Faculty of Computer &
Information Sciences
Ain Shams University
Subject: (SCC 234)
Statistical Analysis and
Applications
Year: (2nd year) undergraduate

Examiners: Dr. Hanaa Talha
Dr. Mahmoud Mounir
Offering Dept.: Scientific Computing
Academic year: 2nd term 2019-2020

Research Topic (3)

Title: Using Probability to Build Decision Trees for Classification

1. Research Description

1.1. Research short description

Statistical Analysis and Applications course has introduced you to the methods and procedures that allow us to: Explore Data and make Statistical Inference. This thesis is designed to allow you to deal with both concepts in addition to the application of statistical analysis and probability concepts in computer science field.

1.2. Research requirements

In this research thesis, you should introduce the Decision Trees concept for classification and how it is used in machine learning. After finishing this research, you **must** cover the following items:

- What Decision Tree is?
- Mathematical formulation of Decision Trees.
- What are the terms in the Decision Trees mean and the intuitions behind them?
- Explain the ID3 method to construct the decision tress how it is related to probability?
- Working numerical examples of how Decision Trees are built.
- Working numerical examples of how Decision Trees are used in classification.
- Applications of Decision Trees in real life.
    - Collecting a real dataset to be used for classification using Decision Trees (**you can get it from the internet from Kaggle or Google datasets or any other source you prefer**).
    - Exploratory data analysis for the collected dataset (**you may write a code to help you in this part using Python or R language or any other language you prefer**).
        - Analyze raw data using appropriate graphical and numerical procedures.

- Describe Shape, Outliers, Center, and Spread of datasets in the context of your research.
- Include appropriate graphical displays and numeric summaries.

- **(Bonus)** Use Decision Trees classifier to classify the data based on some aspects in the context of your collected real dataset (**you may write a code to help you in this part using Python or R language or any other language you prefer**).

1.3. Research deliverables

You are required to **submit**

1. A research thesis document as a Microsoft WORD document.
2. A Compressed (ZIP or RAR File) contains the source file of your implementation code.

Your Report should consist of a summary of your research and experiment/survey as well as your personal conclusions. The goal is to enlighten the reader with words, numeric summaries, and appropriate graphs. Use the following format:

- Title Page
- Table of Contents (Optional)
- Introduction
  - The introduction has three goals:
    a) To introduce the topic of the research thesis;
    b) To present your thesis (which is to say the particular approach or argument the thesis will make);
    c) To tell the reader how the thesis will be structured.
- Body
  - In the body of the thesis you will present the evidence and analysis that will substantiate your research. It is essential that the body of the thesis be developed in a logical and orderly fashion following the preview that you presented in the introduction. The overall goal of this section is to develop your analysis and defend your argument – it is the main part of the research thesis.
  - This logically ordered body of the thesis will consist of a series of paragraphs. Each paragraph should develop one central theme that helps you further your argument. Introduce this theme in a topic sentence; expand on the theme through the use of evidence or examples; and analyze the evidence to show how it contributes to the specific point you are making in the paragraph and to the research as a whole. Paragraphs should consist of several sentences rather than one, long sentence.

- Conclusion
    - The conclusion is designed to bring together your thesis main points and to reassert or emphasize the strength of the research. A conclusion is more than a summary, in that it is important to indicate why there is merit to your research – what has been shown as a result of your investigation or exploration of the topic.
- References

1.4 Research References

The following references may be useful for your research thesis:

- Ronald E. Walpole et al., "Probability and Statistics for Engineers and Scientists", 9th Edition, Pearson Education International.
- Gareth James, et al., "An Introduction to Statistical Learning with Applications in R", Springer.
- Brian Caffo, "Statistical Inference for Data Science", Leanpub.
- https://www.kaggle.com/
- https://datasetsearch.research.google.com/
- https://www.tensorflow.org/

NB:

The form of the research should follow the following rules:

a. Writing using Microsoft Word.

b. using Time New Roman font.

c. 14 Font size for text and 16 Bold for the title.

d. single space between the lines.

e. Page margins are 2.5 cm from top, bottom, right and left.

f. The number of words is not less than 3000 words.

h. English language formulation should be correct and sound.

i. Clarity of texts, pictures or drawings.

With My Best Wishes,
Dr. Hanaa Talha
Dr. Mahmoud Mounir