We build classification model using decision tree algorithm to predict variable “Play Golf”.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Input | | | | Output |
| Outlook | Temperature | Humidity | Windy | Play Golf |
| Rainy | Hot | High | False | No |
| Rainy | Hot | High | True | No |
| Overcast | Hot | High | False | Yes |
| Sunny | Mild | High | False | Yes |
| Sunny | Cool | Normal | False | Yes |
| Sunny | Cool | Normal | True | No |
| Overcast | Cool | Normal | True | Yes |
| Rainy | Mild | High | False | No |
| Rainy | Cool | Normal | False | Yes |
| Sunny | Mild | Normal | False | Yes |
| Rainy | Mild | Normal | True | Yes |
| Overcast | Mild | High | True | Yes |
| Overcast | Hot | Normal | False | Yes |
| Sunny | Mild | High | True | No |

1. At initial state, calculate Gini impurity of variable “Play Golf”
2. At first, root node is split by variable “Outlook” and each children node corresponds to each value. After split, calculate impurity of children nodes.
3. Calculate information gain of this split.
4. Which variable is the best for further split of children node with “Rainy”?

Decision boundary is the region of a problem space in which the output label of a classifier is ambiguous. Draw the decision boundaries produced by 1-nearest neighbor classifier on the following dataset. Use Euclidean distance and shade regions where points are classified as req square.

A graph with red and blue dots

AI-generated content may be incorrect.

We want to use k-nearest neighbor classifier to predict output class of new data based on following training data set.

|  |  |  |  |
| --- | --- | --- | --- |
| Index | X1 | X2 | Class |
| 1 | 7.0 | 3.2 | 1 |
| 2 | 6.3 | 3.3 | 2 |
| 3 | 6.4 | 3.2 | 1 |
| 4 | 5.5 | 2.3 | 1 |
| 5 | 6.3 | 2.9 | 2 |
| 6 | 7.6 | 3.0 | 2 |
| 7 | 6.5 | 2.8 | 1 |
| 8 | 5.2 | 2.7 | 1 |
| 9 | 7.3 | 2.9 | 2 |
| 10 | 6.7 | 2.5 | 2 |

1. Predict output class of new data point, x1= 6.75, x2=2.85 based on given training samples when k=3 and distance measure is Euclidean distance.
2. If we change distance measure from Euclidean distant to Manhattan distance, list up all nearest neighbors and how is the answer of problem (2) changed?

In a certain day care class, 30% of the children have grey eyes, 50% of them have blue and the other 20%’s eyes are in other colors. One day they play a game together. In the first run, 65% of the grey eye ones, 82% of the blue-eyed ones and 50% of the children with other eye colors were selected. Now, if a child is selected randomly from the class, and we know that he/she was not in the first game, what is the probability that the child has blue eyes?

Logistic regression model was trained on the dataset to classify whether a patient has diabetes or not(diabetes = 1, otherwise = 0). The total number of train samples is 768.

|  |  |
| --- | --- |
| Variable | Coefficient |
| Intercept | -7.8141 |
| Pregnant | 0.1449 |
| Glucose | 0.0363 |
| Blood | -0.0118 |
| Insulin | 0.0010 |
| BMI | 0.0907 |

1. What are the degree of freedom of the model and the degree of freedom of error(residuals)?
2. Calculate odds ratio of variables “pregnant” and “glucose”.
3. Write the link function of the trained logit model with trained coefficients.
4. Calculate the probability of diabetes for the following sample

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | pregnant | glucose | Blood | Insulin | BMI |
| Value | 3 | 78 | 50 | 88 | 31 |

Using fixed radius neighbor regression, regression model is trained based on given train data.

1. When Euclidean distance is used and r=4, how many points are included into neighbor set of point (0,0)?

A graph of a blue circle with a red arrow pointing at the center

AI-generated content may be incorrect.

1. When Manhattan distance is used and r = 4, how many points are included into the neighbor set of point (0,0)?

A blue circle with black dots

AI-generated content may be incorrect.

We built Naïve Bayes classifier on the data whose 4 input variables and all binary and output variables is categorical variable with 3 different classes. Trained Naïve Bayes classifier is described in following table.

A table with numbers and a number in it

AI-generated content may be incorrect.

When new input data is (x1, x2, x3, x4) = (1, 0, 1, 0), determine output class based on the Naïve Bayes classifier.

We want to build Gaussian Naïve Bayes model.

1. Write the likelihood function and log likelihood function of Gaussian Naïve Bayes model.
2. Calculate maximum likelihood estimates of mu\_ki and sigma\_ki for Gaussian Naïve Bayes model. k represents different class, t represent feature.

Using k-means clustering, we want to group data points into two classes. Following table describes a certain iteration of k-means clustering.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 1 | 2 | 3 | 4 | 5 |
| X1 | 1 | 2 | 3 | 3 | 4 |
| X2 | 1 | 1 | 2 | 4 | 2 |
| group | 1 | 1 | 1 | 2 | 2 |

1. Calculate centroids of two groups.
2. Using new centroids, update group based on Manhattan distance.

Based on following transaction data, we want to create association rules using Apriori algorithm. Set s\_min = 0.6.

|  |  |
| --- | --- |
| TID | Items |
| 1 | M O N K E Y |
| 2 | D O N K E Y |
| 3 | M A K E |
| 4 | M U C K Y |
| 5 | C O O K I E |

1. Generate C1, L1
2. Generate C2, L2
3. What is support and confidence of rule “if K, then E”