FUNDAMENTALS OF DATA SCIENCE – DSC 441

HOMEWORK 5

Turn your assignment in as a PDF file to the D2L submission folder by the due date associated with that folder (unless otherwise specified explicitly by the Instructor). Late assignments will incur a penalty. Please ask if you need an extension and come to office hours if you need assistance.

Academic integrity is key to maintaining the impact of your degree. You may discuss the problems and methods with others, but everything you submit must be your own work. For violating, you may incur significant sanctions including failing the course or penalties from the University. If in doubt, ask first.

To get credit, you must clearly label your responses with the questions they answer. Include results and visualizations that you generate, and clearly written answers to any questions included in the problem. Your code file must be submitted alongside your document in D2L.

This final homework assignment is more open ended, covering material from the entire course, notably including the advanced evaluation techniques of Week 9. You will clean data, perform exploratory analysis, discover clusters, make predictions and evaluate them.

Single Problem (100 pts – see rubric)

For this assignment, you may pick data of your choice (given the rules below). You will perform all the major steps of data mining and report your results. The goal here is not an extensive report about the data, but practice applying the whole pipeline from start to finish. You have done all these steps already, but now you get to make choices yourself. That will take longer per step, but we are only going through the pipeline once on one set of data. This assignment is not meant to take longer than the previous ones. If you find it is substantially more work, consider switching to a simpler dataset or asking for guidance.

The parts of this *Problem* correspond to the parts of the data mining pipeline, explaining the requirements of the assignment for each one. There is a rubric below that will be used to grade your submissions. The actual deliverable is just the report with each of the steps labeled and described (i.e., include the labels *a-i* in your report document.)

a. *Data gathering and integration*

The first part is to get the data you will use. You may use anything that has not been used in an assignment or tutorial. It must have at least 100 data points and must include both numerical and categorial (or ordinal) variables. I recommend keeping this relatively straightforward because data cleaning can take a lot of time if you choose a large, messy dataset. *Kaggle* (https://www.kaggle.com/datasets) and the *University of California at Irvine (UCI)* (https://archive.ics.uci.edu/ml/index.php) maintain collections of datasets, some even telling you if they are good examples for testing specific machine learning techniques. You may also choose to join together more than one dataset, for example to merge data on health outcomes by US state with a dataset on food statistics per state. Merging data is not required and will earn you a bonus point in this step.

b. Data Exploration

Using data exploration to understand what is happening is important throughout the pipeline, and is not limited to this step. However, it is important to use some exploration early on to make sure you understand your data. You must at least consider the distributions of each variable and at least some of the relationships between pairs of variables.

c. Data Cleaning

Don't forget – this can take a lot of the time of the whole process. Your cleaning process must ensure that there are no missing values and all outliers must be considered. It may be reasonable to just remove rows with missing values, however, if your data or small or that would change the distributions of the variables, that will not be adequate and you will need to consider other options, as discussed in the modules on cleaning.

Depending on your data and what you plan to do with it, you may also need to apply other processes we discussed. For example, clean up strings for consistency, deal with date formatting, change variable types between categorical and numeric, bin, smooth, group, aggregate or reshape.

Make the case with visualization or by showing resulting summary statistics that your data are clean enough to continue with your analysis.

d. Data Preprocessing

In some cases, preprocessing is absolutely necessary. It is rarely a bad idea. Make the case for what is and is not necessary given what you plan to do with the data. This could include making dummy variables, applying normalization, binning and/or smoothing, and other transformations (see course module).

e. Clustering

Remove any labels from your data and use clustering to discover any built-in structure. Use an appropriate method to determine the number of clusters. If your data have labels, compare the clusters to those labels. If not, visualize the clustering results by making a PCA projection and coloring the points by cluster assignment. Note that PCA only works for numerical variables, so if your data have just a few categoricals, you may skip them. If there are many, use dummy variables or choose a different method for making a projection. One way is to make the distance matrix first (we covered a method for distance matrices using categorical variables in the clustering tutorial) and then apply PCA to that matrix. This is actually a way to calculate an MDS projection, a very popular method.

f. Classification

Use at least two classifiers to predict a label in your data. If a label was not provided with the data, use the clustering from the previous part. Follow the process for choosing the best parameters for your choice of classifier. Compare the accuracy of the two.

g. Evaluation

Using the better classifier from the previous step, perform a more sophisticated evaluation using the tools of Week 9. Specifically, (1) produce a 2x2 confusion matrix (if your dataset has more than two classes, bin the classes into two groups and rebuild the model), (2) calculate the precision and recall manually, and finally (3) produce an ROC plot (see Tutorial 9). Explain how these performance measures makes your classifier look compared to accuracy.

h. Report

In a single document, include the answers to all of the parts of this *Problem*, including this one. The *report* component specifically is about your overall takeaways from your data. What was interesting from your analysis?

i. Reflection

The final section of the report is a (short) paragraph reflecting on the course as a whole and what you have learned. The goal is not actually feedback for the course but to get you to think back about what you have learned and how your perspective on data science has changed.

Rubric

CATEGORY	PROFICIENT	MODERATE	INSUFFICIENT
DATA	Appropriately cleaned data with verification of outliers and missing values processed. Discussion and execution of other necessities (8-10) +1 for merging	Fairly clean but missing some information on the process or verification is unclear. Discussion of further measures insufficient. (4-7)	Data are not properly cleaned and may still contain missing values or outliers. Process not adequately described or implemented. (0-3)
EXPLORATION	Visualizations and summary statistics have been used to evaluate individual distributions and relationships between pairs. Appropriate visualizations chosen and properly executed (8-10)	Visualizations and summary statistics were used but some distributions or relationships that were necessary were missing and/or there were choice or implementation errors (4-7)	Significant errors in using visualization and summary statistics to explore the data, including completeness or execution (0-3)

CLEANING	Appropriately chosen, justified and executed cleaning operations (16-20)	Data have been cleaned reasonably well but there are issues with choice or justification of methods or execution (7-15)	Significant issues in choice or execution, justification missing or not appropriate (0-6)
PREPROCESSING	Appropriate choice of preprocessing methods, properly justified and executed (8-10)	Reasonable choices, justification and execution but some incompleteness or errors in application (4-7)	Significant issues in preprocessing, including missing important steps or executing incorrectly (0-3)
CLUSTERING	Correct use of clustering and choice of parameters. Necessary preprocessing justified and executed properly. (8-10)	Issues with the application of clustering or method of choosing parameters. Potentially insufficient processing for chosen clustering (4-7)	Significant issues with the application of clustering, choice of parameters. Lack of necessary processing. (0-3)
CLASSIFICATION	Two models used correctly and properly tuned (8-10)	A model is missing or the tuning was not done properly (4-7)	Significant issues in both applying models and tuning. (0-3)
EVALUATION	Correct 2x2, Precision Recall and ROC with complete explanation of the difference (8-10)	Improper calculation or graphing. Incomplete or incorrect explanation. (4-7)	Calculations are incorrect and explanation is incorrect or unclear (0-3)
REPORT	Submitted document includes all components, labeled for clarity. Overall report clarifies results from the rest of the analysis. (8-10)	Components are missing or they are not easy to find. Overall report has some issues with incompleteness or inaccuracy (4-7)	Components are missing or unlabeled leading to confusion. Report is missing, unclear or incorrect (0-3)
REFLECTION	Considered reflection demonstrating learning from the course (8-10)	Reflection demonstrates misunderstanding or is incomplete (4-7)	Reflection missing or points to significant misunderstanding of the course (0-3)
TOTAL	100		