-

Homework Assignment 5 (week 5)

Faculty Feedback

Score   Last published: 8/11/2019 5:31 PM PDT

100 / 100 (100.00%)

Comments

Overview.

WOW -- Very good work and please review the following.

While you did a really excellent job - all assignments can always be improved. The following will help you to think about where you might add or increase.

Below, I have included some general notes and some specific notes. For the general notes, the most useful method to employ is to compare your assignment to the notes below and to think about what is there, what is missing, and what is medium. In other words - even if its all there, good, and correct, can it be better?

The specific notes will tell you what I recommend and/or what I see.

Introduction: (2 - 3 paragraphs)

Specific:

WOW! Great intro. You have really set the stage and have clearly identified the problem and have provided a clear segue to the investigated method – Decision Trees.

General: An Introduction is about the area or topic, not about the data or models. The introduction helps the reader to understand what the assignment area is about. For example, support the assignment is about schools. In this case, the introduction is about school systems, why schools are measured and ranked, who might be concerned with school measures and rankings (such as students, parents, states, governments, and funding agencies), and the value of comparing schools.

An introduction is a like a warm-up or like dating. It allows the reader to "get to know" the area of interest. :)

The Intro should not contain any detailed information about the dataset or the data cleaning, prep, processing,etc. Everything about the dataset goes into the Analysis section under the "About the Data" subsection.

Data, Analysis, and Models

Specific:

GREAT VIZ and discussion of data format, cleaning, and MORE!!

AMAZING --- GREAT INVESTIGATION OF MULTIPLE MODELS. Model results can vary significantly based on parameterization and learning schemes! Great job here!!

General: The Analysis section contains subsections.

The first subsection is "**About the Data"** which contains all the information about the dataset, the variables, the cleaning and prep, checking for an dealing with missing values, checking for and dealing with incorrect values, checking for an dealing with outliers, feature generation, normalization (if needed), etc. In this subsection, you will also "explore" the data. This means that you write about each variable, visualize each variable, and talk about what the variable represents. Tables are great for this as well.

The second and remaining subsections of Analysis are the model(s).

In some cases, there may only be one model. A model is any method used to analyze the data. Each Assignment specifies which models to use. Always include model details and parameter values when applicable.

Results

Specific: You did a good job here!

Great Viz of results as well!

General: The Results section of the Assignment will have a subsection for results for each model (assuming that you have more than one).

Results are technical. They offer technical information about what was found in the analysis. For example, if you performed a correlation in the analysis between all pairs of numeric variables, then your results would discuss the r-value and relationship of each pair. Similarly, if you looked at measures of center and variation, the results talk about what those

measures are and what they reveal. For example, if the mean is less than the median, the data is skewed, which means....

Each model we will use in this class has results and parameters associated with it. For example, association rule mining will offer the top ten rules for sup, conf, and/or lift if you you code it to do so. These would go into the results along with the sup, conf, and lift for each rule. The meaning would also be discussed.

Conclusions

Specific:

WOW!!! Great work here. Short but sweet!

General : 2 - 3 paragraphs.

This area is not technical at all.

This area explains what was actually found in a way that would make sense to anyone. For example, if your discovered in the analysis that association rule mining with a conf of .2 and a sup of .3 offered 10 rules, you would talk about the measures and values and rules in the *results*. In the Conclusions, you would talk about what it all means. So you would not include the rules themselves or mention of technical measures such as conf or sup. Rather, you would say that you found (as a random example) that people who buy diapers are very likely to by beer and that this means that a store should consider placing these items "near" to each other.

Basis For Grades:

100: This means that your Assignment was amazing and so perfect that nothing can be improved. It covered everything – cleaning, prep – analysis that makes sense – visualizations – results (that are true) – etc. There is nothing really left to improve.

95: This means that your Assignment is really good! You covered most of the items noted below and perhaps a few others not noted. You can make some improvement on pre-processing and results analysis, as well as perhaps other visualizations. Overall – you have the idea and you did well.

90: This means that your Assignment is good, but could be a little better. Perhaps add items such as further data cleaning and pre-processing, data normalization, better or more visualizations, and/or more robust conclusions. Many students forgot to change Section to a

factor for example. Very few students summed and normalized the data to look at the percentages for each attribute.

85: This means that your Assignment is a good start and largely meets the more general and overall requirements. Here, you used R, you did some analysis, you did some cleaning, you made some graphs, and your reached some conclusions. However, there is room for improvement.

Below 85 means that the level of 85 above was not quite met and many elements were missing.

Extra Assignment – Specific Notes:

You will need to separate the original data set to **training and testing data** for classification experiments.

Build and tune decision tree models

First build a DT model using default setting,

and then tune the parameters to see if better model can be generated.

Compare these models using appropriate evaluation measures.

Describe and compare the patterns learned in these models.

Use prediction After building the classification model,

apply it to the disputed papers to find out the authorship.

Does the DT model reach the same conclusion as the clustering algorithms did?

Student Submission | Homework Assignment 5 (week 5)

Response   Last submitted: 8/5/2019 8:34 PM PDT

No response

Files | Download all (3)

File Name	Uploaded	Feedback
Ryan_Timbrook_HW5.docx	8/5/2019 8:33 PM PDT	!
ist707_week5_hw.R	8/5/2019 8:33 PM PDT	?
rtimbroo_util.R	8/5/2019 8:34 PM PDT	?