Reconhecimento de Padrões
Inteligência Geoespacial

2018/2019

Project Assignment

Rain Prediction@Australia

1 Background

Prediction of the next day weather is important for the general population and companies, both for comfort and planning. For example, the general population are interested ti decide about the clothes to use in the next day. On the other side, for example, agricultural companies are interested in deciding about the next day activities. Precipitation for sure the weather parameter that causes more impact in daily activities.

Your job in this assignment is to predict if rain will occur or not in the next day at 44 Australian locations!

2 Dataset Description

Consider the dataset available at https://www.kaggle.com/jsphyg/weather-dataset-rattle-package. This dataset contains daily weather observations from 44 different Australian weather stations.For each station the following annotations and measures were recorded:

- **Date:** The date of observation
- **Location:** The common name of the location of the weather station
- **MinTemp:** The minimum temperature in degrees celsius
- **MaxTemp:** The maximum temperature in degrees celsius

- **Rainfall:** The amount of rainfall recorded for the day in mm
- **Evaporation:** The so-called Class A pan evaporation (mm) in the 24 hours to 9am
- **Sunshine:** The number of hours of bright sunshine in the day.
- **WindGustDir:** The direction of the strongest wind gust in the 24 hours to midnight
- **WindGustSpeed:** The speed (km/h) of the strongest wind gust in the 24 hours to midnight
- **WindDir9am:** Direction of the wind at 9am
- **WindDir3pm:** Direction of the wind at 3pm
- **WindSpeed9am:** Wind speed (km/hr) averaged over 10 minutes prior to 9am
- **WindSpeed3pm:** Wind speed (km/hr) averaged over 10 minutes prior to 3pm
- **Humidity9am:** Humidity (percent) at 9am
- **Humidity3pm:** Humidity (percent) at 3pm
- **Pressure9am:** Atmospheric pressure (hpa) reduced to mean sea level at 9am
- **Pressure3pm:** Atmospheric pressure (hpa) reduced to mean sea level at 3pm
- **Cloud9am:** Fraction of sky obscured by cloud at 9am. This is measured in "oktas", which are a unit of eigths. It records how many eigths of the sky are obscured by cloud. A 0 measure indicates completely clear sky whilst an 8 indicates that it is completely overcast.
- **Cloud3pm:** Fraction of sky obscured by cloud (in "oktas": eighths) at 3pm. See Cload9am for a description of the values
- **Temp9am:** Temperature (degrees C) at 9am
- **Temp3pm:** Temperature (degrees C) at 3pm
- **RainToday:** Boolean: 1 if precipitation (mm) in the 24 hours to 9am exceeds 1mm, otherwise 0
- **RISK_MM:** The amount of rain. A kind of measure of the "risk".
- **RainTomorrow:** The target variable. Did it rain tomorrow? Yes or No.

Note: You should exclude the variable Risk-MM. Because this feature contains information about the future, i.e., it contains information directly about the target variable. More information is available at https://www.kaggle.com/jsphyg/weather-dataset-rattle-package/discussion/78316.

Do not restrict yourself to this data! You may use other information as features, for example geographic information (for example obtained from http://www.geonames.org).

3 Objective

Your task is to develop classifiers for rain prediction. Consider two scenarios:

- **Scenario A (General Classifier)**: where a single classifier should be used to predict rain for all locations;
- **Scenario B (Specific Classifiers)**: where classifiers should be developed for the **six Australian states** (New South Wales, Victoria, Queensland, Western Australia, South Australia and Tasmania).

4 Practical Assignment

4.1 Missing values

It is suggested to not consider features that has more than 20% of missing values. It is also suggested to eliminate measurements (patterns) that has at least one missing value.

4.2 Feature Selection and Reduction

Some of the supplied features may be useless, redundant or highly correlated with others. In this phase, you should consider to use feature selection and dimensionality reduction techniques, and see how they affect the performance of the pattern recognition algorithms. Analyze the distribution of the values of your features and compute the correlation between them. Make sure you know your features! Do not forget to present your findings in the final report.

4.3 Experimental Analysis

You should be able to design experiences in order to run the pattern recognition algorithms in the given data and evaluate their results. Define the appropriate performance metrics and justify your choices!

Run the experiments multiple times and to be able to present average results and standard deviations (of the metrics used) you should split the training set in parts and use cross-validation. At the end you should be able to choose the best classifier and evaluate them in a testing set .
Do not forget that manually inspecting the predictions of your algorithms can give you precious insights of where they might be failing (and why), and what you can do to improve them (e.g. what makes the algorithm fail in this particular case? what special characteristic does it have that makes it so hard? how can I make the algorithm better deal with those cases?). Go back and forward to the Pre-processing, Feature reduction and Feature Selection phases until you are satisfied with the results. It is a good idea to keep track of evolution of the performance of your algorithm during this process. Try to show these trends in your final report, to be able to fundament all the issues involved (choosing parameters, model fit, etc.)

4.4 Pattern Recognition Methods

You can write your own code in your language of choice or use the functions and methods available in Matlab and in the Statistical Pattern Recognition STPRTool used in the classes (since you are already familiarized with it). The methods used in your work should be described as well as discussion of the parameters used. Try out different pattern recognition algorithms. You should try to understand how they perform differently in your data.

4.5 Results and Discussion

Present and discuss final results obtained in your Project assignment. This problem was already studied by other authors. Compare your results with the results from other sources.

4.6 Code & Graphical User Interface (GUI)

You should deliver your software code in MATLAB, or in any other programming language you used during the project.

For your project you should write code for a graphical user interface (GUI). The GUI should improve the interaction of the user with the code by providing options for data-loading, feature selection/dimensionality reduction, classification, post-processing, validation and visualization.

Remember to comment your code. Write also a help section to your code that tells the purpose of the function, usage, and explanation of parameters.

5 Documentation

Write documentation (in Portuguese or in English) about your project. The documentation should include a cover page where course name, project title, date, names and student numbers of the authors are mentioned.

Describe the methods used for classification in such detail that the reader would be able to implement the same kind of functions for feature extraction and classification just based on your documentation and some basic background in pattern recognition. Always justify your choices, even when their are based on intuition. Do not forget to verify your assumptions! Include classification results with the given data to your documentation. At the end of your documentation you should have a list of all references used.

5.1 Requirements

Practical assignment is meant to be done in groups of two persons. If someone wants to work alone, this is also possible. Larger groups are not allowed.

5.2 Project Submission & Deadlines

1. **Project First Milestone (Deadline: 26th April 2019!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Minimum Distance classifier, Fisher LDA for Scenario A.
   - Code + short report.
2. **Project Final Goal (Deadline: 22th May 2019!)**
   Deliverables:
   - Data Preprocessing (Scaling, Feature Reduction (PCA & LDA), Feature Selection, etc.);
   - Several classifiers;
   - Final Report
   - Matlab code + GUI.
3. **Presentation and Discussion (27th May 2019!)**

Acknowledgments

Credits to Kaggle for data supply and description. Observations were drawn from numerous weather stations. The daily observations are available from http://www.bom.gov.au/climate/data. Copyright Commonwealth of Australia 2010, Bureau of Meteorology.

Definitions adapted from http://www.bom.gov.au/climate/dwo/IDCJDW0000.shtml.