Outliers and Anomalies

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Outliers

Outlier: anything too extreme in value, that is oftentimes better removed from the sample

- In a normal distribution, outliers are considered to be all the points that are beyond 2 standards deviations from the mean (sometimes 3)

The exact threshold beyond which something is an outlier varies and relates to a p-value (e.g. 0.05)

- This is a heuristical approach, so any outliers you identify this way are better off being examined further, before removing

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

ANOMALY DETECTION

Involves identifying data points that are very different to the bulk of the dataset (aka anomalies)

- Anomalies can be outlier or inliers

Anomalies tend to be quite different from each other

- Anomaly detection methods rely on figuring out "what's normal"

Anything deviating from the "normal" data points is considered an anomaly

- What's normal can be a combination of different things (i.e. different clusters of data points)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Importance of anomaly detection

Essential for figuring out:

- Fraud in monetary transactions
- Fraud in web traffic
- Network hacking (intrusion detection) and other cyber security issues
- Potential terrorists
- Spam emails
- Diseases based on diagnostics

Useful for cleaning up data

A great research topic as it pushes the envelope of what's possible through data science

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Examples of anomalies in a common data science application

NLP = Natural Language Processing (analyzing and processing text written in plain English or some other language the system is trained on)

Common anomalies in NLP data:

- Very rare words or phrases (appearing only once or twice in the whole corpus)
- Very common words or phrases (aka stopwords)
- Irrelevant words, appearing in normal frequencies (e.g. "1990s", common abbreviations, etc.)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Outliers and inliers as anomalies

Anomalies can be seen as points with very low density

Outliers

- Points having very high values
- Points having very low values
- Easy to identify using their p-values

Inliers

- Points within the main body of the distribution
- Their neighbors tend to be far, in general
- Harder to identify (their p-values appear normal)

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Methods for identifying anomalies

Single dimensional data

- P-value
- Other statistical methods

Multiple dimensional data

- One-class SVMs
- Parametric statical methods (e.g. multivariate Gaussian distribution)
- Other methods (e.g. kNN, ANNs, Rule-based systems, etc.)

One-Class SVMs

- Trained on "normal" data
- Classify a given data point as whether it belongs to that class or not
- Need to define a cut-off threshold (i.e. below which probability score a data point is considered an "outsider" of the normal class)
- Very effective for highly complex datasets

Statistical methods

For 1 dimensional data:

- Ensure distribution is normal
- identify data points with p-value <= th (e.g. 0.01)

For n dimensional data

- Ensure that each feature's distribution is normal
- Calculate n-dimensional p-value for every data point using the formula $p(x) = \frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$ where μ = mean, n = number of features, Σ = covariance matrix

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Some considerations

- Often it is a good idea to perform dimensionality reduction before anomaly detection to reduce computational cost.
  - If number of original features > number of data points, multivariate anomaly detection won't work
- Not every anomaly is a data point that should be removed.
  - Sometimes anomalous data is useful, especially if it's large enough to constitute clusters beyond the ones of normal data
- It is recommended you try different anomaly detection methods before labeling a data point as an anomaly
- For more information on multivariate anomaly detection, check out http://bit.ly/2lgTpFl

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Python functions and classes

Stats-based Anomaly Detection:

- Single dimensional data: descriptive stats functions from *numpy* package (e.g. *mean*, *std*, etc.) as well as from the *scipy.stats.norm* class, various functions (e.g. *sf*, *ppf*, etc.)
- Multiple dimensional data: same as for single dimensional data, but with aggregation of p-values, using a product

One-class SVMs: *OneClassSVM* function in *sklearn.svm* class

PROFESSIONAL & CONTINUING EDUCATION
UNIVERSITY of WASHINGTON

Summary

- Anomalies are points with low density
  - Outliers usually more than 2 standard deviations
  - Inliers usually far from their neighbors
- Look for normal distributions
  - 1-dimensional data look at p-value
  - N-dimensional data check at all the features
- Perform dimensionality reduction before anomaly detection