Norman Chung
5550926
CS 165A MIDTERM

1a. $P(C=1) = P(A) + P(B) = 0.4 + 0.5 = 0.9$

1b. $P(D=1) = P(A)P(B) + P(\sim A)P(\sim B) = 0.4*0.5 + 0.6*0.5 = 0.5$

1c. $P(D|A)$, since we are given A, it depends on whether B is 0 or 1, so = 0.5

1d. A and D are dependent.

2a. Yes, because given Y, the value of W does not affect the values of X or Z.

2b. No, because both W and V affect Y, or have a direct relation to Y.

3a.

3b. $P(Z|W) = P(W,Z)/P(W) = P(V,W,X,Y,Z)/P(W) = P(Z|V,X,Y)P(X)P(V)P(Y|V,W)P(W)$

4. For a self-driving car, the performance measure would be the time it takes to travel to a location, and how safely it got there(whether it hit anything along the way). The environment would be the roads that the car travels on, and the weather it is traveling in. The actuators include the steering wheel, the brakes, the accelerator, the windshield wipers, the turning lights, etc. The sensors would be the distance measuring that determines how far the car is from a wall or another car, the speedometer, the gps to determine where its location is, etc.

5. The strategy for creating branches is based off of information gain. Whichever possible branch would give us the most information for the next branch is used first, and this process continues until the stop condition occurs, which is when there is no more data to go through, in which the decision tree is finished.

6a. H, L, N, R, G

6b. Class 1: GHIL Class 2: NQR, so M is class 1.

6c. No. This is because Euclidean distance is calculated using all features to determine the distance from point M. Since the z feature has the possibility to be large, there is the possibility that the z feature would make the x and y feature irrelevant in some calculations. However, I believe that if the range for Z was similar to that of x and y, it would be an appropriate distance metric.

Norman Chung
5550926
CS 165A MIDTERM

7. For using k-means, k-means is known to have a weakpoint when dealing with outliers. This is because if it is considered one cluster, the calculated point would be skewed towards the outlier, and wouldn't be representative of the actual cluster without the outlier. A solution to this would be to use k-medoids instead, as for k-medoids, the point selected is a data point that already exists.

8a. The vector space model is used to determine how similar a document is to the query. There are n-dimensions, where n is the amount of terms used to determine similarity. If the query vector is heavily towards one axis, and away from all the others, any document that has their vector heavily towards the same axis means that it is very similar to the query. The farther away a document's vector is from the query vector, the less similar it is.

8b. Cosine distance is used over dot product similarity, because cosine distance takes into account how many times a specific word was said, along with how long the document is, and calculates the distance with these things in mind. For dot product similarity, it doesn't take into account when the same word is said multiple times in a document, and calculates it based off of each word.

8c. For TF-IDF, it takes into account words that appear multiple times in the document, along with words that are rare in the document. With this, it gives these words a higher weight, and the reasoning is because a word is considered more important if it shows up multiple time, and for IDF, a word is considered more important when it shows up rarely.

9a. $P(A) = 4/12 = 1/3$
$P(B) = 4/12 = 1/3$
$P(C) = 4/12 = 1/3$

9b. $P(TRUE|A) = (1/2)/1/3 = 3/2$
$P(FALSE|A) = (1/2)/1/3 = 3/2$
$P(TRUE|B) = (1/2)/1/3 = 3/2$
$P(FALSE\ |B) = (1/2)/1/3 = 3/2$
$P(TRUE|C) = (1/2)/1/3 = 3/2$
$P(FALSE\ |C) = (1/2)/1/3 = 3/2$

9c.

9d.