Data Mining Course Project

Instructor: Lei Yang

Department of Computer Science and Engineering, UNR – Fall 2016

Mon, Wed 5:30PM - 6:45PM, MS 227

Project

- **One key goal** of this course is to take advantage of your intelligence and (limited) experience (so you're audacious and creative) to expand your knowledge in creating something useful and interesting
- Group project
  - 3 students per group
  - 2 tasks per group
    - Task 1: classification
      - All groups will work on the same problem
    - Task 2: open problem based on the data set (http://sensor.nevada.edu/SENSORDataSearch/)
      - 2 problems. 3 groups per problem
  - You can apply whatever techniques you learnt from data mining course and other courses

Evaluation

- Final report (due Dec 12, 2016 in class) (**15%**)
- Class presentation and/or demo (**5%**)
  - Nov. 28, 2016 and Nov. 30, 2016 (tentative)

Task 1: Classification

- Provided data
  - The training set and its label information
  - The testing set
- Hidden data
  - The label information of the testing data
  - The data will be used for the purpose of evaluation

Data Format

- The training set
  - training.txt
  - The first column is the information ID
  - The second column is the feature ID
  - The third column is the value of the feature
  - The default values of features are zeros

```
1  16  1
1  23  4
1  27  1
1  29  8
1  30  2
1  33  3
1  42  1
1  54  1
1  72  1
1  81  1
```

Data Format

- The label information of the training set
  - label_training.txt
  - Each row represents a data point in the training set
  - 1 is true information while -1 is misinformation

```
1
-1
1
-1
1
1
1
-1
-1
```

Data Format

- The testing set
  - testing.txt
  - It has the same format as the training set

```
1 16 1
1 23 1
1 27 1
1 29 2
1 50 1
1 245 1
1 340 1
1 388 1
1 589 1
1 638 1
1 764 1
1 902 1
1 905 1
1 2774 1
1 8066 1
1 10762 2
```

Model Challenge from Model Selection

- There are so many classifiers
  - Which one is better?

- There may be parameters in classifiers
  - How to determine the optimal values?

Evaluation

- Each group needs to submit label_test_groupID.txt in the same format as the label_training.txt
  - For example, if you are in group 1, the file name will be label_test_group1.txt

- Classification accuracy will be used to evaluate the quality of the predicted labels

  - Comparing the hidden labels with your predicted labels

- Your final grades will strongly depend on the rankings of the quality of the predicted labels you provide

Task 2: Open Problem

Group 1

Matthew, Robert, Aaron

Irradiation forecast

- Problem Description: Solar energy farms undergo challenges in efficiency, production, and deployment to the sporadic nature of solar energy generation.
- Goal: Predict an irradiation forecast to provide solar farms information on deployment of new solar panels, and how to tune solar panels for optimal energy generation.
- Data Set:
  - Data Type: Radiation: Photosynthetically Active, Solar; Temperature, Barometric Pressure
  - Training Data: 01/01/2013 - 12/31/2013
  - Test Data: 01/01/2014 - 12/31/2014

Group 2

Amir, Biplav, Masood

Problem 1

- We take in the following parameters:
  - Barometric Pressure
  - Radiation: Solar
  - Relative Humidity
  - Temperature
  - Wind Direction
  - Wind Speed
- And design a model to predict **Combined precipitation**. The class for combination could be defined as a binary variable.

Problem 2

- We take in Solar Radiation and calculate the time cyclic correlation model and predict the **solar radiation** in some time in future by determining a model using neural networks.

Problem 3

- We take in the following parameters:
  - Photosynthetically active radiation
  - Solar Radiation
  - Soil Temperature
  - Soil Volumetric Water Content
  - Permittivity
  - Sap Flow Differential Temperature

- And determine a prediction model for **Trunk Radial Growth**.

