Week 5: Classification Models and Clustering

Introduction

Contents of the Week

- Introduction
- Readings
- Supplemental Materials
- 5.1 Discussion/Participation
- 5.2 Exercise: Build your own Sentiment Analysis Model

Topics

- A Data preparation/cleaning
- B Training classification models
- C Evaluating classification models
- D Interpreting classification models
- E Clustering

Readings

Read the following:

- Chapters 14-19 of *Machine Learning with Python Cookbook*

Supplemental Materials

All of the materials below are from external sources. Authorship and ownership are indicated within the sources themselves.

Readings | **Videos**

- How to Choose Evaluation Metrics for Classification Models
- Introduction to Classification Algorithms
- The Best Metric to Measure Accuracy of Classification Models
- Classification Accuracy is Not Enough: More Performance Measures You Can Use
- Scikit-Learn Logistic Regression
- An Introduction to Clustering and Different Methods of Clustering
- Clustering in Machine Learning

5.1 Discussion/Participation

Here are optional topics for discussion via Teams this week. Remember, these topics aren't required, but if you are struggling to know what to post about, these can be used to initiate discussion!

1. What is meant by a classification model?
2. What are some common types of classification models?
3. What is the difference between a regression model and a classification model?
4. What are some common metrics for evaluating classification models?
5. Is accuracy always the best metric for evaluating classification models?

6. What are precision, recall, and the F1-score of a classification model?
7. Why is it important to take the context of the problem into consideration when deciding the most important metric(s) to evaluate you model?
8. What is the difference between supervised and unsupervised learning?
9. How do you evaluate unsupervised learning models?
10. What is clustering?
11. What are the different types of clustering?
12. Compare KNN and the K-means algorithms?

5.2 Exercise: Build your own Sentiment Analysis Model

You will build a model with the movie reviews dataset that you worked with in Week 3: Bag of Words Meets Bags of Popcorn.

1. Get the stemmed data using the same process you did in Week 3.
2. Split this into a training and test set.
3. Fit and apply the tf-idf vectorization to the training set.
4. Apply but **DO NOT FIT** the tf-idf vectorization to the test set (Why?).
5. Train a logistic regression using the training data.
6. Find the model accuracy on test set.
7. Create a confusion matrix for the test set predictions.
8. Get the precision, recall, and F1-score for the test set predictions.
9. Create a ROC curve for the test set.
10. Pick another classification model you learned about this week and repeat steps (5) – (9).

Submission Instructions

Click the title above to submit your assignment.

This exercise is due by Sunday 11:59 PM.

Submit your code, output, and answers at the link above. Comment all your code and answer any questions that are asked in the instructions. It is perfectly fine to answer a question by displaying output from your code, but make sure you are displaying the appropriate output to answer the question. I would recommend using and submitting a Jupyter Notebook, but this is not required.

View the rubric for this Assignment by clicking on the link below:

Exercise Rubric