Statistics 202: Data Mining

Outliers
Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

December 2, 2012
Outliers

Concepts

- What is an outlier? *The set of data points that are considerably different than the remainder of the data...*
- When do they appear in data mining tasks?
  - Given a data matrix $X$, find all the cases $x_i \in X$ with anomaly/outlier scores greater than some threshold $t$. Or, the top $n$ outlier scores.
  - Given a data matrix $X$, containing mostly normal (but unlabeled) data points, and a test case $x_{\text{new}}$, compute an anomaly/outlier score of $x_{\text{new}}$ with respect to $X$.

Applications

- Credit card fraud detection;
- Network intrusion detection;
- Misspecification of a model.
What is an outlier?
Outliers

**Issues**

- How many outliers are there in the data?
- Method is unsupervised, similar to clustering or finding clusters with only 1 point in them.
- Usual assumption: *There are considerably more “normal” observations than “abnormal” observations (outliers/anomalies) in the data.*
Outliers

General steps

- Build a profile of the “normal” behavior. The profile generally consists of summary statistics of this “normal” population.
- Use these summary statistics to detect anomalies, i.e. points whose characteristics are very far from the normal profile.
- General types of schemes involve a statistical model of “normal”, and “far” is measured in terms of likelihood.
- Other schemes based on distances can be quasi-motivated by such statistical techniques . . .
Outliers

Statistical approach

- Assume a parametric model describing the distribution of the data (e.g., normal distribution)
- Apply a statistical test that depends on:
  - Data distribution (e.g., normal)
  - Parameter of distribution (e.g., mean, variance)
  - Number of expected outliers (confidence limit, $\alpha$ or Type I error)
Outliers

Grubbs’ Test

- Suppose we have a sample of \( n \) numbers \( Z = \{Z_1, \ldots, Z_n\} \), i.e. a \( n \times 1 \) data matrix.
- Assuming data is from normal distribution, Grubbs’ tests uses distribution of

\[
\frac{\max_{1 \leq i \leq n} Z_i - \overline{Z}}{SD(Z)}
\]

to search for outlying large values.
Outliers

Grubbs’ Test

- Lower tail variant:
  \[
  \min_{1 \leq i \leq n} \frac{Z_i - \bar{Z}}{SD(Z)}
  \]

- Two-sided variant:
  \[
  \max_{1 \leq i \leq n} \frac{|Z_i - \bar{Z}|}{SD(Z)}
  \]
Outliers

Grubbs’ Test

- Having chosen a test-statistic, we must determine a threshold that sets our “threshold” rule.
- Often this is set via a hypothesis test to control Type I error.
- For large positive outlier, threshold is based on choosing some acceptable Type I error $\alpha$ and finding $c_\alpha$ so that

$$P_0 \left( \frac{\max_{1 \leq i \leq n} |Z_i - \bar{Z}|}{\text{SD}(Z)} \geq c_\alpha \right) \approx \alpha$$

- Above, $P_0$ denotes the distribution of $Z$ under the assumption there are no outliers.
- If $Z$ are IID $N(\mu, \sigma^2)$ it is generally possible to compute a decent approximation of this probability using Bonferonni.
Outliers

**Grubbs’ Test**

- Two sided critical level has the form

\[ c_\alpha = \frac{n-1}{\sqrt{n}} \sqrt{\frac{t^2_{\alpha/(2n),n-2}}{n-2 + t^2_{\alpha/(2n),n-2}}} \]

where

\[ P(T_k \geq t_{\gamma,k}) = \gamma \]

is the upper tail quantile of \( T_k \).

