02.10.2018

|  |  |  |
| --- | --- | --- |
|  | **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.

09.10.2018

1. 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.

23.10.2018

1. 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 |