.5 & .25 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2$$

Hence, the answer is C.