COMP8325: Applications of Machine Learning in Cyber Security

Assignment Part-I Description

Macquarie University
Session 1, 2022

1 Assignment Deadline

Assignment Part-1 (13%): using machine learning models for data analysis

- **Sunday 27 March 2022 17:00PM: deliverables due**
- Individual

2 Learning Outcomes

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

- Understand and practically adopt some basic machine learning models regression, classification, clustering and other anomaly detection methods to conduct anomaly analysis in data
- Learn how to perform various data pre-processing and basic feature selection and extraction methods for different machine learning models
- Learn to evaluate machine learning models with different performance metrics
- Understand and practically tune hyperparameters to select an appropriate machine learning model for a specific application dataset
- Learn to use the Python packages for machine learning, e.g., scikit-learn, numpy, pandas, matplotlib, etc.

3 Submission

You will submit a zip / tar file which includes:

- Four Jupyter Notebook files discussing the process and results of applying the machine learning models for the four 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.
- At the beginning of each Jupyter Notebook file, please write down your name and student ID. All submissions should be done via iLearn.

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 reserved data.

5 Late Submissions

Late submissions follow the polices specified in Unit Guideline.

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

6 Assignment Details

In Assignment Part-1, you are required to conduct several data analysis tasks on four different data sets with the corresponding machine learning models for regression, classification, clustering and anomaly detection. 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 may need to pre-process the original data sets by feature selection and extraction to cater for your machine learning models. The decision-makings in data pre-processing feature preparation should be justified in your report. Then, you need to split the data sets released to you into training and testing data sets. The training data sets will be used to build the machine learning models and the testing data sets will be used to test the trained models. Usually, you need to tune the model hyperparameters by using the K-fold cross-validation method with appropriate justification. The performance of the trained models with respect to different feature selection or hyperparameter tuning should be reported, visualized, and interpreted appropriately. To ensure the robustness of the performance indicators, the averaged results from multiple executions of model training should be used. After tuning hyperparameters, you should report the performance metrics from the execution of the trained model on training data (i.e., for training performance) and testing data (i.e., for testing performance). Both quantitative and qualitative comparisons among different methods that can be applied to the same data analysis tasks are highly expected. Data sets as well as the analysis tasks and methods are detailed as follows.

Data Analysis Task: We have four data sets of different characteristics for different machine learning models. You may need to have different considerations for individual data sets in terms of their characteristics. The tasks for each data set are specified below:

Data Sets	Data Analysis Tasks	Machine Learning Models
Power Consumption Time Series	Regression	Ordinary regression
Landsat Satellite Data	Classification	Nearest neighbour classifier Decision tree classifier
Power Consumption Time Series	Clustering	KMeans
Predictive Maintenance Data	Anomaly Detection	Mahalanobis distance Distortion (in clustering) KNN distance LoF iForest One-class SVM (optional)
Network Traffic Data		

Notes:

- You are required to choose more than one performance metrics (if possible) to evaluate the performance of a trained model. You also need to justify your selection of the evaluation measures.
- You can include more relevant machine learning models and data sets if you wish for comparison study.
- Refer to https://scikit-learn.org/stable/ for scikit-learn documentations

Task-1 Regression on Power Consumption Time Series (3 marks): The time series dataset consists of samples about the weather factors and the power consumption of City A in 2017 (for assignment purpose, some specific information is hidden). Every ten minutes, the weather information including Temperature, Humidity, and Wind Speed is recorded together with the power consumption. You are required to train a regression prediction model for the next day's power consumption prediction based on the historical records of the past week (7 days).
Notes:

1). The data set has not be pre-processed, and you might need to consider pre-processing issues like missing value imputation and standardization before training the model.

2). Please note that while the goal is to build the model for predicting daily power consumption, the data is sampled every ten minutes.

3). So, you should perform basic feature extraction for time series by the sliding window technique. This is a very important step for the success of the built model. Let W be the

window size. If the task is to predict the power consumption at time stamp (i.e., day) t, you could construct features with the historical data at time stamps from $t - W$ to $t - 1$. Data instances after the feature extraction can be regarded as independent of each other. Also, it is also noted that the data set is a multi-dimensional time series.

4). With obtained feature vectors, a linear regression model can be trained for prediction.

Task-2 Classification on Landsat Satellite Data (2.5 marks): The 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. To facilitate your data analysis process, we have done sampling and data pre-processing for you. All the features are named as "Feature #", and the last attribute is the label (each number corresponds to a class). This is a multi-class classification problem.

Notes:

1). You need to tune the hyperparameter K for the KNN classifier, and impurity heuristics (information gain or Gini) for the decision tree classifier. Also, the pruning methods could be considered to tune the decision trees.

Task-3 Clustering on Power Consumption Data (2.5 marks).

Notes:

1). This task requires to perform clustering on the Power Consumption data used in Task-1. But you need to have a different feature extraction manner. Specifically, the clustering is daily based, i.e., you need to extract features for each day, and perform clustering on these features. There are many possible options. A simple one could be the concatenation of the samples belonging to the same day. But you could use other options.

2). Note that clustering is unsupervised learning and the no label information is specified or used in the training stage. So, all the attributes including Temperature, Humidity, Wind Speed, and Power Consumption could be used as the feature of a day.

3). You need to tune the hyperparameter K for KMeans clustering. Or, you could run the hierarchical agglomerative clustering on a small sample set to select the value for K with the help of dendrogram visualization.

4). Try to interpretate the clusters. For example, different clusters might correspond to different months or seasons.

Task-4 Anomaly Detection on two data sets (5 marks):

Notes:

1). Predictive Maintenance Data: The data set tries to build an anomaly detection to predict and detect the machine failures. The dataset consists of 10000 data points stored as rows with 14 attributes in columns. You still need pre-processing and feature extraction for this data set. Particularly, the target is the attribute "Machine Failure" (0 for normal points and 1 for anomaly), and you could simple ignore the last five columns as they have been merged into the attribute "Machine Failure". There are a several columns that might not be suitable

to included. Note that the encoding of the target classes might affect the calculation of AUC scores.

2). Network Traffic Data: This is the data set used for a competition. The competition task was to build a network intrusion detector, a predictive model capable of distinguishing between “bad” connections, called intrusions or attacks, and “good” normal connections. This database contains a standard set of data to be audited, which includes a wide variety of intrusions simulated in a military network environment. To facilitate your data analysis process, we have done sampling and data pre-processing for you. All the features are named as “Feature #”, and the last attribute is the label. The label just has two values (1 for normal data and -1 for anomalies).

3). You are required to perform anomaly detection on the two data sets with multiple methods. One-class SVM is optional while we encourage you to complete it by learning it from the sklearn online document.