

[← BACK TO GRADES](#)

Homework 4

Faculty Feedback

Score

Last published: 11/1/2019 7:10 AM PDT

95 / 100 (95.00%)

Comments

Week 4 Assignment Rubric and Explanations

Ryan - Well done and please see my scores and comments below. This assignment offers an excellent opportunity to practice with reading in and vectorizing data. It also requires you to make sure that the data is cleaned - such as lowercase, punctuation removal, etc.

Next, Naive Bayes is a supervised learning method. So, it requires labeled data. As a next step, after vectoring the data, one must be sure that the format of the data is Data Frame (pandas in Python) and has an added column that is the label.

Remember that the data is all numeric and it is also best to normalize it. The Label (not part of the dataset itself) will be categorical - such as Pos, Neg, Med (or any other).

Finally, to train and then test any supervised method, such as NB, the dataset must be separated into a Training and a Testing set. The model is trained on the training data and then tested on the test data. A confusion matrix is best to see and describe and discuss the results.

You can also start to think about ways to update the data to make it perform better. NB assumes independence and so columns that are highly correlated can be removed or combined.

See below for **Basis for Grades – what grades mean.**

Grade: 95/100 A**Grade explanation....**

Telling a cohesive, professional, clear, and well-thought out story.	
---	--

Here, a 9/10 means that you told a very nice story. A 9/10 is a very excellent grade and means that you are right on track.

9.5/10

Part of this grade is an over measure of flow, look and feel, and readability. Have you created clear separations (subsections and headings)?

Do you use a lot of tables for clarity?

Do you use a lot of visualization for illustration?

Are you not placing code or code output in your paper? Code and code output is generally not appropriate in papers. You can sometimes do this IF you create a figure and properly note key elements, etc. It should never be a direct copy/paste.

Does your paper tell a cohesive story, does it have a nice flow, do you look at, discuss, stats-analyze, and visualize (explore) the variables, etc. etc.

Introduction: (2 - 3 paragraphs: 3 is better.)

For a relative comparison, an 18.5/20 means you met all the key and core requirements and did a very nice job. Can you get 19 or 20/20? Sure – be creative, be more awesome, exceed, think about side the box.

19/20

An Introduction is about the **area or topic**, not about the datasets, variables, methods, or models.

For example, if your data focuses on restaurant reviews, then this is your topic. You might first talk about how often American eat at restaurants. Then you can talk about how much money is spent eating out. Then you might navigate into how the use of reviews has become common and how such reviews affect restaurant choices. From here, you can talk about the dangers of fake reviews to both restaurant goers and restaurant owners, etc. etc...

The introduction helps the reader to understand what the topic of the data area is about, who it affects, and why it matters.

An introduction and conclusion should always tie together. *You* should always understand what information you discovered (or predicted), what it means, who it affects, why it matters, what the

relevance and applications might be, and how to explain this to non-tech people.

Writing skills are very important in real-life. In fact, it is often the case that a first impression is based on writing. Writing in this class should be professional, at a high level of skill, and never about you. Avoid using “I”, “me”, “my”, “you”, “us”, “we”, etc. Write about the topic, methods, models, results, etc. Do not write about yourself.

A professional or graduate-level paper should never have a “studenty” or “undergrad” feel, such as using sentences like “My Assignment is about...”, or “I am going to talk about...”, or “In this Assignment, I will....”. At the graduate level, write Assignments as though they are technical papers.

The Intro should never contain any information about the dataset variables, where the dataset came from, the data types, or the data cleaning, prep, processing, etc. Everything about the dataset and data wrangling goes into the Analysis section under the “About the Data” subsection.

Analysis and Models

A 23/25 means that you did well – very well – but you can do more. Generally, more includes more data EDA, feature generation, normalization (with discussion), transformation, aggregation, detailed cleaning WITH the before and after measures, an easy-to-follow flow – so the viewer can quickly see (so tables) what you did – etc...

The Analysis section always contains two or more subsections.

The first subsection in Analysis is “**About the Data**” which contains all the information about the dataset, the variables, the types, visual and statistical EDA, the cleaning and prep, checking for and dealing with missing values, checking for and dealing with incorrect values, checking for and 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).

