University of Technology - Iraq

Department of Computer Sciences – Information Systems

Branch

Data Analysis Methods

Fourth Class – Second Course

Lecturer: Asst. Prof. Dr. Hiba Basim Alwan

2024 – 2025

University of Technology - Iraq

Department of Computer Science

Data Analysis Methods | Information Systems Branch | Fourth Stage

2024 - 2025

Asst. Prof. Dr. Hiba Basim Alwan

Table of Contents

Experiment One: Types of Distributions

Introduction

There are many types of distributions, among them the following distributions are the most commonly used:

- **Frequency Distributions:** One of the most common ways to summarize a set of data is to construct a frequency table or frequency distribution. The process begins with recording the number of times a particular value of a variable occurs. This is the frequency of that value.

- **Percentage Distribution:** A frequency distribution organized into a table (or graph) that summarizes percentage values associated with particular values of a variable.

- **Probability Distribution:** It is the long-run relative frequency with which an event will occur. Inferential statistics uses the concept of a probability distribution, which is conceptually the same as a percentage distribution except that the data are converted into probabilities.

The Aim of the Experiment

This experiment aims to write a suitable code for the different types of distribution and show the student what will happen if the values of the data are changed.

The Algorithm

Begin

Read input data

Compute the frequency of the input data

Calculate frequency distribution (fd)

$$fd = frequency\ of\ each\ element\ in\ the\ input\ data$$

Disp (fd)

Calculate probability distribution (pro_d)

$$pro\_d = \frac{fd}{size\ of\ input\ data}$$

Disp (pro_d)

Calculate percentage distribution (per_d)

$$per_d = pro_d \times 100\%$$

Disp (per_d)

End

Experiment Two: Hypothesis Test about μ: σ Known

Introduction

There are three cases to perform a test of hypothesis for the population mean μ when the population standard deviation σ is known. Here there are three possible cases as follows:

Case I. If the following three conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is small (i.e., $n < 30$).
- The population from which the sample is selected is normally distributed.

then we use the normal distribution to perform a test of the hypothesis about μ.

Case II. If the following two conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is large (i.e., $n \geq 30$).
- The population from which the sample is selected is normally distributed.

then we use the normal distribution to perform a test of the hypothesis about μ.

Case III. If the following three conditions are fulfilled:

- The population standard deviation σ is known.
- The sample size is small (i.e., $n < 30$).
- The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown).

then we use a nonparametric method to perform a test of the hypothesis about μ.

The Aim of the Experiment

This experiment aims to write a suitable code for the hypothesis test about μ when σ Known and explain to the student the results of outputs.

The Algorithm

Step 1. State the null and alternative hypotheses.

Step 2. Select the distribution to use.

Step 3. Calculate the p-value.

$$\sigma\bar{x} = \sigma \,/\, \sqrt{n}$$

$$z = (\bar{x} - \mu) \,/\, \sigma\bar{x}$$

Step 4. Make a decision.

Experiment Three: Hypothesis Test about μ: σ Unknown

Introduction

There are three cases to perform a test of the hypothesis for the population mean μ when the population standard deviation σ is unknown. Here there are three possible cases as follows:

Case I. If the following three conditions are fulfilled:

- The population standard deviation σ is unknown.
- The sample size is small (i.e., $n < 30$).
- The population from which the sample is selected is normally distributed.

then we use the normal distribution to perform a test of the hypothesis about μ.

Case II. If the following two conditions are fulfilled:

- The population standard deviation σ is unknown.
- The sample size is large (i.e., $n \geq 30$).
- The population from which the sample is selected is normally distributed.

then we use the normal distribution to perform a test of the hypothesis about μ.

Case III. If the following three conditions are fulfilled:

- The population standard deviation σ is unknown.
- The sample size is small (i.e., $n < 30$).
- The population from which the sample is selected is not normally distributed (or the shape of its distribution is unknown).

then we use a nonparametric method to perform a test of the hypothesis about μ.

The Aim of the Experiment

This experiment aims to write a suitable code for the hypothesis test about μ when σ unknown and explain to the student the results of outputs.

