CS/ECE 528: Embedded Systems and Machine Learning
Fall 2023

Homework/Lab 5: Anomaly Detection

Assigned: 24 October 2023
Due: 31 October 2023

Instructions:

- Submit your solutions via Canvas.
- Submissions should include your jupyter notebooks in a zip file, with notebooks names q1.ipynb, q2.ipynb, etc. in a single folder. You can include comments in your notebooks to explain your design choices.
- **"Save and checkpoint" your notebook after running your notebook, so that cell outputs are preserved.** If you are using Colab, make sure to 'Save' your notebook after running it, before downloading it and submitting.

Q1. (50 points) In this question, you will perform unsupervised anomaly detection for IoT network intrusion detection. The dataset we will use can be found here: http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html You will need to download kddcup.data.gz which has the full data set (18M; 743M Uncompressed). Starting from the notebook ***iot-intrusion.ipynb***, fill out the missing code to create and use an Isolation Forest on the dataset, for anomaly detection. You should explore different hyperparameters for the classifier to achieve the best performance. Your score will depend on the highest AUC-ROC value achieved by your model on the test set. You can read up on AUC-ROC here: https://developers.google.com/machine-learning/crash-course/classification/roc-and-auc.

Q2. (50 points) In this question, you will perform supervised anomaly detection for IoT network intrusion detection. Unlike the unsupervised anomaly detection in Q1 which was only able to differentiate between normal and anomalous data, a supervised anomaly detector can also predict the class to which an anomaly belongs to. The dataset we will use is the same as in Q1. Starting from the notebook ***iot-intrusion-rf.ipynb***, fill out the missing code to create and use a Random Forest Classifier on the dataset, for anomaly detection. You should explore different hyperparameters for the classifier to achieve the best performance. Your score will depend on the highest recall value achieved by your model on the test set.

Q3. (75 points) Sometimes it is required to predict anomalies that occur in time series data. RNNs can be used to predict anomalies in such time series data. In this question, you will predict anomalies in a time series data of an IoT device utilization. When an anomalous observation occurs in such a time series, there will likely be a large deviation between the predicted and observed series, because an anomalous event is often difficult to model. An example of metric that can capture the extent of this deviation is a root mean square error (RMSE) plot between predictions and observations. A large enough RMSE can be used to signal when anomalies occur. The dataset for this question is in the file iot-util.csv. Starting from the notebook ***time-series.ipynb***, fill out the missing code to create an RNN model for predicting the time series, and detect anomalies in that series. Note that if your model is poor, you will end up with more false positives, as well as false negatives. Note also that the threshold specified for anomaly detection in the notebook has been empirically determined. Show your RMSE plot generated from the notebook as well as the times at which your model detects the anomalies.