24/25

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. You can always use more, but never fewer. If a model or method is not specified, start simple and go as deep as you can.

Always include model details and parameter values when applicable.

***** Have Visualizations throughout the assignment. Use color.**

There are 1000s of vis options so while bars, pies, and box-plots are great, have many others as well.

Visualizations should have titles, proper labeling, and should have a Fig. # with a very short explanation.

Results

For a relative comparison, an 23.5/25 means you met all the key and core requirements and did a very nice job. Can you get a higher grade here? Sure – be creative, exceed, think about the story, the flow, the readability, the measures the use of tables and vis (and color and clarity), etc.

If your results are a great start, but are just way to short, this means that you did well with what you have, but did not cover a lot of things you could have. Results are generally a few pages (not paragraphs). Results have tables, comparisons, vis, discussion, exploration of attributes, measures of accuracy, etc, etc,

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....

24.5/25

Each model we will use in this class has results and parameters associated with it.

Use a lot of vis and tables for clarity and illustration.

Results should include comparison tables and/or confusion matrices.

** Always have lots of visualizations

Conclusions

3 paragraphs. This is directly connected to the Introduction – but – it now explains what was learned via the analysis and results. However, unlike the Results, this is non-technical and very focused on why humans care, how humans can use this information, what this all means to humans, etc.

Throughout this class, your conclusions (and your intro) should get better and better. Think about what that means. Think about where you are now. (If you got an 18/20 – this means you met all the requirements and did a good job. A higher grade means you did that and a bit more – like better flow – interesting comments – more detail, better detail, more creativity, better use of visualization – etc. etc.) If you are below 18/20, you are not quite there yet. This could mean that your conclusions contain some technical matter that really belong in the Results. It could mean that your conclusion is a good start, but really could do a better and more thorough job of telling the story. Etc. Think about it.

A data analysis story is like a picture that you have painted. Everyone can paint a different picture. Everyone can choose colors, styles, flow, use of space, etc. Interestingly, some pictures represent better than others. Why? Learning to answer this is part of the learning experience. So, think about where you are, what you did well, and what you can improve. Be creative – you are the artist!

This area is not technical at all. If you are technical, such as using works like k means, kernel, linearly related, rules, sup, etc. You can lose points. When in doubt, make it more clear and less complicated. Tell the reader what happened and how it affects real-life. Keep the “results” in the Results section.

18/20

Ryan - everything here is perfect for Results.

Always ask yourself if a 10-year can read and understand (and appreciate) what you did.

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 buy beer and that this means that a store should consider placing these items “near” to each other.

Conclusions also help readers to make decisions based on the research that you did. For example, a university group might perform protein assay studies to understand amino acid digestion systems in the gut. However, the conclusions will not talk about assays J Instead, it will talk about what the results IMPLY...such as the need for those with ulcers or GERD to take glutamine alone (not with other protein) to assist in healing...

Basis For Grades:

99 - 100: This means that your Assignment was so 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. I generally do not give 100% grades as this does not help you to think about what you can add or do better based on where you are. Everyone can always get better – this is why we seek education.

92 - 98: This means that your Assignment is really good! You covered most of the items and areas noted 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. A grade of 97 or 98 means that dazzled me with your awesomeness. This is GOOD GRADE! A 93 is an A grade that means that you did very well, met the requirements, put in the time, etc. If you want more than a 93/A, think about what

more you can do – how you can exceed and base-expectations, how and where you can apply more (and better) vis options, how you can use tables, etc. etc.

89 - 91: This means that your Assignment is very 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 - 88: 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.




Student Submission | [Homework 4](#)

Response

Last submitted: 10/29/2019 10:09 PM PDT

Note that the jupyter notebook uses functions from the rtimbroo_utils_hw4.py code file

Files | [Download all \(4\)](#)

File Name	Uploaded	Feedback
HW_4_Timbrook_Ryan.docx	10/29/2019 10:07 PM PDT	
hw4_multinomial_nb.ipynb	10/29/2019 10:08 PM PDT	
rtimbroo_utils_hw4.py	10/29/2019 10:08 PM PDT	
Timbrook_HW4_data_code.zip	10/29/2019 10:08 PM PDT	