Anomaly Detection: Statistical Methods

Statistical Methods

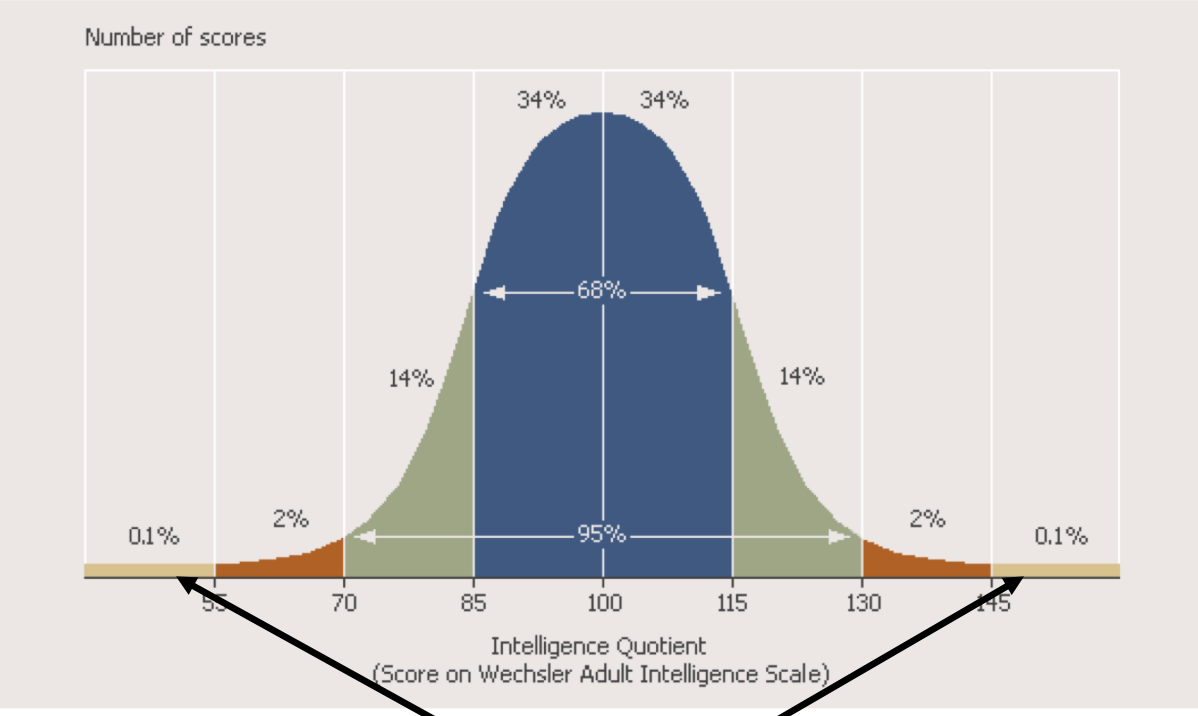
- Statistical methods assume that the data is generated by an underlying statistical model i.e., generative model.

- Outline of a statistical outlier detection method:
  - Fit a generative model to the given data
  - Data in regions of low probability are potential outliers

- How to fit a model to data?
  - **Parametric Approach**: Assume a parametric distribution e.g., Gaussian distribution and learn the parameters using the given dataset
  - **Non-parametric Approach**: No a-priori assumed model, learn the model from data e.g., a histogram

Parametric Methods: Univariate Normal Distribution

- Suitable for univariate data i.e., having only one feature
- Fit a Normal distribution to the data:
  - Use the data to determine the mean and variance
  - Data with low probability are labeled as outliers
  - Usually done by calculating ***z-scores*** and then thresholding



Anomalies in IQ

$\mu \pm 3\sigma$ region contains 99.7% data

Parametric Methods: Grubb's Test

- Used for univariate outlier detection, with the assumption that the data comes from a Normal distribution

- Description:
  - For each data instance compute its z-score
  - An instance will be an outlier if:

$$z \geq \frac{N-1}{\sqrt{N}} \sqrt{\frac{t^2_{\alpha/(2N),N-2}}{N-2+t^2_{\alpha/(2N),N-2}}}$$

where, N is the number of data instances, $t^2_{\alpha/(2N),N-2}$ is the value taken by a t-distribution at a significance level of $\alpha/(2N)$

Parametric Methods: Multivariate Data

- Most real-world datasets have multiple features, how can we use univariate methods to detect anomalies in multivariate data?