The Algorithm

Step 1. State the null and alternative hypotheses.

Step 2. Select the distribution to use.

Step 3. Calculate the p-value.

$$s\bar{x} = s / \sqrt{n}$$

$$t = (\bar{x} - \mu) / s\bar{x}$$

Step 4. Make a decision.

Experiment Four: Linear Regression

Introduction

Linear regression is the most widely used statistical technique; it is a way to model a relationship between two sets of variables. The result is a linear regression equation that can be used to make predictions about data.

Simple linear regression is a measure of linear association that investigates straight-line relationships between a dependent variable and an independent variable. Simple linear regression investigates a straight-line relationship of the type:

$$Y = \alpha + \beta X$$

where Y is a dependent variable and X is an independent variable. Alpha (α) and beta (β) are two parameters that must be estimated so that the equation best represents a given set of data. These two parameters determine the height of the regression line and the angle of the line relative to the horizontal.

The Aim of the Experiment

This experiment aims to write a suitable code for the linear regression analysis method and show the student what will happen if the values of the input are changed.

The Algorithm

Step 1: Find the sum of independent variable data.

Step 2: Find the sum of the dependent variable data.

Step 3: Find the multiplication result between the independent variable and dependent variable then sum them.

Step 4: Find the result of squaring independent variables the sum them.

Step 5: Calculate Alpha value:

$$\alpha = \frac{(\sum y \sum x^2) - (\sum x \sum xy)}{(n \sum x^2) - (\sum x)^2}$$

Step 6: Calculate the Beta value

$$\beta = \frac{(\sum xy) - (\sum x \sum y)}{(n \sum x^2) - (\sum x)^2}$$

Step 7: Calculate linear equation

$$Y = \alpha + \beta X$$

Experiment Five: ANOVA

Introduction

ANOVA is the appropriate statistical technique to examine the effect of a less-than-interval independent variable on an at least interval-dependent variable. An independent samples t-test can be thought of as a special case of ANOVA in which the independent variable has only two levels. When more levels exist, the *t*-test alone cannot handle the problem.

The Aim of the Experiment

This experiment aims to write a suitable code for the ANOVA analysis method and explain to the student the output result.

The Algorithm

Step 1: State the hypothesis

Step 2: Calculate Means

Step 3: Calculate SSW

Step 4: Calculate SSB

Step 5: Make an ANOVA table

Step 6: Make a decision

If Calculated F value < Tabulated F value

Post hoc tests

Experiment Six: Linear Discriminant Analysis

Introduction

Dimensionality reduction techniques are important in many applications related to machine learning, data mining, bioinformatics, biometrics and information retrieval. The main goal of the dimensionality reduction techniques is to reduce the dimensions by removing the redundant and dependent features by transforming the features from a higher dimensional space that may lead to a curse of dimensionality problem, to a space with lower dimensions. There are two major approaches of the dimensionality reduction techniques, namely, unsupervised and supervised approaches.

Linear Discriminant Analysis (LDA) is a very common technique for dimensionality reduction problems as a pre-processing step for machine learning and pattern classification applications. The goal of the LDA technique is to project the original data matrix onto a lower dimensional space.

The Aim of the Experiment

This experiment aims to write a suitable code for the LDA analysis method and explain to the student the output result.

The Algorithm

Step 1: Compute the Mean for all the given data.

Step 2: Compute the statistics for the given data.

- Mean vector (M_i).
- Covariance matrix (C_i).

Step 3: Compute within-class scatter matrix C.

C = (no. of first sample / no. of all given data) × C_1 + (no. of second sample / no. of all given data) × C_2

Step 4: Generate discriminant functions (F_i).

$$F_i = M_i \times C^{-1} \times X^T - 0.5 \times M_i \times C^{-1} \times Mi^T + \ln(P_i)$$

Experiment Seven: Decision Tree

Introduction

A decision tree is a hierarchical data structure that represents data through a divide-and-conquer strategy. In classification, the goal is to learn a decision tree that represents the training data such that labels for new examples can be determined. Decision trees are classifiers for instances represented as feature vectors (e.g. colour = ?; shape = ?; label = ?;). Nodes are tests for feature values, leaves specify the label, and at each node, there must be one branch for each value of the feature. One of the most common approaches to decision trees is the ID3 approach.

