CS6301: R For Data Scientists

LECTURE 30: ANOMALY DETECTION

The Problem …

We have a dataset and would like to identify observations which we can consider “anomalies”

What is an anomaly?

- An observation which is considerably “far away” from most observations
- In one dimension, we can use 1.5IQR rule …
- Tougher in higher dimensions

Notice, this is not a *prediction* (classification) problem

- We are not trying to classify future observations as anomalies
- This is more of a static analysis – given a set of datapoints, can we identify those which are “far outside the norm”

One Approach

One approach we can take to this problem would be to do a cluster analysis on the data, and then find data points that are on the edges of the clusters

- This will probably require K-means – we need a way to measure how “far out” an observation is, that is distance from the center of the cluster

We will first need to clean the data set and convert to numerical

We may also do dimension reduction (PCA) …

Finally, we can do K-Means clustering, use the elbow method to pick the optimal number of clusters

One Approach

Once we have the data points assigned to clusters, how do we find the “outliers”?

We have the cluster centroid coordinates; for each data point in the cluster, we can calculate the distance to the centroid.

We can then take the data points that are at the edges

- Compute a sample mean and standard deviation for the distances, pick points that are in the 1% percentile (i.e. the 1% that are furthest away)
- Can also do 1.5 IQR at this point, since the distance measure is one dimensional

The Algorithm

1. Clean the dataset
2. Use model.matrix() (or some other function) to convert all data to numeric
3. Do PCA to reduce dimensions
4. Do K-Means clustering, identify optimal number of clusters
5. Calculate distances from each data point to the centroid of its cluster
6. Use some measure to select anomalies based upon distance

Predicting Anomalies

Can we build a model that will predict an anomaly, based upon input variables?

This is a classification problem (techniques will be introduced later)

Would help if we had a dataset with each datapoint labelled (anomaly/not anomaly)

If we do not have the labels, we can create them using the above algorithm …

- Find the anomalies, and create a new column. Assign all datapoints a ‘0’ except for the ones identified as anomalies in the algorithm

AnomalyDetection Package in R

Developed by data scientists at Twitter

Works with *time series* data

Detects anomalies in a time series using an algorithm called Seasonal Hybrid ESD

We will study time series in another lecture, but for now understand that a time series is usually broken into parts: *trend, cyclical, seasonal,* and *noise.*

The package is relatively new (Sept 2017)

Links

https://medium.com/@xenonstack/anomaly-detection-of-time-series-data-using-machine-learning-deep-learning-c248061ea4f5