Lecture 8
Histogram and Regression

Ref: Outlier Analysis, Charu C Agrawal
Ref: Outlier Analysis: A Review, Chandola et al.

Histogram-based Anomaly Detection

- Build histogram of the normal data based on the feature value
- If a test data falls in one of the bins of the histogram, it is normal
- Alternatively, the inverse of the height of the bin in which data falls is taken as anomaly score
- Choosing bin width is challenging

Dealing with Multivariate Data

- Construct histogram for each attribute
- During testing phase, calculate attribute-wise anomaly score
- Overall score may be by combining individual scores

Histogram-based Outlier Score for Multivariate Analysis

$$HBOS(\mathbf{x}) = \sum_{d=1}^{D} log\left(\frac{1}{hist_d(\mathbf{x})}\right)$$

- D is the number of variables (or dimensions)
- $hist_d(\mathbf{x})$ is the height of normalised histogram in which dth variable of the given data point falls

Effect of the bin Width

- Smaller width bins lead to a lot of empty or small height bins, leading to more false positives
- A larger bin width will force many anomalous data into normal bins, leading to high false negative rate
- Rule of thumb is to create k bins where k is square root of N, the total number of data points

Freedman–Diaconis rule

$$\text{Bin width} = 2\frac{IQR(X)}{\sqrt[3]{n}}$$

IQR is inter-quartile range of the dataset, n is the number of data points.

IQR

- Arrange the data points in increasing order
- Q1 is the middle of the first half of the data
- Q3 is the middle of the second half of the data
- IQR = Q3-Q1
- For even number of data points, the middle value if the average of the two middle values

Dynamic Bin Width

- Sort the values
- Group N/k values in one single bin
- Hence dense areas will have more bins
- Bin height is calculated as k/w, where w is bin width (it assumes that the area should be equal to the number of instances, i.e., k)

Learning Histogram

- Labelled normal data is needed to learn normal histogram
- If available, anomaly data histogram can also be learned
- Alternatively, bin height can be analysed to tag a bin is anomaly or normal, e.g., larger height bins are normal

Detecting anomaly in dependency oriented dataset, aka contextual dataset!

Simple Linear Regression

$$y = mx + c$$

or

$$y = \beta_0 + \beta_1 x$$

Dependent variable

Explanatory variable

Training Regression Model

- If there are n attributes, n equations are enough to learn the coefficients
- But the number of equations is more, hence it is an over-determined system
- No exact solution is possible

The data points do not fall on a perfect straight line!

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Least Squares Method

Minimize E= $\sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (\hat{y}_i - \beta_0 - \beta_1 x_i)^2$

Setting $\frac{\partial E}{\partial \beta_0} = 0$ and $\frac{\partial E}{\partial \beta_1} = 0$, we get

$$\beta_0 = \bar{y} - \beta_1 \bar{x} \qquad \beta_1 = \frac{\sum_{i=1}^{n} (x_i y_i - \bar{y} x_i)}{\sum_{i=1}^{n} (x_i^2 - \bar{x} x_i)}$$

Regression-based Anomaly Score

- Different dimensions are related to each other with a linear equation
- Fit a regression model to the training data
- The residual score of each training data is the anomaly score (you can use z-value of residual score)

What if the dependent variable is not known?

- Take each variable as dependent variable and obtain d models
- Calculate d outlier scores
- Average score can be used for outlier analysis