HW 3 Machine Learning

Deadline: July 16, 11:59pm

Homework description (100pts)

In this homework, each student will solve their own machine learning problem using one of the three given datasets and explain the results of 3 different machine learning algorithms. You will work in groups using several machine learning algorithms.

1. 3 students should form a team, and sign up on the Google Form for one data set (choice of either 1, 2 or 3). We only allow 15 teams per data set. If there are already 15 teams signed up for one data set, no more teams can select that. If you have problems forming a group, please use Piazza to find team members.
   Please use the following Google Sheet to sign up your team:
   https://docs.google.com/spreadsheets/d/1VKf-uFAcerdW2eFAqH2uSkhglTB1eJrcl0BpE0ZLIk8/edit?usp=sharing

2. **40pts** For every dataset, you must apply *neural networks* and *decision tree* algorithms. Each member in the team should try one more different machine learning algorithm that has not been applied by another team member, such that each team runs 5 different prediction algorithms in total.
   a. The whole assignment must be implemented in Wolfram/Mathematica. We recommend using the Mathematica Desktop version on your machine for best performance. Please use the newest version to ensure everyone can run each other's code.
   b. Provide a discussion of the problem you are solving, how you set up the data, comparison of the results from the three learning algorithms.
   c. Visualizations of your results including these 3 performance metrics (ROC curve, accuracy, precision/recall)
   d. Explain why you chose the third machine learning algorithm and what conclusions you were able to prove based on your results

3. **20pts** Each student must prepare a video (max. 5 min duration) to present the findings using a presentation of 3 slides.

4. **35pts** Each student must prepare a poster and present it either on July 18 or on July 20.

5. **5pts** Each student must prepare 3 true/false or multiple-choice questions that can be used to assess a student's knowledge of machine learning techniques used in this project. We will select the best questions to appear on the final exam.

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.

For example, if `data` is a `List`, the two following Mathematica commands separate it into test and training data:

```
test  = Take[data, {1, -1, 5}]
train = Drop[data, {1, -1, 5}]
```

1. CIFAR-100

http://datarepository.wolframcloud.com/resources/CIFAR-100_1

This is a computer vision problem where you need to classify 50,000 images (32 by 32 pixels) into 100 possible categories.

2. Tornadoes in the U.S., 1950-2015

https://datarepository.wolframcloud.com/resources/United%2BStates%2Btornadoes%2B1950-2015

The goal of this problem is to predict the magnitude of tornados based on the F-scale rating (F0 to F9). While there exists a clear description of the F-scale methodology, it will be interesting to see how the different classification algorithms stack up against the defined metric.

3. 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.

The CSV file can be read using the command:

```
Import["reviews.csv", "Table", FieldSeparators -> "|"]
```

Software tools

Mathematica installation:
https://itservices.usc.edu/mathematica/
Wolfram Language:
https://wolframlanguage.org
Mathematica tutorial:
http://www.math.mtu.edu/~msgocken/pdebook2/mathtut2.pdf

Submission:

You must submit by July 16 11:59pm:

1. the link to your 5-min video
2. Mathematica notebook with 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)
4. your exam questions (TF or multiple choice).

We will later send out a Google Form where you can submit your homework.