OUTLIER DETECTION

MYRON OUYANG

u16008368

THE PROBLEM

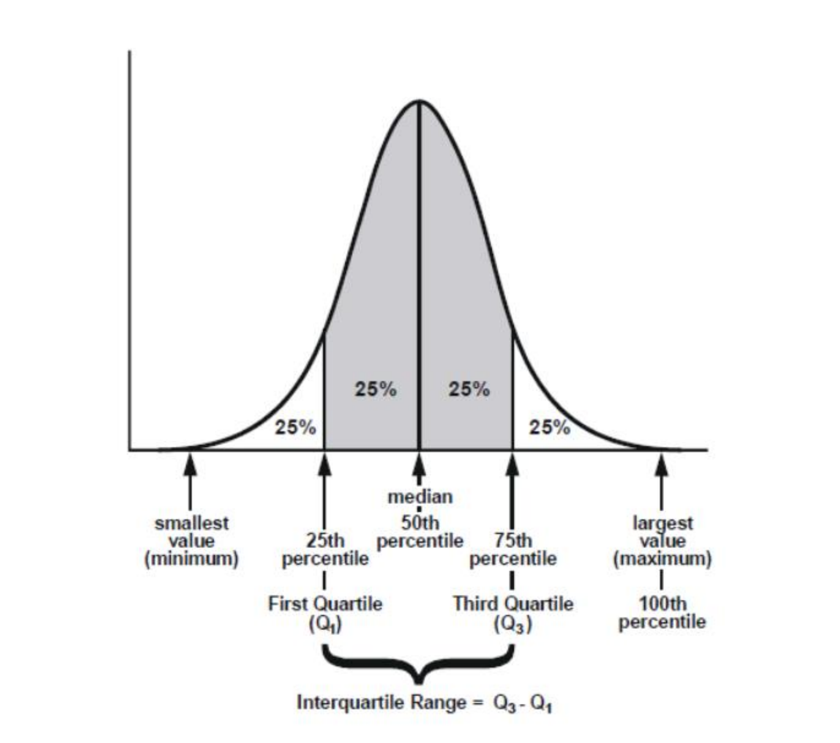
What is the problem?	• Existence of data points that deviate from the rest of the data
Why should this problem be solved?	• Can cause bias in results of data analysis • Affects the mean value of data
Causes of problem	• Errors • Genuine extreme values

PROBLEM SOLUTION

1. Detect Outliers in data
   - Exteme Value Analysis
   - Distance measures
   - Angle-based Outlier Degree
   - SmartSifter
   - AutoEncoder
2. Treat Outliers in data

EXTREME VALUE ANALYSIS

- Basic outlier detection method
- Find statistical tails in distribution of data
- Data points at extreme ends of tails are the outliers
- Use Gaussian distribution or
- Calculate inter-quantile range of data



DISTANCE MEASURES

Mahalanobis distance (MD)

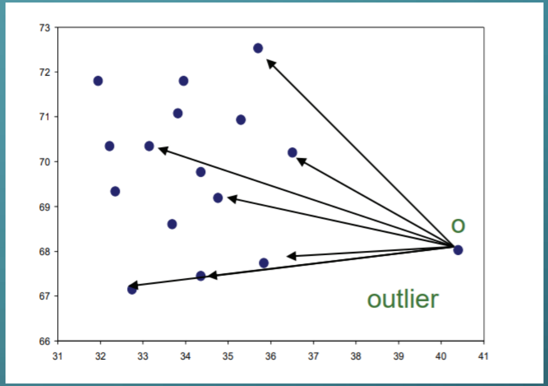
- "Distance between two points in multivariate space"
- Measures distance from central point
  - Central point – total mean of **ALL** the multivariate data
- The further away a data point is from the central point, the greater the MD

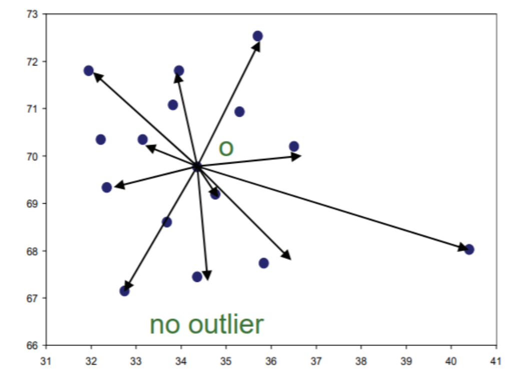
Cook's Distance (CD)

- Measures how much the expected outcome will change provided the current data point is dropped
- Common threshold value is 4 times the mean
- Therefore any change in expected outcome greater than the threshold can be considered influential

