CPSC 340:
Machine Learning and Data Mining

Outlier Detection
Fall 2016
Admin

• Assignment 1 solutions will be posted after class.
• Assignment 2 is out:
  – Due next Friday, but start early!
• Calculus and linear algebra terms to review for next week:
  – Vector addition and multiplication: $ax + \beta y$.
  – Inner-product: $x^T y$.
  – Matrix multiplication: $Xw$.
  – Solving linear systems: $Ax = b$.
  – Matrix inverse: $X^{-1}$.
  – Norms: $\|x\|$.
  – Gradient: $\nabla f(x)$.
  – Stationary points: $\nabla f(x) = 0$.
  – Convex functions: $f''(x) \geq 0$. 
Last Time: Hierarchical Clustering

- We discussed **hierarchical clustering**:  
  - Perform clustering at multiple scales.
  - Output is usually a **tree diagram** ("dendrogram").
  - Reveals much more structure in data.
  - Usually non-parametric:
    - At finest scale, every point is its own clusters.

- Important application is **phylogenetics**.
  - Scientific American yesterday:
    - “Scientists Trace Society’s Myths to Primordial Origins”
    - “Cosmic Hunt”: Man hunts animal that becomes constellation.

http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html
Motivating Example: Finding Holes in Ozone Layer

• The huge Antarctic ozone hole was “discovered” in 1985.

• It had been in satellite data since 1976:
  – But it was flagged and filtered out by quality-control algorithm.

https://en.wikipedia.org/wiki/Ozone_depletion
Outlier Detection

• **Outlier detection:**
  - Find observations that are “unusually different” from the others.
  - Also known as “anomaly detection”.
  - May want to remove outliers, or be interested in the outliers themselves.

• Some sources of outliers:
  - Measurement errors.
  - Data entry errors.
  - Contamination of data from different sources.
  - Rare events.
Applications of Outlier Detection

• Data cleaning.
• Security and fault detection (network intrusion, DOS attacks).
• Fraud detection (credit cards, stocks, voting irregularities).
• Detecting natural disasters (earthquakes, particularly underwater).
• Astronomy (find new classes of stars/planets).
• Genetics (identifying individuals with new/ancient genes).
Classes of Methods for Outlier Detection

1. Model-based methods.
2. Graphical approaches.
3. Cluster-based methods.
4. Distance-based methods.
5. Supervised-learning methods.

• Warning: this is the topic with the most ambiguous “solutions”.
  – Next week we’ll get back to topics with more concrete solutions.
Model-Based Outlier Detection

- **Model-based outlier detection:**
  1. Fit a probabilistic model.
  2. Outliers are **examples with low probability**.

- Simplest approach is **z-score**:
  - If $z_i > 3$, 97% of data is larger than $x_i$.

\[
  z_i = \frac{x_i - \mu}{\sigma}
\]

[Link to Outlier page on Wolfram MathWorld](http://mathworld.wolfram.com/Outlier.html)
Problems with Z-Score

• The z-score relies on mean and standard deviation:
  – These measure are sensitive to outliers.
  – Possible fixes: use quantiles, or sequentially remove worse outlier.

• The z-score also assumes that data is uni-modal...

http://mathworld.wolfram.com/Outlier.html
Global vs. Local Outliers

• Is the red point an outlier?
Global vs. Local Outliers

• Is the red point an outlier? What if add the blue points?
Global vs. Local Outliers

• Is the red point an outlier? What if add the blue points?

• Red point has the lowest $z$-score.
  – In the first case it was a “global” outlier.
  – In this second case it’s a “local” outlier:
    • It’s within the range of the data, but is far away from other points.

• In general, hard to give precise definition of ‘outliers’.
Global vs. Local Outliers

• Is the red point an outlier? What if add the blue points?

- Red point has the lowest z-score.
  - In the first case it was a “global” outlier.
  - In this second case it’s a “local” outlier:
    • It’s within the range of the data, but is far away from other points.

• In general, hard to give precise definition of ‘outliers’.
  - Can we have outlier groups?
Global vs. Local Outliers

• Is the red point an outlier? What if add the blue points?

• Red point has the lowest z-score.
  – In the first case it was a “global” outlier.
  – In this second case it’s a “local” outlier:
    • It’s within the range of the data, but is far away from other points.

• In general, hard to give precise definition of ‘outliers’.
  – Can we have outlier groups?
  – What about repeating patterns?
Graphical Outlier Detection

• Graphical approach to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.

• Examples:
  1. Box plot:
     • Visualization of quantiles/outliers.
     • Only 1 variable at a time.
Graphical Outlier Detection

• **Graphical approach** to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.

• Examples:
  1. Box plot.
  2. Scatterplot:
     • Can detect complex patterns.
     • Only 2 variables at a time.

http://mathworld.wolfram.com/Outlier.html
Graphical Outlier Detection

• **Graphical approach** to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.