- We can apply the univariate method to each individual feature and then combine the results

- For an instance $x_i \in \mathbb{R}^m$, calculate the z-score for each feature independently:

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \qquad 1 \leq j \leq m$$

Parametric Methods: Multivariate Data

- The aggregate deviation from the norm can then be calculated as the sum of squares of these deviations (i.e., the z-scores)

- Since $z_{ij} \sim N(0,1)$, then the sum of squares is a random variable $V_i = \sum_{j=1}^{m} z_{ij}^2$

- For data with ***m*** features V_i is distributed as a chi-square distribution with ***m*** degrees of freedom

$$V_i \sim \chi^2(m)$$

- Calculate the probability of V_i to decide whether x_i is anomalous or not

Multivariate Gaussian for Anomaly Detection

- **Assumption**: data is normally distributed as a multivariate Gaussian distribution

- Recall that,

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

- Mahalanobis Distance:
  - The term in the exponent is half the squared Mahalanobis distance

$$Mahalanobis(x, \mu, \Sigma) = \sqrt{(x-\mu)\Sigma^{-1}(x-\mu)^T}$$

  - Similar to Euclidean distance from the mean, but scales the distance based on inter-feature correlations

Multivariate Gaussian for Anomaly Detection

- Instances that have a higher Mahalanobis distance, are candidate outliers

- Mahalanobis distance is robust to increasing dimensionality, because it takes into account the relative relevance of different directions through the covariance matrix

- However, calculating the Mahalanobis distance requires inverting the covariance matrix (Σ), and in cases where Σ has low rank regularization is required

Parametric Methods: Multivariate Data

- For high dimensional data we can model the data distribution using Gaussian Mixture Models (GMM)

- GMMs attempts to find a mixture of multivariate Gaussian probability distributions that best model any input dataset

- Fit a GMM to data (EM algorithm):
  - Data that resides in low probability regions are outliers
  - GMMs can also be viewed as clustering the data, therefore data that does not belong to any cluster (has low probability) is an outlier

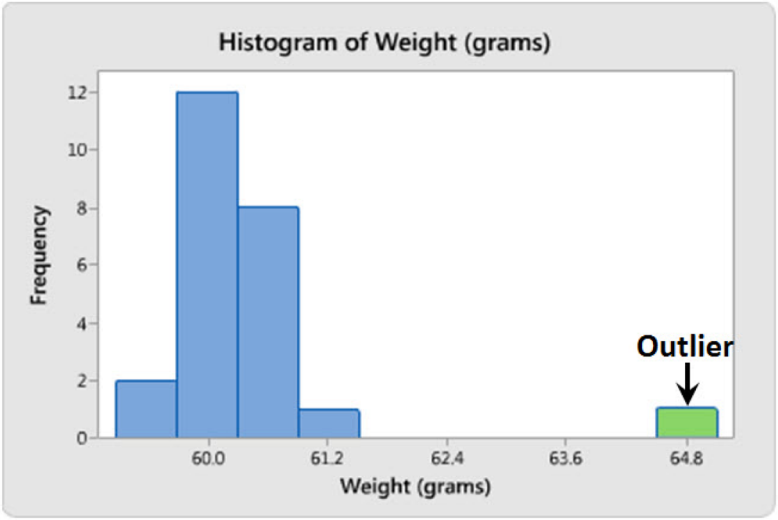
Parametric Methods: Drawbacks

- Parametric methods impose a specific distribution on the data, when this assumption does not hold the detected anomalies are not credible
- If the model is too general i.e., number of model parameters is large it can overfit the data, resulting in missing outliers (or group of outliers)
- Example: consider a mixture model with a large number of components: more than what is needed to describe the data. In this case a tight group of outliers might be completely described by one of the components, leading us to believe that the group of outliers is normal data

Non-parametric Methods: Univariate Data

- Instead of assuming an underlying statistical model, the model of normal data is learned from the data
  - Fewer assumptions about the data lead to wider applicability

- Outlier detection using histogram:
  - Figure shows the weights of driver shafts (golf)
  - We can see that 64.8g is an outlier
  - What about 61.2g?



