**Anomaly Detection**

**Assignment 4**

**(10 points)**

**1. How is credit-card fraud different from insurance fraud from the perspective of anomaly detection? (2 marks)**

For the credit-card fraud, it has application fraud and transaction fraud. If the goal is to find abnormal transactions in a short period time, we can aggregate transactions over a period of time. We can aggregate of transactions, either min, max, mean or sum etc by time, by category code, by location, by transaction method etc for the goal of anomaly detection. The company may also have labeled data containing previous fraudulent transations. Also supervised methods are better able to distinguish between  anomalies and noise too.

But the insurance fraud user-specific profiles cannot be constructed because some user may rarely make a claim. Repeated claims by a single user is often an indicator. Feature extraction is highly domain specific.

**2. Suppose you wish to detect outbreaks of an infectious but uncommon disease in your state. Sketch how an anomaly detection system might be used for this. Your answer should address the data to be used (and how it might be collected), preprocessing, modeling and system deployment. (5 marks)**

To detect the outbreaks, we need to monitor surveillance data based on disease characteristic that starts to spike and abnormally behave and the records significantly change when compared to normal history record.Huge cases are reported than expected over a particular period of time.

The data can include person profile, time, location, number of cases, symptom etc. So we can investigate under time duration, location, and number of cases. The time or temporal detection is to find abnormal spike in time for case aggregation. Spatial detection is to look for area of abnormally record high cases. Other than diagnostic reports, we can use pre-diagnostic, pharmacy sales, emergency department visits etc as data to help monitor outbreaks. All the data need to be preprocessed for analysis. In the multidimensional Data streams, we need to detect the rare class outlier breaks. The identification of such outliers is similar to that in the static supervised scenario and needs to be done more efficiently. ny pattern reches over the upper limit control based on eash disease is detected as outbreak.

For the model, we can monitor outbreak pattern over time series using moving average, general linear model etc. of course we can model the relationships among event of interest and those being observed. We can use clustering for set mining. Outbreak is detected based on the outlier cluster pattern.

**3. Section 11 of Chapter 13 lists several resources for open source and commercial anomaly detection software. Study the documentation of any one of them or some other anomaly detection software and assess it from the standpoint of the guidelines listed in Section 10. (3 marks)**

Python Scikit Learn package has quite complete code base for the anomaly detection purpose. sklearn.preprocessing package provides several common utility functions and transformer classes to change raw feature vectors. Standardization, or mean removal and variance scaling can be done through scale function. Normalization can be done through function normalize. Feature binarization is the process of thresholding numerical features to get boolean values. scikitplot package can be used for visual analysis at all stages of outlier analysis. Of course we can always use matplotlib instead of scikitplot. sklearn.covariance covariance estimation can help concentrating on a relevant cluster. decision_function method is also defined from the scoring function that negative values are outliers. svm.OneClassSVM finds a new, but regular, observation outside the frontier. sklearn.neighbors has LocalOutlierFactor. sklearn.ensemble has IsolationForest. Overall Python Scikit learn package has quite complete code base for anomaly dectection.