|  |  |  |
| --- | --- | --- |
|  | **Cube** | **Cylinder** |
| **Red** | 10 | 30 |
| **Blue** | 20 | 15 |

1. Following table presents frequencies of objects with regard to shape and color. Is there any association between color and the shape of the objects?
2. Suppose there are n observations , with mean . When a new observation is added to the dataset, the value of the new mean becomes . Show that the change in mean, when the new observation xn+1 is added, is proportional to the difference between and with the proportionality constant .
3. X and Y are continuous variables both with the same mean and no skewness. X has a higher variance then Y.

a) Draw typical distributions of X and Y on the same graph.

b) Draw box plots of X and Y on the same graph.

1. For a student studying 10 terms in a university, her GPA falls in the first 4 terms then remains the same for the next 2 terms and finally increases in the last 4 terms. Assume that the students gets the same number of credits in each term and does not repeat any courses.
   1. Draw a possible GPA versus term graph: evolution of her GPA over time.
   2. Draw a compatible SPA (semester point average) graph as a function of term on the same graph compatible with her GPA.
2. Each of the following plots shows two continuous variables X and Y. For each plot, please specify
   1. whether there exists a functional relationship between the two variables, and
   2. whether principal component analysis can reduce the dimensionality from two to one.

1. Give an example
   1. where outliers are useful and essential patterns to be mined, and
   2. where outliers are useless and caused by error or noise.

16.10.2018

1. Consider a data set of two continuous variables X and Y. X is right skewed and Y is left skewed. Both represent measures about same quantity.
2. Draw typical distributions of X and Y separately.
3. Draw box plots of X and Y separately.
4. Draw q-plots (quantile) of X and Y separately.
5. Draw q-q plot of X and Y.
6. A dice is thrown 60 times. Numbers 1, 2, and 3 are obtained 5 times each; and numbers 4, 5, and 6 are obtained 15 times each. Test the fairness of the dice using Chi-Square test.
7. Consider X and Y as points in a multidimensional space with given coordinates: X (1, 2, 4), Y (2, 2, 1). Calculate Manhattan, Euclidian, and Chebyshev distances between the two points.
8. Construct a dataset with X and Y variables where k-means is not suitable
   1. because the clusters are not spherical,
   2. because clusters are spherical but have different sizes.
   3. because the final clusters are affected by initial cluster centers

Show your work using scatter plots.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **1** | **2** | **3** | **4** | **5** |
| **1** | 0 |  |  |  |  |
| **2** | 1 | 0 |  |  |  |
| **3** | 4 | 5 | 0 |  |  |
| **4** | 3 | 9 | 2 | 0 |  |
| **5** | 6 | 7 | 10 | 8 | 0 |

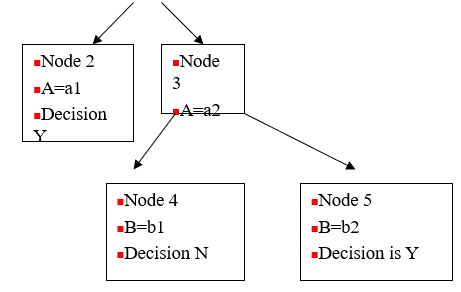
1. The following table provides the dissimilarity (distance) matrix for 5 data points. Draw the dendogram which shows the steps of Agglomerative Hierarchical Clustering using the single-link distance method.
2. Using the given transaction database, find strong association rules using Apriori algorithm with minimum support of 40% and confidence of 60%. Report and interpret values of the lift measure for the obtained rules.

|  |  |
| --- | --- |
| 1 | A, B, C, D, F |
| 2 | A, C |
| 3 | A, B, F |
| 4 | B, D, F |
| 5 | B, C, D, F |
| 6 | B, E, G |
| 7 | B, D, E, F, G |
| 8 | D, E, G |
| 9 | F |
| 10 | F, G |

1. A rule is deemed strong based on its support and confidence values. For each of the following cases, generate some simple datasets and specify your assumed support and confidence values.
   1. A ⇒ B and B ⇒ C are strong rules but A ⇒ C is not.
   2. A ⇒ C and B ⇒ C are strong rules but A AND B ⇒ C is not.
   3. A ⇒ B and A ⇒ C are strong rules but A ⇒ B AND C is not.
   4. A AND B ⇒ C is a strong rule but A ⇒ C and B ⇒ C is not.,

*Bonus:* If A ⇒ B AND C is a strong rule, are A ⇒ B and A ⇒ C necessarily strong rules?

1. Answer and briefly explain or give examples for each of the following.
   1. Suppose {A,B,C} is a frequent 3 itemset. Does it imply that {A,B} and {A,C} are frequent 2 itemsets?
   2. Suppose {A,B}, {A,C}, and {B,C} are frequent 2 itemsets. Does it imply that {A,B,C} is a frequent 3 itemset?
   3. Suppose {A,B} is a frequent 2 itemset. Does it imply that, A ⇒ B and B ⇒ A are strong rules?
2. Given that C4: {(A, B, C, D), (B, D, E, F)}
   1. Write an L3.
   2. Write a C3.
3. In a transactional database, there are 100 transactions. {A,B} is a frequent 2-itemset with a support count of 20. For both part a) and b) separately draw contingency tables and find the asked quantities.
   1. For the rules A⇒B confidence is 0.4 and for the rule B⇒A it is 0.5. Fill the contingency table. What is the number of null transactions with respect to {A,B}?
   2. Critical confidence level is 0.5, A⇒B is a strong rule but B⇒A is not. What is the maximum possible number of null transactions with respect to {A.B}?
4. Consider a data set of two attributes *A* and *B*. *A* is continuous. *B* is categorical and has two values as “Y” and “N”. When *A* is discretized into two equi-width intervals, no information is provided by *B* but when discretized into three equi-width intervals there is perfect information provided by *B*. Construct a simple dataset obeying these characteristics.
5. Construct a data set that generates the tree shown below.



