CMSC 478 — Spring 2017 — C. S. Marron
Projects

Contents

Challenge Problems

- Challenge 1: Noisy Sensor Data

The object of the first challenge problem is to predict the location of an object from noisy sensor data. You will need to build two predictive models: a regression model and a classification model. The regression model should predict the *distance from the origin* of the object, and the classification model should determine whether the object is close to the origin (within three miles).

Training Data
There are seven training files, `path1_data.csv` through `path7_data.csv`, corresponding to seven different paths through the sensor area. Links to the files are given below. Each file has the following variables:

1. $S1, S2, \ldots, S13$ are measurements from the 13 sensors
2. x, y are the x and y coordinates of the object at the time of the measurements
3. r is the distance to the origin ($\sqrt{x^2 + y^2}$) at the time of the measurements

Your best predictive model and code to create the model are **due on Tuesday, November 21**.

- path1_data.csv
- path2_data.csv
- path3_data.csv
- path4_data.csv
- path5_data.csv
- path6_data.csv
- path7_data.csv

Test Data

The test data consists of a single file of sensor observations ($S1$ through $S13$) *without* the x, y, and r values. Use your predictive models to estimate the distance of each observation from the origin and to classify the observation as “close” if it is within three miles of the origin.

For each observation in the test data, predict r and classify the observation as “close” (`TRUE`) or “not close” (`FALSE`). Save your predictions in a data frame with variables `r` and `close` and export the data frame to a csv file. The file must be readable in R. Your predictions are **due on Thursday, November 30**. If you have changed your predictive model substantially since it was originally submitted, turn-in the new model as well.

- test_path_data.csv

- Challenge 2: Sequence Classification

Sequence classification is a predictive modeling problem in which you have some sequence of inputs over space or time and the task is to predict a class label for the sequence. So the challenge is a type of pattern recognition task that involves the assignment of a class label to each sequence in a large collection of observed sequences.

We have a multi-label representation of a sequence of characters from a text line followed by a class label. We need to develop a sequence modeling classifier to predict the class label of the sequence. A class label may be “Regular-Text”, “Top-level Section Header”, “Subsection Header” or “Sub-subsection Header.”

Training Data
The dataset is generated from PDF documents. Each row is a text line with its class label. Each row has 101 columns: the first 100 columns are the predictors or features generated from a text line based on the character sequence, and the last column is the class label. In the training data file, the features are named `f0` to `f99` and a class label is named `class`. Each of the sequences will belong to one of the following four classes:

0: Regular-Text
1: Top-level Section Header
2: Subsection Header
3: Sub-subsection Header

Train a model to classify each text sequence. Your best predictive model and code to create the model are **due on Thursday, December 7**.

- train.csv

Test Data
Unlabeled test data will be provided at a later date. You are to use your model to predict the class for each line in the the test set. Your predictions for the test data are **due on Tuesday, December 12**.

Project Policies

General Project Rules

1. You *may* freely discuss concepts and ideas for the project, short of sharing actual code.
2. You *must* document all help that you receive.
3. You *must* cite all reference materials used.

You MAY NOT

1. Copy anyone else's code,
2. Have someone else write your code for you,
3. Submit someone else's code as your own,
4. Look at someone else's code, or
5. Have someone else's code in your possession at any time.

See the Acadmeic Conduct portion of the syllabus for more more informtion, including the penalties for academic misconduct.

Project Submission

Projects will be submitted on Blackboard.

Late Policy

The due date and time for each project will be found on the course schedule and in the project description.

If a project is late, the following grade penalties will apply:

- Up to 24 hours late: 15 percent reduction
- 24 to 48 hours late: 40 percent reduction
- More than 48 hours late: no credit

> All due dates and times, and assessments of late penalties, are firm.