- In \( R \), you can use the functions \( \text{pnorm}, \text{qnorm}, \text{pt}, \text{qt} \) for these quantities.
Model based: linear regression with outliers

**Figure**: Residuals from model can be fed into Grubbs’ test or Bonferroni (variant)
Outliers

**Multivariate data**

- If the non-outlying data is assumed to be multivariate Gaussian, what is the analogy of Grubbs’ statistic

  \[
  \max_{1 \leq i \leq n} \frac{|Z_i - \bar{Z}|}{\text{SD}(Z)}
  \]

- Answer: use Mahalanobis distance

  \[
  \max_{1 \leq i \leq n} (Z_i - \bar{Z})^T \Sigma^{-1}(Z_i - \bar{Z})
  \]

- Above, each individual statistic has what looks like a Hotelling’s \( T^2 \) distribution.
Outliers

**Likelihood approach**

- Assume data is a mixture
  
  \[ F = (1 - \lambda)M + \lambda A. \]

- Above, \( M \) is the distribution of “most of the data.”
- The distribution \( A \) is an “outlier” distribution, could be uniform on a bounding box for the data.
- This is a mixture model. If \( M \) is parametric, then the EM algorithm fits naturally here.
- Any points assigned to \( A \) are “outliers.”
Outliers

Likelihood approach

- Do we estimate $\lambda$ or fix it?
- The book starts describing an algorithm that tries to maximize the equivalent *classification likelihood*

$$L(\theta_M, \theta_A; l) = \left( (1 - \lambda)^{\#l_M} \prod_{i \in l_M} f_M(x_i, \theta_M) \right)$$

$$\times \left( \lambda^{\#l_A} \prod_{i \in l_A} f_A(x_i; \theta_A) \right)$$
Outliers

Likelihood approach: Algorithm

- Algorithm tries to maximize this by forming iterative estimates \((M_t, A_t)\) of “normal” and “outlying” data points.

1. At each stage, tries to place individual points of \(M_t\) to \(A_t\).
2. Find \((\hat{\theta}_M, \hat{\theta}_A)\) based on partition new partition (if necessary).
3. If increase in likelihood is large enough, call these new set \((M_{t+1}, A_{t+1})\).
4. Repeat until no further changes.
Outliers

Nearest neighbour approach

- Many ways to define outliers.
- Example: data points for which there are fewer than $k$ neighboring points within a distance $\epsilon$.
- Example: the $n$ points whose distance to $k$-th nearest neighbour is largest.
- The $n$ points whose average distance to the first $k$ nearest neighbours is largest.
- Each of these methods all depend on choice of some parameters: $k$, $n$, $\epsilon$. Difficult to choose these in a systematic way.
Outliers

Density approach

- For each point, $x_i$ compute a density estimate $f_{x_i,k}$ using its $k$ nearest neighbours.
- Density estimate used is

$$f_{x_i,k} = \left( \frac{\sum_{y \in N(x_i,k)} d(x_i,y)}{\#N(x_i,k)} \right)^{-1}$$

- Define

$$LOF(x_i) = \frac{f_{x_i,k}}{\left( \sum_{y \in N(x_i,k)} f_{y,k} \right)/\#N(x_i,k)}$$
Outliers

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample \( p \) as the average of the ratios of the density of sample \( p \) and the density of its nearest neighbors
- Outliers are points with largest LOF value

In the NN approach, \( p_2 \) is not considered as outlier, while LOF approach finds both \( p_1 \) and \( p_2 \) as outliers

**Figure**: Nearest neighbour vs. density based
Outliers

Detection rate

- Set $P(O)$ to be the proportion of outliers or anomalies.
- Set $P(D|O)$ to be the probability of declaring an outlier if it truly is an outlier. This is the detection rate.
- Set $P(D|O^c)$ to the probability of declaring an outlier if it is truly not an outlier.
Outliers

Bayesian detection rate

- Bayesian detection rate is

\[ P(O|D) = \frac{P(D|O)P(O)}{P(D|O)P(O) + P(D|O^c)P(O^c)}. \]

- The false alarm rate or false discovery rate is

\[ P(O^c|D) = \frac{P(D|O^c)P(O^c)}{P(D|O^c)P(O^c) + P(D|O)P(O)}. \]