Last updated on Sunday, September 18, 2016

2016 IFN680 - Assignment Two (Machine Learning)

Assessment information

- Code and report submission at the end of **Week 13 (Monday of Week 14, 08.30am)**
- Use **Blackboard** to submit your work
- Group size: 2 people per submission

Overview

The aim of this assignment is to compare alternative techniques for predicting forest cover types from cartographic variables. You will be using the ***sklearn*** machine learning library to perform this study. The dataset is based on four wilderness areas in the Roosevelt National Forest, located in the Front Range of northern Colorado. Cover type data came from US Forest Service inventory information, while the cartographic variables used to predict cover type consisted of elevation, aspect, and other information derived from standard digital spatial data processed in a geographic information system (GIS).

Data Set Information:

The prediction of forest cover type is from cartographic variables only (no remotely sensed data). The actual forest cover type for a given observation (30 x 30 meter cell) was determined from US Forest Service (USFS) Resource Information System (RIS) data. Independent variables were derived from data originally obtained from US Geological Survey (USGS) and USFS data. Data is in raw form (not scaled) and contains binary (0 or 1) columns of data for qualitative independent variables (wilderness areas and soil types).

This study area includes four wilderness areas located in the Roosevelt National Forest of northern Colorado. These areas represent forests with minimal human-caused disturbances, so that existing forest cover types are more a result of ecological processes rather than forest management practices.

Some background information for these four wilderness areas: Neota (area 2) probably has the highest mean elevational value of the 4 wilderness areas. Rawah (area 1) and Comanche Peak (area 3) would have a lower mean elevational value, while Cache la Poudre (area 4) would have the lowest mean elevational value.

As for primary major tree species in these areas, Neota would have spruce/fir (type 1), while Rawah and Comanche Peak would probably have lodgepole pine (type 2) as their primary species, followed by spruce/fir and aspen (type 5). Cache la Poudre would tend to have Ponderosa pine (type 3), Douglas-fir (type 6), and cottonwood/willow (type 4).

The Rawah and Comanche Peak areas would tend to be more typical of the overall dataset than either the Neota or Cache la Poudre, due to their assortment of tree species and range of predictive variable values (elevation, etc.) Cache la Poudre would probably be more unique than the others, due to its relatively low elevation range and species composition.

Attribute Information:

Given is the attribute name, attribute type, the measurement unit and a brief description. The forest ***cover type*** is the **classification problem.** The order of this listing corresponds to the order of entries along the vectors of the database.

Name / Data Type / Measurement / Description

- Elevation / quantitative /meters / Elevation in meters
- Aspect / quantitative / azimuth / Aspect in degrees azimuth
- Slope / quantitative / degrees / Slope in degrees
- Horizontal_Distance_To_Hydrology / quantitative / meters / Horz Dist to nearest surface water features
- Vertical_Distance_To_Hydrology / quantitative / meters / Vert Dist to nearest surface water features
- Horizontal_Distance_To_Roadways / quantitative / meters / Horz Dist to nearest roadway
- Hillshade_9am / quantitative / 0 to 255 index / Hillshade index at 9am, summer solstice
- Hillshade_Noon / quantitative / 0 to 255 index / Hillshade index at noon, summer soltice
- Hillshade_3pm / quantitative / 0 to 255 index / Hillshade index at 3pm, summer solstice
- Horizontal_Distance_To_Fire_Points / quantitative / meters / Horz Dist to nearest wildfire ignition points
- Wilderness_Area (4 binary columns) / qualitative / 0 (absence) or 1 (presence) / Wilderness area designation
- Soil_Type (40 binary columns) / qualitative / 0 (absence) or 1 (presence) / Soil Type designation
- **Cover_Type (7 types) / integer / 1 to 7 / Forest Cover Type designation**

Your tasks

Preprocessing (5 marks)

Write a Python script (with the file named ***task_1.py***), that

- loads the dataset shuffles the data
- deals with any missing value (if any)
- performs normalization of the data
- creates numpy arrays *train_data, validation_data* and *test_data* which contain respectively 70%, 15% and 15% of the shuffled data.
- saves these preprocessed arrays

Whenever applicable use the validation set to select the architectural parameters needed to build the following classifiers;

Naive Bayes (5 marks)

Write a Python script (with the file named ***task_2.py***), that

- builds naïve Bayes classifiers
- evaluates the prediction errors
- reports these errors with a *confusing matrix*

Decision Trees (5 marks)

Write a Python script (with the file named ***task_3.py***), that

- builds decision tree classifiers
- evaluates the classifiers

Nearest Neighbours (5 marks)

Write a Python script (with the file named ***task_4.py***), that

- builds kNN classifiers
- evaluates the classifiers

Support Vector Machines (5 marks)

Write a Python script (with the file named ***task_5.py***), that

- builds SVM classifiers
- evaluates the classifiers

Deliverables

You should submit via Blackboard a zip file containing

- A report in pdf format limited to 10 pages of text (be concise!)
    - describing the design of your experiments
    - reporting on the performance of your classifiers
    - a statement of completeness (concise description of what you have implemented, what works, what doesn't).

- All your Python script files with proper commenting. Good code documentation is critical.

Draft Marking Scheme

- Code quality; readability, simplicity, structure, genericity, in line documentation, header comments.   (15 marks)
- Tasks as indicated
- Report; structure, clarity, flow, grammar, typos, tables, figures

Final Remarks

- Start early. You are strongly encouraged to ask questions during the practical sessions.
- Email questions to f.maire@qut.edu.au