Q) Consider the training examples shown in the Table below for a binary classification problem.

Maintenance	Persons	Luggage Boot	Safety	Buy
High	More	Medium	High	No
High	More	Medium	Medium	No
Medium	More	Medium	High	Yes
Low	5	Medium	High	Yes
Low	4	Big	High	Yes
Low	4	Big	Medium	No
Medium	4	Big	Medium	Yes
High	5	Medium	High	No
High	4	Big	High	Yes
Low	5	Big	High	Yes
High	5	Big	Medium	Yes
Medium	5	Medium	Medium	Yes
Medium	More	Big	High	Yes
Low	5	Medium	Medium	No

a Using the ID3 Algorithm, construct the full decision tree and motivate your answer by showing the steps.

b Evaluate the accuracy of your decision tree using the following testing examples.

Maintenance	Persons	Luggage Boot	safety	Buy
High	More	Big	High	Yes
High	More	Big	Medium	No
Medium	More	Big	High	Yes
Low	5	Big	High	No
Low	4	Medium	High	Yes
Low	4	Medium	Medium	No
Low	4	Medium	Medium	No
High	5	Big	High	Yes
High	4	Medium	High	Yes

Q) It is desired to partition customers into similar groups on the basis of their demographic profile. Which data mining problem is best suited to this task?

Clustering

Q) Suppose in previous exercise, the merchant already knows for some of the customers whether or not they have bought widgets. Which data mining problem would be suited to the task of identifying groups among the remaining customers, who might buy widgets in the future?

Classification

Q) Suppose in previous exercise, the merchant also has information for other items bought by the customers (beyond widgets). Which data mining problem would be best suited to finding sets of items that are often bought together with widgets?

We are looking for association rules

Q) If a source send n messages, with equal probability, the receiver needs to ask $\log_2 n$ yes/no questions to know the message. Explain the concept of Entropy in light of this sentence

Q) Explain with examples the concept of Entropy given its formula

$$E(S) = -\sum_{j=1}^{k} p_j \log_2(p_j)$$

Q) Describe a situation (or give an example) that motivates the use of GainRatio instead of Gain while inducing decision trees.

Q) Do you agree or disagree, and why:

a) holdout, as a method for classification model validation, is a special case of k-fold cross validation No, in hold out we only have one iteration of training and test set which is not changed later on like in cross validation

b) leave-one-out, as a method for classification model validation, is a special case of k-fold cross validation Yes, if there are n tuples and n folds

c) k-fold cross validation helps improving the classification accuracy of a model

Solution says no, confirm this

Q) Describe how to assess the data skew using mean, median, mode summaries.

Q) Given two strings of lengths n and m, what are the time, space, and backtrace complexities of the recursive Edit distance algorithm ? Sketch (e.g., as pseduo code) a way to improve.

Confirm example given in solution

Q) Assume that Edit(X, Y) represents the cost of transforming the string X to Y . Show that Edit(X, Y) and Edit(Y , X) are the same, as long as the insertion and deletion costs are the same. (Edit(.,.) is the string edit distance function)

Q) Compute the edit distance between:
(a) ababcabc and babcbc and
(b) cbacbacba and acbacbacb.
For the edit distance, assume equal cost of insertion, deletion, or replacement.

Q) One of the following terms does not fit with the others. Choose it, and concisely explain why

1. Euclidean.
2. Binning
3. Manhattan.
4. L_∞-norm.

Q) Consider the following database of student grades.

Student	Grade
Student 1	7
Student 2	6
Student 3	10
Student 4	14
Student 5	10
Student 6	19
Student 7	2
Student 8	21
Student 9	0
Student 10	17
Student 11	1
Student 12	6
Student 13	16
Student 14	5

a Use K-Medians to cluster them according to their grades into three cluster. Start with this set as initial cluster representatives {4,6,10}. Show the steps of your answer until convergence.

For K-medians, Dist(X_i, Y_j) is the Manhattan distance:

$$Dist(\overline{X_i}, \overline{Y_j}) = ||X_i - Y_j||_1.$$

Q) Explain the complexity of DBSCAN (you may look at the pseduo code)

Q) Give an example that motivates the need for probabilistic clustering.

Q) Given the following database of car theft incidents:

Example No	Color	Type	Origin	Stolen ?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	White	Sports	Domestic	No
5	White	Sports	Imported	Yes
6	White	SUV	Imported	No
7	White	SUV	Imported	Yes
8	White	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

a Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Red, type=SUV, Origin= Domestic).

b Use the naive Bayes classifier to predict whether or not the following car will be stolen (color= Yellow, type=SUV, Origin= Domestic). (10 pts)

In solution, it says 1/6, is it 1/6 or 1/8?

Naive Bayes classifier

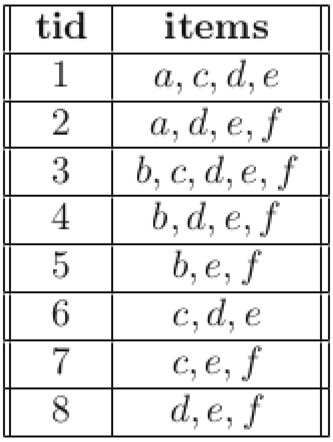
$$P(X \vee C_i) = \prod_{k=1}^{n} P(x_k \vee C_i)$$

Q) Explain the conditions for an ensemble of classifiers to yields more accurate predictions than its individuals.

Q) Which of the following is true in an itemset lattice?

1. Every subset of a frequent itemset is also frequent.
2. Every superset of an infrequent itemset is also infrequent.
3. The lattice size is $2^{|U|}$, where U is the itemset.
4. The support of an itemset is always greater than or equal to the support of any of its supersets.

Q) Consider the transaction database in the table below:



tid	items
1	a, c, d, e
2	a, d, e, f
3	b, c, d, e, f
4	b, d, e, f
5	b, e, f
6	c, d, e
7	c, e, f
8	d, e, f

a) Determine all frequent patterns and maximal patterns at support counts of 3, 4, and 5.

b) Determine the confidence of the rules {a} ⇒ {f }, and {a, e} ⇒ {f } in this transaction database

c) Show the candidate itemsets and the frequent itemsets in each level-wise pass of the Apriori algorithm. Assume a minimum support count of 2.

Q) Consider the 1-dimensional data set with 10 data points {1, 2, 3, . . . 10}.

a) Show three iterations of the k-means algorithms when k = 2, and the random seeds are initialized to {1, 2}.

b) Repeat previous exercise with an initial seed set of {2, 9}. How did the different choice of the seed set affect the quality of the results?

Q) Consider a 1-dimensional data set with three natural clusters. The first cluster contains the consecutive integers {1 . . . 5}. The second cluster contains the consecutive integers {8 . . . 12}. The third cluster contains the data points {24, 28, 32, 36, 40}. Apply a k-means algorithm with initial centers of 1, 11, and 28. Does the algorithm determine the correct clusters?

Q) If the initial centers are changed to 1, 2, and 3, does the algorithm discover the correct clusters? What does this tell you? It will but after multiple iterations. So, initial seed is important

Q) The Apriori algorithm builds C_{k+1} by combining pairs in F_k that have k-1 items in common. Prove that with this strategy $F_{k+1} \subseteq C_{k+1}$

Credits

Some questions are copied from

- Charu C.Aggarwal. Data Mining The Textbook, Springer.
- Jiawei Han, Micheline Kamber and Jian Pe, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2012.