• Examples:
  1. Box plot.
  2. Scatterplot.
  3. Scatterplot array:
     • Look at all combinations of variables.
     • But laborious in high-dimensions.
     • Still only 2 variables at a time.

Graphical Outlier Detection

- **Graphical approach** to outlier detection:
  1. Look at a plot of the data.
  2. Human decides if data is an outlier.

- **Examples:**
  1. Box plot.
  2. Scatterplot.
  4. Scatterplot of 2-dimensional PCA:
     - ‘See’ high-dimensional structure.
     - But **PCA** is sensitive to outliers.
     - There might be info in higher PCs.

We’ll cover PCA later in this course.

http://scienceblogs.com/gnxp/2008/08/14/the-genetic-map-of-europe/
Cluster-Based Outlier Detection

• Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don’t belong to clusters.

• Examples:
  1. K-means:
     • Find points that are far away from any mean.
     • Find clusters with a small number of points.
Cluster-Based Outlier Detection

• Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don’t belong to clusters.

• Examples:
  2. Density-based clustering:
     • Outliers are points not assigned to cluster.

Cluster-Based Outlier Detection

• Detect outliers based on clustering:
  1. Cluster the data.
  2. Find points that don’t belong to clusters.

• Examples:
  2. Density-based clustering.
  3. Hierarchical clustering:
     • Outliers take longer to join other groups.
     • Also good for outlier groups.

http://www.nature.com/nature/journal/v438/n7069/fig_tab/nature04338_F10.html
Distance-Based Outlier Detection

• Most of these approaches are based on distances.
• Can we skip the models/plot/clusters and directly use distances?
  – Directly measure of how close objects are to their neighbours.
• Examples:
  – How many points lie in a radius ‘r’?
  – What is distance to $k^{th}$ nearest neighbour?

https://en.wikipedia.org/wiki/Local_outlier_factor
Global Distance-Based Outlier Detection: KNN

• **KNN outlier detection:**
  – For each point, compute the *average distance to its KNN*.
  – Sort these values.
  – Choose the biggest values as outliers.

• Goldstein and Uchida [2016]:
  – Compared 19 methods on 10 datasets.
  – KNN best for finding “global” outliers.
  – “Local” outliers better detected by LOF...

http://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0152173
Local Distance-Based Outlier Detection

• As with density-based clustering, problem with differing densities:
  - Outlier $o_2$ has similar density as elements of cluster $C_1$.
  - Solution: “local outlier factor” (LOF) and variations like outlierness:
    - Is point “relatively” far away from its neighbours?

http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf
Outlierness

- Let $N_k(x_i)$ be the $k$-nearest neighbours of $x_i$.
- Let $D_k(x_i)$ be the average distance to $k$-nearest neighbours:
  \[ D_k(x_i) = \frac{1}{k} \sum_{j \in N_k(x_i)} \|x_i - x_j\| \]
- **Outlierness** is ratio of $D_k(x_i)$ to average $D_k(x_j)$ for its neighbours ‘$j$’:
  \[ O_k(x_i) = \frac{D_k(x_i)}{\frac{1}{k} \sum_{j \in N_k(x_i)} D_k(x_j)} \]
- If outlierness > 1, $x_i$ is further away from neighbours than expected.
Outlierness Ratio

- Outlierness finds $o_1$ and $o_2$:

- More complicated data:

http://www.dbs.ifi.lmu.de/Publikationen/Papers/LOF.pdf
https://en.wikipedia.org/wiki/Local_outlier_factor
Outlierness with Close Clusters

• If clusters are close, outlierness gives unintuitive results:

• In this example, ‘p’ has higher outlierness than ‘q’ and ‘r’:
  – The green points are not part of the KNN list of ‘p’ for small ‘k’.

Outlierness with Close Clusters

• ‘Influenced outlierness’ (INFLO) ratio:
  – Include in denominator the ‘reverse’ k-nearest neighbours:
    • Points that have ‘p’ in KNN list.
  – Adds ‘s’ and ‘t’ from bigger cluster that includes ‘p’:

But still has problems:
  – Dealing with hierarchical clusters.
  – Yields many false positives if you have “global” outliers.
  – Goldstein and Uchida [2016] recommend just using KNN.

Supervised Outlier Detection

• Final approach to outlier detection is to use supervised learning:
  • $y_i = 1$ if $x_i$ is an outlier.
  • $y_i = 0$ if $x_i$ is a regular point.

• Let’s us use our great methods for supervised learning:
  – We can find very complicated outlier patterns.

• But it needs supervision:
  – We need to know what outliers look like.
  – We may not detect new “types” of outliers.
Summary

- **Outlier detection** is task of finding unusually different object.
  - A concept that is very difficult to define.
- **Model-based** methods check if objects are unlikely in fitted model.
- **Graphical** methods plot data and use human to find outliers.
- **Cluster-based** methods check whether objects belong to clusters.
- **Distance-based** methods measure relative distance to neighbours.
- **Supervised-learning** methods just turn it into supervised learning.

- Next time: “customers who bought this item also bought”.