Group 3

Christopher, Emily, Edward

Classification Problem

- We can use
  - Barometric Pressure
  - Relative Humidity
  - Temperature
  - Wind Speed
  - Radiation: Solar
- To see if there was precipitation that day, and what type of precipitation fell.
- Time Frame:
  - Learning Set: 2015
  - Test Set: 2016

Group 4

Vinh Le, Hannah Munoz, Daniel Goodnow
CS 491/691

Project Description & Goal

- The idea is to forecast weather conditions for one research site by using data from other research sites.
- The goal is to forecast across space rather than time
- For examples, by using the research sites:
  - Spring 0
  - Spring 1
  - Snake 1
  - Snake 2
- Forecast the following:
  - Spring 2
  - Spring 3
  - Spring 4

Data

The initial idea is to use:

- Data from the sites mentioned in the previous slide
- Approximately 2-3 years worth of Data from several research sites to train on
- To test, use current data to predict weather conditions
- Primarily testing on Windspeed
  - Potentially more

Group 5

Andy Singh, Zeeshan Sajid, Alex Ward

Overview

- Description: Do predictive analysis for snowfall and rainfall.
- Goal: To use data mining on the data-set to predict the current years weather forecast.
- Data:
  - Data set: Combined liquid and solid precipitation, barometric pressure, wind speed and temperature.
  - Training Data: From September- November, over the last three years, excluding 2015.
  - Test Data: From September-November, from last year.

Group 6

Yuan Sun Jiajun Xin Robert Martinez

A Data-Driven Approach to Predict the Success of Bank Telemarketing

- Description: Using data mining (DM) approach to predict the success of telemarketing calls for selling bank long-term deposits.
- Goal: Using data mining to improve the accuracy of targeting clients.
- Dataset:
  - Source:http://archive.ics.uci.edu/ml/datasets/Bank+Marketing
  - Data type: input variables include 20 attributes of a client and output variable is whether the client subscribed the term deposit.
  - Training Data/Test Data: ⅔ randomly selected sets as training data and the rest
  - as test data.

Task 2: Open Problem

- Group 1: Irradiation forecast
- Group 2: a) predict precipitation; b) predict the solar radiation; c) predict Trunk Radial Growth.
- Group 3: precipitation classification
- Group 4: forecast windspeed for one research site by using data from other research sites
- Group 5: Do predictive analysis for snowfall and rainfall.
- Group 6: Predict the Success of Bank Telemarketing

Task 2: Open Problem

- Problem: wind speed prediction
  - All Groups work on this problem

Problem: wind speed prediction

- Description: forecast wind speed for one research site by using data from other research sites.
- Goal: improve forecast accuracy of hourly wind speed at multi-sites

Problem: wind speed prediction

- Sites:
  - Snake Range West Salt Desert Shrub
  - Snake Range West Sagebrush
  - Snake Range East Salt Desert Shrub
  - Snake Range East Sagebrush
  - Snake Range West Pinyon-Juniper
  - Snake Range West Montane
  - Snake Range West Subalpine

- Training Data: **1-Hour aggregation of 10 minutes average** wind speed and wind direction
  - Starting from: January 1, 2013
  - Ending at: Dec. 31, 2015
- Test Data:
  - Sites:
    - Snake Range West Pinyon-Juniper
    - Snake Range West Montane
    - Snake Range West Subalpine
  - Starting from: Jan. 1, 2016
  - Ending at: June 30, 2016

Problem: wind speed prediction

- For each test site, you need to provide prediction accuracy of your proposed approach based on the following measure:
  - Mean absolute error (MAE)

$$MAE = \frac{1}{Number\ of\ points} \sum |forecast - actual|$$

  - Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{Number\ of\ points} \sum |forecast - actual|^2}$$

Evaluation

- Your final grades will strongly depend on the rankings of the prediction accuracy you provide

Report

- Team members and their contribution in %
- Introduction
- Literature review
- Your approach (for each task)
  - Preprocessing
  - Model selection
  - Parameter selection
  - Your solution
- List of document you submitted

Presentation

- Each group has 20 minutes to present their work and 3 minutes for questions.
- Your presentation will be evaluated by the other groups.