Matthew Barnes
0555121

COIS 4400 Data Mining | Assignment 3

Question 1

Given the data shown in Figure 8.2 (pg. 496), explain how each of the following learning techniques would perform on the data sets: decision tree, k-nearest neighbours, genetic algorithms, multilayer feedforward neural networks.

a) **Decision Tree →**

Because the data is well defined, it should be very simple for a decision tree.

kNN →

The two clusters are well-separated, so the kNN classifier would perform excellently in this instance. The definition for each point is clear.

Genetic Algorithms →

Due to the nature of the data, being in a cluster, without outliers, the genetic algorithms may have to run through many generations before they are accurately classified.

Multilayer Feedforward Neural Networks →

Because the data is well-defined, an artificial neural network should be able to classify the data quite easily.

b) **Decision Tree →**

This is also well defined, so the decision tree will not be too complex.

kNN →

Although the data is now center-based, there should not be too much of a difference between this data set and the previous ones. The only cause for concern would be if the data was in directly in the center of the two clusters, but that would marginally raise the miss rate of the classification algorithm.

Genetic Algorithms →

The hit rate for these populations should be higher, due to the smaller range of data, but other than that, the same properties from the previous data apply.

Multilayer Feedforward Neural Networks →

Same as the last set of data, because the it is well-defined, the ANN should have no problem classifying data.

c) **Decision Tree →**

A decision tree will be more complex due to the irregularly shaped clusters of data.

kNN →

Because each point is closer to at least one point in its cluster than to any point in another cluster, the kNN shouldn't have too much of a problem classifying the data.

Genetic Algorithms →

Due to the odd shape of the data, a genetic algorithm would not perform extremely well consistently. If a generation was to populate some of the smaller clusters, those traits would be passed down, and it would take much longer to accurately classify the data.

Multilayer Feedforward Neural Networks →

The artificial neural network may struggle a bit in this case, as the calculations are not as simple and data is not as well spread apart as the last instances. The weightings of connections between nodes may be peculiar, and because of this, the ANN may not accurately classify data.

d) **Decision Tree →**

The decision tree will be quite simple. This is because of the low density region. It would be quite simple to map out the rules for this instance.

kNN →

kNN will have a little bit more trouble due to the low density data. If a point lands near a high density cluster, in the low density region, it may be misclassified due to the low density near to the high density cluster.

Genetic Algorithms →

The genetic algorithm should perform well enough in this case. There will have to be a high density of points in and around those high density clusters.

Multilayer Feedforward Neural Networks →

Weightings would be very clear cut in the ANN due to the low density region. The algorithm could generalize a large portion of records to be classified in the low density region. ANNs are also good at separating important data from noise.

e) **Decision Tree →**

If the data is well-defined, or there is a catch-all, (such as the low-density region in the previous instance) then the decision tree will perform well and be relatively simple. With a dataset that doesn’t display these characteristics, the tree can become lengthy and complex. In the case of overlapping data, a decision tree would become quite complex as well.

kNN →

Depending on the shape of the conceptual cluster the kNN could do well. Clusters that are very close to each other, or overlapping are a lot tougher for kNN.

Genetic Algorithms →

If the data is well separated, then a genetic algorithm shouldn’t struggle to classify the data. If the data is not well-defined, then it will struggle, and there will need to be many generations created.

Multilayer Feedforward Neural Networks →

Data that is well organized will be very simple for an ANN to sort. Those that are overlapping will create much more complex relations between nodes.

Question 2

Consider the market basket transactions shown in Table 6.23 of your textbook.

a) What is the maximum number of association rules that can be extracted from this data (including rules that have zero support)?

Given *d* unique items in the itemset, the total number of possible association rules is

$$3^d - 2^{d+1} + 1 \text{ or } 3^6 - 2^7 + 1 = 602$$

b) What is the maximum size of frequent itemsets that can be extracted?

$$2^d \text{ or } 2^6 = 64$$

64 itemsets can be extracted with the largest being of size 4: {Milk, Diapers, Bread, Butter}

c) Write an expression for the maximum number of size-3 itemsets that can be derived from this data set.

$$6\ choose\ 3 = \binom{6}{3} = 20$$

20 size-3 itemsets can be derived from the current data set.

