German University in Cairo
Faculty of Media Engineering and Technology
Spring 2019

CSEN1083 – Data Mining

Problem Set #1

Problem 1

Discuss whether or not each of the following activities is a data mining task:

(a) Dividing the customers of a company according to their gender.

(b) Dividing the customers of a company according to their profitability.

(c) Computing the total sales of a company.

(d) Sorting a student database based on student identification numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

(f) Predicting the future stock price of a company using historical records.

(g) Monitoring the heart rate of a patient for abnormalities.

(h) Monitoring seismic waves for earthquake activities.

(i) Extracting the frequencies of a sound wave.

Solution:

(a) No. This is a simple database query.

(b) This is an accounting calculation, followed by the application of a threshold. However, predicting the profitability of a new customer would be data mining.

(c) No. Again, this is simple accounting.

(d) No. Again, this is a simple database query.

(e) No. Since the die is fair, this is a probability calculation. If the die were not fair, and we needed to estimate the probabilities of each outcome from the data, then this is more like the problems considered by data mining. However, in this specific case, solutions to this problem were developed by mathematicians a long time ago, and thus, we wouldn't consider it to be data mining.

(f) Yes. We would attempt to create a model that can predict the continuous value of the stock price. This is an example of the area of data mining known as predictive modelling. We could use regression for this modelling, although researchers in many fields have developed a wide variety of techniques for predicting time series.

(g) Yes. We would build a model of the normal behavior of heart rate and raise an alarm when an unusual heart behavior occurred. This would involve the area of data mining known as anomaly detection. This

could also be considered as a classification problem if we had examples of both normal and abnormal heart behavior.

(h) Yes. In this case, we would build a model of different types of seismic wave behavior associated with earthquake activities and raise an alarm when one of these different types of seismic activity was observed. This is an example of the area of data mining known as classification.

(i) No. This is signal processing.

Problem 2

You have two bags. The first bag contains 12 white marbles and 7 black marbles. The second contains 9 white marbles and 4 black marbles. One bag is chosen uniformly at random, and then a marble is selected from the bag. The marble selected is white. What is the probability it came from the first bag?

Solution:

Let B denote the bag where $B = \{1, 2\}$ and let C denote the marble color $C = \{w, b\}$.

The conditional probabilities are:

$p(C = w \mid B = 1) = 12/19$

$p(C = b \mid B = 1) = 7/19$

$p(C = w \mid B = 2) = 9/13$

$p(C = b \mid B = 2) = 4/13$

Given that the bags are chosen uniformly, then $p(B = 1) = 0.5$ and $p(B = 2) = 0.5$.

We are supposed to compute $p(B = 1 \mid C = w)$. Using Bayes' rule

$$p(B=1|C=w) = \frac{p(C=w|B=1)p(B=1)}{p(C=w)}$$

$$\begin{aligned} p(C=w) &= p(C=w, B=1) + p(C=w, B=2) \\ &= p(C=w|B=1)p(B=1) + p(C=w|B=2)p(B=2) \\ &= \frac{12}{19} \times 0.5 + \frac{9}{13} \times 0.5 = 0.66 \end{aligned}$$

$$\therefore p(B=1|C=w) = \frac{(12/19)0.5}{0.66} = 0.48$$

Problem 3

Consider a chess tournament in which your robot is playing against two groups of robots. Group 1 consists of 30 robots while Group 2 consists of 50 robots. The probability that your robot wins against any robot from Group 1 is 0.3 and against any robot from Group 2 is 0.4.

(a) Choose the correct answer:

i) The probability that your robot plays against a robot from Group 1 is the

1) Prior probability 2) Posterior probability 3) Likelihood

ii) Knowing that your robot won the game, the probability that this game was against a robot from Group 1 is the

1) Prior probability 2) Posterior probability 3) Likelihood

(b) Compute the probability of winning a game.

Solution:

(a) i → 1, ii → 2

(b) $p(\text{Win}) = p(\text{Win, Group 1}) + p(\text{Win, Group 2})$

$= p(\text{Win}|\text{Group 1})\, p(\text{Group 1}) + p(\text{Win}|\text{Group 2})\, p(\text{Group 2})$

$= 0.3 \times 3/8 + 0.4 \times 5/8 = 0.3625$