- Alternative:
  - Estimate data density using kernel density estimation
  - Evaluate the resulting density function to label instances with low scores as outliers

Non-parametric Methods: Multivariate Data

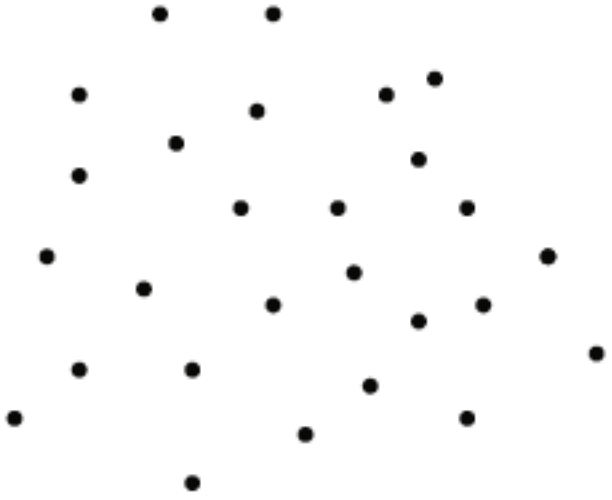
- Histograms are easy to interpret when we have univariate data, however even with two features (bivariate), histograms lose their ease of interpretability

- Depth-Based anomaly detection:
  - Use convex hull analysis to find outliers
  - Intuition: data points on the outer boundaries of the data that form the convex hull are more likely to be anomalous
  - An iterative procedure that progressively removes instances that form the convex hull

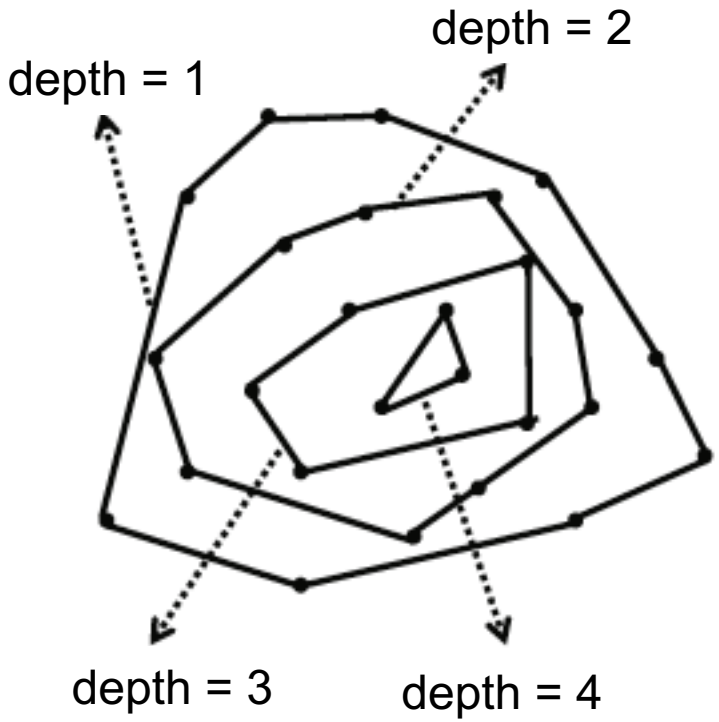
Depth-Based Anomaly Detection

Algorithm: DepthBasedAnomalyDetection($\mathcal{D}, r$)
begin:
    $k = 1$
    repeat
        Find S: set of corners of convex hull
        Depth(S) = k
        $\mathcal{D} = \mathcal{D} - S$
        $k = k + 1$
    until $\mathcal{D}$ is empty
    Outliers = all points with depth r
end

Depth-Based Anomaly Detection



Data



Successive Convex Hulls

Non-parametric Methods

- Drawbacks of histogram based anomaly detection:
  - Highly dependent on bin size and number of bins
  - Small bins cause normal data to appear sparse leading to false positives
  - Large bins can lump the outliers with frequent data leading to false negatives

- Non-Parametric Density Estimation:
  - For non-parametric methods estimating the probability distribution in high dimensional spaces is difficult

Non-parametric Methods

- Depth-Based Anomaly Detection:
  - Computational complexity increases exponentially with increase in dimensionality
  - In higher dimensions, most data will lie at the corners of a convex hull because the number of points that form the corners can be exponentially related to dimensionality
  - Only effective for finding boundary outliers, cannot find internal outliers

References

Based on slides by Carla E. Brodley

J. Han, M. Kamber and J. Pei, Data Mining: Concepts and Techniques. 2011

C. C. Aggarwal, Outlier Analysis. 2012