The Aim of the Experiment

This experiment aims to write a suitable code for the ID3 analysis method and explain to the student the output result.

The Algorithm

Step 1: Selecting the Root node by calculating the entropy of the target variable (class labels) based on the dataset. The formula for entropy is:

$$\text{Entropy}(S) = -\Sigma\,(p_i \times \log_2(p_i))$$

Step 2: Calculating Information Gain (IG) by calculating the information gain when the dataset is split on that attribute. The formula for information gain is:

$$IG = \text{Entropy}(S) - \Sigma\,((|S - v| \,/\, |S|) \times \text{Entropy}(S - v))$$

Step 3: Select the best attribute by choosing the attribute with the highest IG as the decision node for the tree.

Step 4: Splitting the dataset by splitting the dataset based on the values of the selected attribute.

Step 5: Repeat the process by recursively repeating steps 1 to 4 for each subset until a stopping criterion is met (e.g., the tree depth reaches a maximum limit or all instances in a subset belong to the same class).

Experiment Eight: Support Vector Machine

Introduction

Support Vector Machine (SVM) is a powerful supervised learning algorithm that works best on smaller datasets but on complex ones. The SVM has been introduced as a successful statistical learning approach for classification. SVM has an extremely good generalization capability and a strong theoretical foundation. Generalization capability can be defined as the ability of SVM to classify unknown data examples correctly through constructed SVM. This is achieved by learning SVM from training examples which is also known as SVM performance. The SVM manipulates the “curse of dimensionality”, which means the computational complexity for the SVM training or testing is not affected by the feature space dimensionality.

The Aim of the Experiment

This experiment aims to write a suitable code for the SVM analysis method and explain to the student the output result.

The Algorithm

Step 1: The SVM algorithm predicts the classes. One of the classes is identified as 1 while the other is identified as -1.

Step 2: All machine learning algorithms convert the business problem into a mathematical equation involving unknowns. These unknowns are then found by converting the problem into an optimization problem. As optimization problems always aim at maximizing or minimizing something while looking and tweaking for the

unknowns, in the case of the SVM classifier, a loss function known as the hinge loss function is used and tweaked to find the maximum margin.

Step 3: For ease of understanding, this loss function can also be called a cost function whose cost is 0 when no class is incorrectly predicted. However, if this is not the case, then error/loss is calculated. The problem with the current scenario is that there is a trade-off between maximizing margin and the loss generated if the margin is maximized to a very large extent. To bring these concepts into theory, a regularization parameter is added.

Step 4: As is the case with most optimization problems, weights are optimized by calculating the gradients using advanced mathematical concepts of calculus viz. partial derivatives.

Step 5: The gradients are updated only by using the regularization parameter when there is no error in the classification while the loss function is also used when misclassification happens.

Step 6: The gradients are updated only by using the regularization parameter when there is no error in the classification, while the loss function is also used when misclassification happens.

Experiment Nine: Apriori Algorithm

Introduction

Apriori algorithms have been popularized through market basket analyses, leading to different recommendation engines for music platforms and online retailers. They are used within transactional datasets to identify frequent item sets, or collections of items, to identify the likelihood of consuming a product given the consumption of another product. For example, if I play Black Sabbath's radio on Spotify, starting with their song "Orchid", one of the other songs on this channel will likely be a Led Zeppelin song, such as "Over the Hills and Far Away." This is based on my prior listening habits as well as the ones of others. Apriori algorithms use a hash tree to count item sets, navigating through the dataset in a breadth-first manner.

The Aim of the Experiment

This experiment aims to write a suitable code for the Apriori algorithm method and explain to the student the output result.

The Algorithm

Step 1: Deciding Threshold

Step 2: Computing Support

Step 3: Forming Candidate Item sets

Step 4: Finding Frequent Combinations

Step 5: Forming Association Rules

Step 6: Calculating other metrics