ANGLE-BASED OUTLIER DETECTION

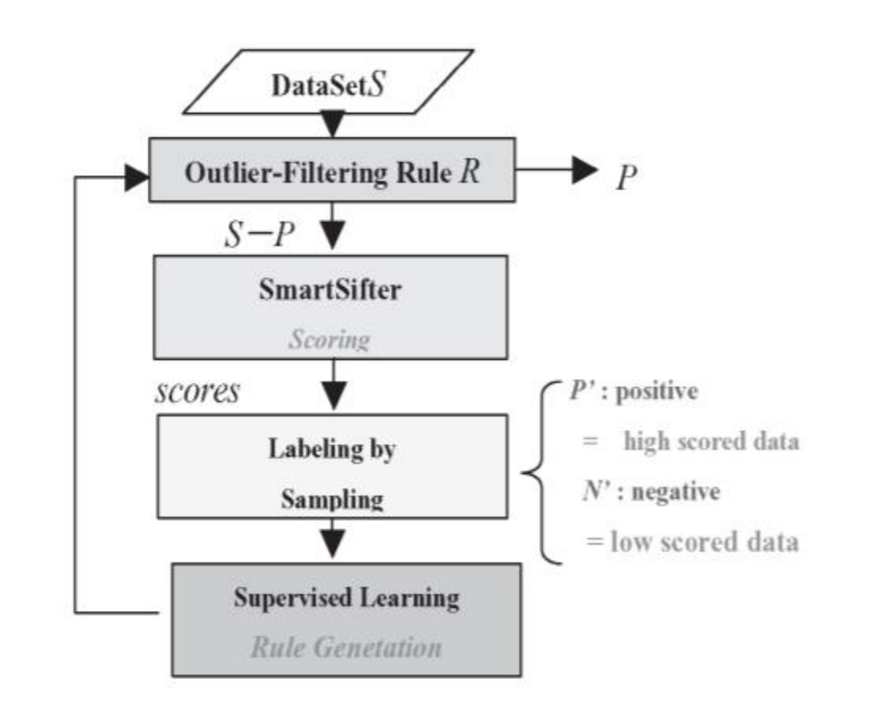
- Angles are more stable than distances when working with high-dimensional data
- Based on the variance of the angles between a point and all the other points in data set
- Outlier if majority of other data points are in similar direction
- Not outlier if other data points are in varying directions





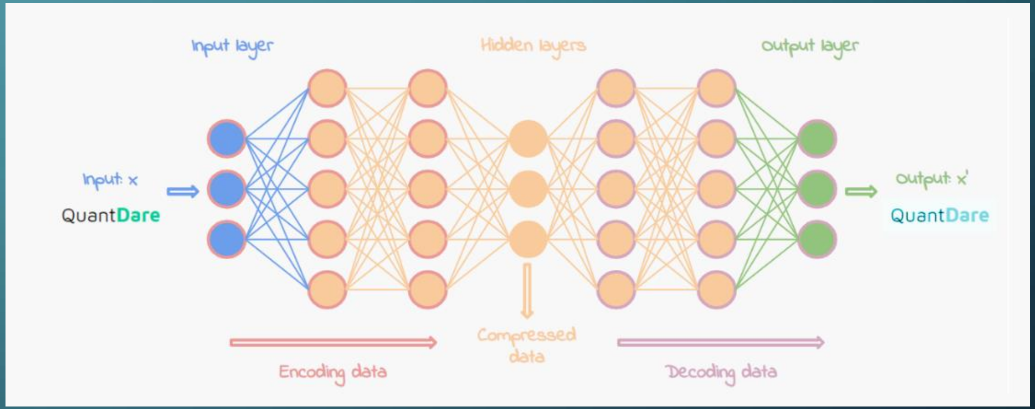
SMARTSIFTER

- Construct a model in which data is filtered through
- The model is probabilistic and it is affected by each data point passing through it
- Filter each data points successively through the model
- Data point is an outlier if the model changes significantly



AUTOENCODER

- Unsupervised artificial neural network
- Two stages: Encode & Decode
- Encode
  - Compresses data, removes noise/ unnecessary information
- Decode
  - Reconstructs data from compressed data
- Compressing data forces NN to only learn the important features (outliers are revealed when the main features are filtered out)



REFERENCES

- Extreme value analysis (EVA) of inspection data and its uncertainties - ScienceDirect [WWW Document], n.d. URL https://www.sciencedirect.com/science/article/pii/S0963869517300488 (accessed 10.27.19).
- Mahalanobis Distance: Simple Definition, Examples - Statistics How To [WWW Document], n.d. URL https://www.statisticshowto.datasciencecentral.com/mahalanobis-distance/ (accessed 10.27.19).
- Cook's Distance / Cook's D: Definition, Interpretation - Statistics How To [WWW Document], n.d. URL https://www.statisticshowto.datasciencecentral.com/cooks-distance/ (accessed 10.27.19).
- Yamanishi, K., Takeuchi, J.I., Williams, G. and Milne, P., 2004. On-line unsupervised outlier detection using finite mixtures with discounting learning algorithms. *Data Mining and Knowledge Discovery*, *8*(3), pp.275-300.
- Chen, J., Sathe, S., Aggarwal, C. and Turaga, D., 2017, June. Outlier detection with autoencoder ensembles. In *Proceedings of the 2017 SIAM International Conference on Data Mining* (pp. 90-98). Society for Industrial and Applied Mathematics.
- Outliers detection with autoencoder, a neural network | Quantdare [WWW Document], n.d. URL https://quantdare.com/outliers-detection-with-autoencoder-neural-network/ (accessed 10.27.19).
- Pham, N., 2018, September. L1-Depth Revisited: A Robust Angle-Based Outlier Factor in High-Dimensional Space. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 105-121). Springer, Cham.