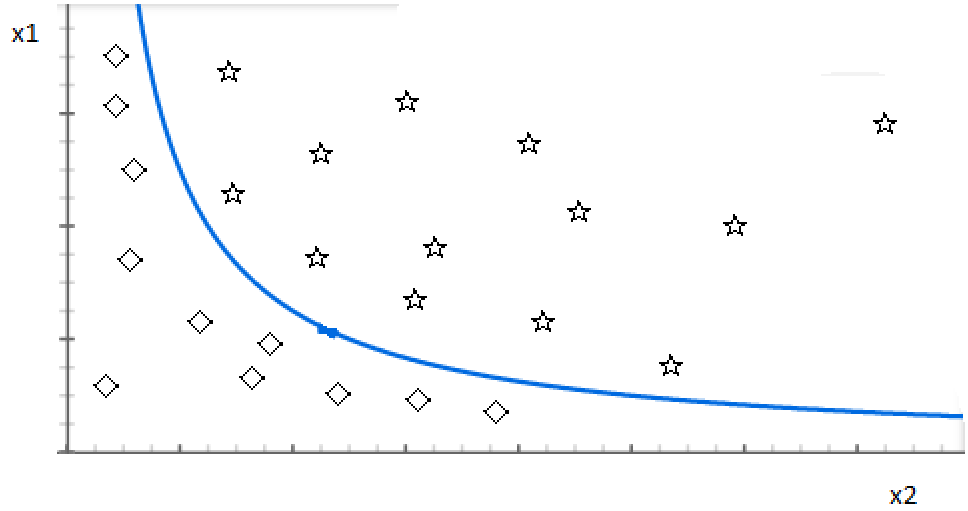
1. Consider a decision tree with only two branches in that the attribute selection measure is entropy. Bearing in mind that each candidate input attribute may have more then two distinct values, how do you modify the ID3 algorithm to handle such a constraint on the number of branches of the tree? What if the input attribute is also ordinal?
2. A classification problem is that: given a student with two input variables, will she pass a course or not? Data about two categorical input variables about students *Repeating* and *Working* are collected. The training data for 100 students are shown in the table below. The first raw of the tables indicates that there are 10 repeating and working students. If the classification problem is to be solved by Naïve Bayesian method, what are the probabilities of passing and not passing for a working student repeating the course?

|  |  |  |  |
| --- | --- | --- | --- |
| Repeating? | Working? | Pass? | number of cases |
| y | y | y | 10 |
| y | n | y | 20 |
| n | y | y | 10 |
| n | n | y | 20 |
| y | y | n | 15 |
| y | n | n | 5 |
| n | y | n | 15 |
| n | n | n | 5 |

|  |  |  |
| --- | --- | --- |
| **A** | **B** | **Class Dist.** |
|  |  | 3 , 2 |
|  |  | 3 , 0 |
|  |  | 1 , 4 |
|  |  | , 2 |

1. Data about two categorical input variables *A* and *B*, and an output variable are collected. Each variable has two possible values: , ; , ; , respectively. The table shows class distribution for different combination of input variable values. Generate the decision tree based on ID3 algorithm.

|  |  |  |
| --- | --- | --- |
| **Department** | **Age** | **Status** |
| Sales | 31-35 | Senior |
| Sales | 26-30 | Junior |
| Sales | 31-35 | Junior |
| Systems | 21-25 | Junior |
| Systems | 31-35 | Senior |
| Systems | 26-30 | Junior |
| Marketing | 36-40 | Senior |
| Marketing | 31-35 | Junior |

1. The table consists of training data from an employee database. Predicted variable is status. Age and Department are inputs. Design a multilayer feedforward neural network for the given data. Label the nodes in the input, hidden and output layers. Describe how you encode the input and output variables. Specify the parameters of the network that can be changed by the backpropagation algorithm.
2. For a typical machine learning task, show the behavior of training set error rate and test set error rate with regard to the model complexity on the same plot (i.e., x-axis: model complexity, y-axis: error rate).
3. Consider a neural network with a single hidden layer. The activation functions in the hidden and output nodes are sigmoid functions. There are 2 input nodes, 2 hidden nodes, 2 output nodes, and relevant bias terms. Assign initial (e.g., random) values to the weight parameters. Given the input values of 0.2 and 0.8, output values of 0 and 1, and learning rate of 0.5; update the values of weight parameters for the first step using backpropagation.
4. Please observe the two input values x1 and x2, and the class (output) which is either star or a quadrangle. How would you transform the input variable so that a single perceptron can separate the two classes?

**More:** From the textbook *Statistics for Business and Economics, P. Newbold, W.L. Carlson, B.M. Thome, 8th ed, Pearson, q*uestions 6.60, 6.61, 7.49, 7.50, 9.49, 10.3, 10.6, 10.12, 10.13, 10.15, 10.23, 11.5, 11.7, 11.12, 11.18, 11.20, 11.25, 11.31, 11.32, 11.33, 11.39, 11.55, 12.11, 12.19, 12.28, 12.33, 12.35, 12.36, 12.41, 12.47, 12.53, 12.62, 12.63

**ps.** Solve Analysis of Variance (Ch.15) problems on your own.