Data selection guidelines

April 18, 2024

1 Guidelines for dataset selection for the final assignment

1.1 Most important criterion for selecting a dataset:

Chose a dataset that you are genuinely passionate and curious about.

Other necessary criteria to consider:

1.2 Data fit for the purpose

In the final assignment you should implement EDA, Descriptive Statistics and all the ML algorithms you have learned in the course.
Therefore, the dataset should be suitable for this. Specifically, you need at least two kinds of datasets:

a) a dataset in which the target variable is a continuous, quantitative variable.

b) a dataset in which the target variable is a categorical variable that you can use for classification. It may be binary or multi-class classification.

c) For clustering, you can use any of the two datasets. Alternatively, you may use a third dataset. Also, you may implement clustering on both of the datasets you used for classification or regression. Just remove the features' labels.
In this case, clustering will become a useful tool for data exploration and understanding of the data. This will need a slight modification of the "table of contents" of each file in the final assignment.

Extra things to consider:
- Ranked Ordered logistic classification has not been covered in the course. If you chose such a dataset, take this into account and select the proper algorithm.
- Athletes or artists names are not a suitable target variable for classification for the purpose of this assignment. Think of them as a unique category on their own. As Monty Pythons say: "*You 're all individuals. You 're all different*".

1.3 Data size

The dataset should be large enough to be able to apply the ML algorithms you have learned.
Be mindful about the width and the height of the dataset. To allow for better explainability there should be at least six features besides the target variable in the dataset.
That means seven columns in total. As a rule of thumb, the number of rows should be at least 10

times the number of columns. That is a barely acceptable starting point. For this assignment, the dataset should have at least 300 rows. It really doesn't matter if is a bit less than 300.

The optimal "observations to features ratio" varies greatly depending on the specific dataset, the complexity of the model, the noise in the data, and factors such as the desired efficiency and error-tolerance of the "users".

Non-linear growth of necessary observations:
In some situations, especially when dealing with high-dimensional feature spaces or highly complex models, the number of observations required may increase **exponentially** with the number of features. This is because more features can lead to a higher-dimensional space where the data become sparser. This requires more observations to effectively capture the underlying patterns.

1.4 Data quality

The dataset should be clean and well-structured.
You should be able to understand the meaning of each column.
For example, if you are using a dataset from Kaggle, you should read the description, and select a dataset that with "Usability Rating" greater than 8.

1.5 Don't select datasets that have been used in the course or are tutorials in python libraries.

Here is the list of data used in the lectures:
- Iris
- Wine
- Breast Cancer
- Boston Housing - Diabetes
- MNIST

There are many datasets concerning the topics above. You may use any of them except the ones used in the course.

1.6 Don't select datasets that have been selected by other students

Please check the list of datasets selected by other students.
The first one to select a dataset will be the one to use it.

1.7 Don't select datasets suitable for algorithms that have not been taught in the course.

No time series data (y as a function of time, and its value at a given time depends on the values at previous times).
No text data, no image data, no audio data.

1.8 Selecting the dataset.

Enter your selection in this link:
Dataset Selection

Email me for permission to edit the document.

Fill in the available fields. There is a distinct sheet for each group of algorithms.

The datasets will NOT be approved by the teacher.
Of course, I am at your disposal to help you with any questions you may have.
But, as stated in the assigment, chosing the dataset is part of the assignment.

Chose wisely and most importantly:
Select a dataset that you are genuinely passionate and curious about.

Thank you very much for your participation, this year's class was memorable and exceptional. I hope you have learned a lot and enjoyed the course. I learned a lot from you, and you helped me improve considerably. I am very grateful for that.