SOFTENG 755: Special Topics – Bayesian Machine Learning

Assignment 1 Description

University of Auckland
Semester 2, 2018

1 Assignment Deadlines

Assignment 1 (30%): Regression and Classification

28 August 11:59PM: deliverables due

2 Learning Outcomes

By completing this assignment, you should demonstrate your ability to:

- Understand and practically adopt the regression and classification machine learning models to conduct data analysis
- Learn how to perform various data pre-processing and basic feature selection and extraction for different machine learning models
- Learn to evaluate machine learning models with different performance measures
- Understand and practically tune the hyperparameters to select an appropriate model
- Learn to use the Python packages for machine learning, e.g., scikit-learn, numpy, etc.

3 Submission

You will submit a zip/tar file which includes:

- A report discussing the process and results of applying the machine learning models for the specified data analysis tasks. You need to briefly depict the machine learning models you use to a data set. Then, you are required to describe and justify the process of handling data pre-processing, feature selection/extraction, and parameter tuning. Data analysis results with tables or figures should be reported, together with your interpretation and critical thinking.
- The source code (in Python) that can be executed with ease. No GUI is required, but clear README or other documentation should be given. This means that we can repeat your data analysis process and results with little effort. Another reason is that we will use the reserved testing datasets to justify your trained model again to see whether the testing errors comply with your results derived from the data sets released to you. You could also directly submit a Github link if you have your code there. In this case, just put the link at a noticeable place of the report, e.g., a separate/dedicated section.

The report should have a title page which has on it the course number and name, and your name. A single pdf file should be submitted and should be named "A1_<*familyname*>_<*UPI*>.pdf". All submissions should be done via Canvas. The maximum number of the report pages is **8**, including all references and appendices. All papers must conform, at time of submission, to the IEEE formatting guidelines available here. All submissions must be in PDF format. All submissions must use the A4 page format. Make sure that you are using the correct IEEE style file: the title should be typeset in 24pt font and the body of the paper should be typeset in 10pt font. Latex users: please use \documentclass[a4paper,10pt,conference]{IEEEtran}.

4 Marking Criteria

- All the required data analysis tasks have been reasonably accomplished.
- The organisation, presentation and readability of the report
- Appropriate justification of which you have chosen and what you have done in the data analysis process, as well as critical thinking and understanding on the related aspects of the machine learning methods
- The quality of source code, especially the ease of using the code to perform prediction/testing on the reserved data.
- The prediction results in the testing stage (based on the reserved data sets).

5 Late Submissions

Late submissions will incur the following penalties:

- 15% penalty for 1 to 24 hours late,
- 30% penalty for 24 to 48 hours late, and
- 100% penalty for over 48 hours late (Canvas assignment automatically closes).

If you have a legitimate reason for submitting late, discuss this with the lecturer well in advance of the assignment due date.

6 Assignment Details

Task Overview: In Assignment 1, you are required to conduct several data analysis tasks on four different data sets with the corresponding machine learning models including regression and classification. You are recommended to use the Python packages such as Pandas, Numpy and Scikit-learn to implement the data analysis programs. To apply the machine learning models, you need to pre-process the original data sets by feature selection and extraction to cater for your machine learning models. These decision-makings should be justified in your report. Then, you need to split the data sets released to you into training and testing data sets, and use the training data sets to build your models and the testing data sets to test the learned models. Hyperparameters tuning using your preferred cross-validation method (e.g., K-folds) should be performed should be performed where applicable with appropriate justification. The performance of the trained models with respect to different feature selection or hyperparameter tuning should be reported and interpreted appropriately. To ensure the robustness of the performance indicators, the averaged results from multiple executions of model training should be used. Both quantitative and qualitative comparisons among different methods that can be applied to the same data analysis tasks are also expected. Data sets and the machine learning methods are detailed as follows, respectively.

Description of Data Sets: We have four data sets with different characteristics for different machine learning models. You might need to have different considerations for individual data sets in terms of their characteristics. We have done the basic data pre-processing to facilitate your further processing. Note that the data we reserve will be of the same format. That is, we will feed your program with the data we reserve directly and expect to obtain the performance output. Each data set is briefly described as follows. More details are documented together with the data sets. You can down load the data files in Canvas at /Files/Data/Data-assignment-1.zip.

