**Problem 1**

**Calculate the information gain of A2 and A3 and determine which is better as the test**

**attribute at the root. You must show all calculations, including the calculation of info and**

**information gain.**

In the whole dataset, there are 7 Y tuples, and 5 N tuples

|  |  |  |  |
| --- | --- | --- | --- |
| A2 | Number | Y | N |
| Mild | 5 | 3 | 2 |
| Cool | 4 | 2 | 2 |
| Hot | 3 | 2 | 1 |

|  |  |  |  |
| --- | --- | --- | --- |
| A3 | Number | Y | N |
| East | 8 | 7 | 1 |
| West | 4 | 0 | 4 |

Because , A3 is better than A2

**Problem 2**

**(1). Calculate the distance between X and all 10 objects. Use the Euclidean distance.**

1. **. Classify X using five nearest neighbors.**

According to the Distance in (1), we can get 5 nearest neighbors, they are:

1. Object 4, distance = 0.3
2. Object 8, distance = 0.447
3. Object 5, distance = 0.583
4. Object 3, distance = 0.781
5. Object 9, distance = 2.657

**Problem 3**

**Coefficient of A1 = 0.045**

**Coefficient of A2 = 0.003**

**Intercept = -3.485**

**O1: <A1 = 47, A2 = 213>**

Because P is 0.325 which is smaller than the threshold 0.5, Classified O1 as “No”

**O2: <A1 = 65, A2 = 276>**

Because P is 0.5666 which is bigger than the threshold 0.5, Classified O2 as “Yes”

**Problem 4**

**(1). Write a brief, one-page description of the method.**

Discriminant Analysis is a statistical approach used for classifying observations into predefined categories. The method is used primarily when the variables are multivariate normally distributed. It is used when the dependent variable is categorical and the independent variables are interval in nature.

There are two main types of Discriminant Analysis: Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA).

Linear Discriminant Analysis (LDA) assumes equal covariance matrices for the classes. It seeks to find a linear combination of the features that characterizes or separates two or more classes. The resulting combination may be used as a linear classifier, or more commonly, for dimensionality reduction before later classification.

Quadratic Discriminant Analysis (QDA) allows for class-specific covariance matrices, meaning each class has its own covariance matrix. It is a more flexible classifier than LDA because it does not make strong assumptions about equal covariance.

Discriminant Analysis works by modeling the distribution of the predictors X separately in each of the response classes (i.e., the multivariate Gaussian distribution), and then uses Bayes' theorem to flip these around into estimates for the probabilities of the response category given the value of X. The observation is then classified into the class for which the posterior probability is greatest.

Steps involved in Discriminant Analysis:

1. Estimate the class-specific mean vector and covariance matrix for each class from the training data.
2. Using these estimated quantities, compute the discriminant function for a new observation.
3. Classify the new observation to the class that maximizes the discriminant function.

Discriminant Analysis is used in various fields including medical imaging, finance, marketing, and more. In finance, it's used to determine the set of characteristics that investors use in assessing the credit risk of a borrower. In marketing, it is used to identify distinct groups of customers for targeted marketing campaigns.

While Discriminant Analysis is a powerful tool, it does have its limitations. It assumes that the predictors are normally distributed and it requires a large number of observations for each predictor. If these assumptions are not met, the model may perform poorly.

In summary, Discriminant Analysis is a robust statistical method for classification, useful in situations where predictive variables are continuous and normally distributed. Its main types, LDA and QDA, offer different levels of flexibility depending on the assumptions one can make about the data.

**(2). Run a linear discriminant analysis method on Accidents1000 dataset using Weka,**

**JMP Pro, R, or Python**

Weka

