HW 2 Machine Learning

Deadline: July 23, 11:59pm

Homework description (100pts)

In this homework, each student will solve their own machine learning problem using Kaggle dataset for Amazon reviews using 3 different machine learning algorithms.

1. **75pts** For the dataset, you must apply *neural networks* and *decision tree* algorithms, and another algorithm of your choice. Create your own classifier in python for sentiment analysis on Amazon reviews. Your classifier will be graded based on its accuracy classifying a new set of test Amazon reviews.

2. **25pts** Each student must prepare a video (max. 5 min duration) to present the findings using a presentation of 3 slides.
    a. Provide a discussion of the problem you are solving, how you set up the data, comparison of the results from the three learning algorithms.
    b. Visualizations of your results including these 3 performance metrics (ROC curve, accuracy, precision/recall)
    c. Explain why you chose the third machine learning algorithm and what conclusions you were able to prove based on your results
    d. Each student must prepare a poster and present it on July 25.

Data Sets

The following **datasets must be separated into test and training data** as follows: every 5th sample belongs to test data, the remaining samples belong to training data.

Amazon review sentiment analysis: *https://www.kaggle.com/bittlingmayer/amazonreviews*

This dataset is an extract from the Amazon Reviews Kaggle competition. The goal is to perform sentiment analysis to determine whether a review is positive or negative. We have provided a CSV file on D2L which contains the binary label (positive/negative) and the corresponding text for the 400,000 reviews.

Submission:

You must submit by July 23 11:59pm:

1. A classifier in python for sentiment analysis on Amazon reviews. (Vocarem submission instructions provided on July 17.)

2. The link to your 5-min video which describes your dataset, preprocessing, classifications, results, and discussion of the problem you are solving, how you set up the data, comparison of the results, and explanation of what conclusions you were able to prove.
3. your 3-slide presentation which covers
    a. the dataset and problem
    b. 3 machine learning techniques you used and why
    c. visualizations of your results including these 3 performance metrics (ROC curve, accuracy, precision/recall)