1) World Cup Final 2018: This data set is collected by our TA from the website https://www.fifa.com/worldcup/matches/. Each data instance contains the basic information about a match held for World Cup Final 2018 in Russia, including locations, dates, time, statistics of the match, and the scores. One characteristic of this data sets is that it is relatively small and feature selection/extraction might be in high demand.
2) SH1 Traffic Volume: Traffic congestion results in significant monetary losses in countries around the world, and short-term prediction continues to be an open problem because the underlying, complex spatiotemporal dependencies are difficult to model. This data set contains the historical data of traffic volume measurements along 45 segments of SH1, one of the major motorways in Auckland (New Zealand), over a period of time. The original form of the data is 45 time series of the volume. To facilitate your data analysis, we have performed basic feature extraction by the sliding window technique. The window size is 10 (including time stamps *t*-9, …, *t*). So, each data instance is the concatenation of the traffic volume of all the road segments at time stamp *t*-9, …, *t*, plus the volume of a segment (Segment 23) we are interested at time stamp *t*+1. So, the number of attributes is 45x10+1=451. Relative high dimensionality is one characteristic and embedding with temporospatial information is another characteristic.
3) Occupancy Sensor Data: Occupancy detection is an important task in Smart Home applications, e.g., the Smart Home control system can use the occupancy information to switch on or off a series of appliances. This data set contains the light, temperature, temperature, humidity and CO2 measurements in an office room. Each data instance contains these information together with date and time information, and the label if the room is occupied or not. The measurements are taken every minute.
4) Landsat Satellite Data: This data set consists of the multi-spectral values of pixels in 3x3 neighbourhoods in a satellite image, and the label information associated with the central pixel in each neighbourhood. This data set was generated from Landsat Multi-Spectral Scanner image data. The data set is used by a researcher in the Department of Statistics and Data Modelling at University of Strathclyde (acknowledged).

Data Analysis Task: All the data files are in the form of CSV (Comma Separate Values) format. The usage of Python packages to analyse the above data sets with machine learning models is highly recommended for the efficiency of marking process. The choice of other implementation languages (e.g., Java) should be your last consideration, preferably. Regression and classification models are mainly involved. The tasks for each data set are specified below:

Data Sets	**Data Analysis Tasks**
World Cup Final 2018	Regression and Classification
SH1 Traffic Volume	Regression
Occupancy Sensor Data	Classification
Landsat Satellite Data	Classification

The machine models will be used for each data analysis task are listed in the following table:

Data Analysis Tasks	**Machine Learning Models**
Regression	Ordinary regression Ridge regression
Classification	Perceptron SVM Decision trees Nearest neighbour classifier Naive Bayes classifier

1) World Cup Final 2018: For regression tasks, the score attribute (the win/loss/draw status attribute is ignored) is the target, while for classification tasks, the win/loss/draw status attribute is the target (the score attribute is ignored). Other attributes can be used as features. Note that the statistics information is obtained after each match, but we are usually interested in predicting the match based the historical match statistics. So feature selection and extraction might be applied in order to improve your prediction accuracy.
2) SH1 Traffic Volume: The last attribute “segment 23(t+1)” is the target for regression. Other attributes are features. Note that although the prediction focuses on Segment 23, other segments might be also correlated to the target segment because of the temporospatial connection of segments. Data instance can be regarded as independent of each other.

3) Occupancy Sensor Data: The “Occupancy” attribute is the target attribute. This is a binary classification problem. Also note that the two classes are not well balanced.
4) Landsat Satellite Data: This is a multi-class classification problem. The last attribute is the target attribute and others are features.

Notes:

- Although we have reserved a part of data for testing purpose, you need to partition the data released to you into training data set and testing data set (not used in model training, 10% of the dataset). That is, you need to report your own prediction performance results based on the testing data set you derived.
- You need to choose more than one performance indicators (if possible) to evaluate the performance of the trained models. You also need to justify your selection of the evaluation measures.
- You can include more regression/classification models if you wish.
- The online documentation/tutorial for Python Scikit-learn package: http://scikit-learn.org/stable/supervised_learning.html#supervised-learning.