d) Find the itemset of size 2 or larger that has the largest support

Beer, Bread, Butter, Cookies, Diapers, Milk

Itemset	**:**	**# of Times bought together**
{Beer, Cookies}	:	2
{Beer, Diapers}	:	3
{Bread, Butter}	:	5
{Bread, Cookies}	:	1
{Bread, Diapers}	:	3
{Bread, Milk}	:	3
{Butter, Cookies}	:	1
{Butter, Diapers}	:	3
{Butter, Milk}	:	3
{Cookies, Diapers}	:	2
{Cookies, Milk}	:	1
{Diapers, Milk}	:	4
{Beer, Cookies, Diapers}	:	1
{Bread, Butter, Milk}	:	3
{Bread, Butter, Cookies}	:	1
{Bread, Butter, Diapers}	:	3
{Butter, Diapers, Milk}	:	2

As Itemsets get more values, support will decrease, therefore, the itemset with the highest support is **{Bread, Butter}**.

Question 3

For each of the following questions, provide an example of an association rule from the market basket domain that satisfies the following conditions. Also, describe whether such rules are interesting.
a) a rule that has high support and high confidence
b) a rule that has reasonably high support but low confidence
c) a rule that has low support and low confidence
d) a rule that has low support but high confidence

Support (s) → Fraction of transactions that contain both X and Y
Confidence (c) → Measures how often items Y appear in transactions that contain X

a) Bread → Butter

$$c = \frac{\sigma(bread, butter)}{\sigma(bread)} = \frac{5}{5} = 1.00$$

b) Bread → Milk

$$s = \frac{\sigma(beer, cookies)}{|T|} = \frac{2}{10} = 0.2$$

$$c = \frac{\sigma(beer, cookies)}{\sigma(beer)} = \frac{2}{4} = 0.5$$

c) Butter → Cookies

$$s = \frac{\sigma(butter, cookies)}{|T|} = \frac{1}{10} = 0.1$$

$$c = \frac{\sigma(butter, cookies)}{\sigma(butter)} = \frac{1}{5} = 0.2$$

d) Beer → Diapers

$$s = \frac{\sigma(beer, diapers)}{|T|} = \frac{3}{10} = 0.3$$

$$c = \frac{\sigma(beer, diapers)}{\sigma(beer)} = \frac{2}{3} = 0.75$$

Question 4 (20 points)
Given the Bayesian network shown in Figure 5.48 (pg. 321) of your textbook, compute the following probabilities:
a) P(B=good, F= empty, G= empty, S= yes)
b) P(B=bad, F= empty, G= not empty, S= no)
c) Given that the battery is bad, compute the probability that the car will start.

a) $P(B = good(0.9),\ F = empty(0.2),\ G = empty(0.8),\ S = yes(0.2)) = 0.9 * 0.2 * 0.2 * 0.8 = 0.0288 = 2.88\%$

b) $P(B = bad(0.1),\ F = empty(0.2),\ G = not\ empty(0.1),\ S = no(1)) = 0.1 * 0.2 * 0.1 * 1 = 0.002 = 0.2\%$

c) $P(S = yes \mid B = bad(0.1),\ F = empty(0.2)) * P(B = bad(0.1)) * P(F = empty(0.2)) +$
$P(S = yes \mid B = bad(0.1),\ F = notempty(0.8)) * P(B = bad(0.1)) * P(F = notempty(0.8))$
$= (0 * 0.1 * 0.2 * 0.1 * 0.2) + (0.1 * 0.1 * 0.2 * 0.1 * 0.2) = 0 + 0.00004 = 0.00004 = 0.004\%$

Question 5 (20 points)
Describe two other supervised learning techniques we did not discuss in class.

Kriging / Gaussian Process Regression
Kriging is optimal interpolation, based on the previously measured values surrounding the unmeasured point. These values are weighted based on their difference from the unmeasured point, and then these weighted values are used to calculate and classify the unmeasured point.
Interpolation Algorithm characteristics
If data is clustered and sparse, there will not be any good results.
If data is dense and distributed evenly, there will be better results, with fairly good estimates.
Kriging specifically helps compensate for data clustering, giving points within clusters less weighting, or treating it as a single point.

PROAFTN
The acronym PROAFTN stands for “PROcédure d'Affectation Floue pour la problématique du Tri Nominal”, which means in English “Fuzzy Assignment Procedure for Nominal Sorting.” PROAFTN compares records to be classified through many criterion, rather than using just the distance between points. This means qualitative and quantitative characteristics can be used to evaluate a record. It also helps compare data that aren’t evaluated in the same way.

[1]N. Belacel, "Multicriteria assignment method PROAFTN: Methodology and medical application", European Journal of Operational Research, vol. 125, no. 1, pp. 175-183, 2000.