Preprocessing Data

Data Transformation and Reduction
Course Topics

Preliminaries
Data Understanding
Data Preprocessing
Statistical Methods
Classification & Regression
Validation & Interpretation
Advanced Topics
Data Preprocessing Tasks

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction
Wrapping up...

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction
Data Transformation

- Distributional Transform
- Normalization / Standardization
- Discretization
- Feature Construction
Various transformations are used, among the most common are Log, Square Root, and Inverse. While the general rules above apply, it’s typically more useful to test a few and examine the normality of the results.
Data Transformation

- Distributional Transform
- Normalization / Standardization
- Discretization
- Feature Construction
Similarity and Dissimilarity

• Similarity measure or similarity function
  – A real-valued function that quantifies the similarity between two objects
  – Measure how two data objects are alike: The higher value, the more alike
  – Often falls in the range $[0,1]$: 0: no similarity; 1: completely similar

• Dissimilarity (or distance) measure
  – Numerical measure of how different two data objects are
  – In some sense, the inverse of similarity: The lower, the more alike
  – Minimum dissimilarity is often 0 (i.e., completely similar)
  – Range $[0, 1]$ or $[0, \infty)$, depending on the definition
Example: Euclidean Distance

\[ d(i, j) = \sqrt{|x_{i1} - x_{j1}|^2 + |x_{i2} - x_{j2}|^2 + \cdots + |x_{il} - x_{jl}|^2} \]

**Dissimilarity Matrix (by Euclidean Distance)**

<table>
<thead>
<tr>
<th>point</th>
<th>attribute</th>
<th>attribute</th>
<th>attribute</th>
</tr>
</thead>
<tbody>
<tr>
<td>x1</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>x2</td>
<td>3</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>x3</td>
<td>2</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>x4</td>
<td>4</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

\[
\begin{array}{cccc}
  x1 & x2 & x3 & x4 \\
  x1 & 0 & & \\
  x2 & 3.61 & 0 & \\
  x3 & 2.24 & 5.1 & 0 \\
  x4 & 4.24 & 1 & 5.39 & 0 \\
\end{array}
\]
# Real World Considerations

<table>
<thead>
<tr>
<th>Patient</th>
<th>Total Eye Blinks</th>
<th>Heart Attacks</th>
<th>Tumor Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>40000</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>55000</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>27000</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>53000</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>
Real World Considerations

<table>
<thead>
<tr>
<th>Patient</th>
<th>Total Eye Blinks</th>
<th>Heart Attacks</th>
<th>Tumor Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>40000</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>55000</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>27000</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>53000</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>

Distance Between Patients 1-2

\[
\sqrt{(25000 - 40000)^2 + (24 - 27)^2 + (4 - 5)^5} = 15000.000333333333
\]
# Real World Considerations

<table>
<thead>
<tr>
<th>Patient</th>
<th>Total Eye Blinks</th>
<th>Heart Attacks</th>
<th>Tumor Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25000</td>
<td>24</td>
<td>4</td>
</tr>
<tr>
<td>2</td>
<td>40000</td>
<td>27</td>
<td>5</td>
</tr>
<tr>
<td>3</td>
<td>55000</td>
<td>32</td>
<td>7</td>
</tr>
<tr>
<td>4</td>
<td>27000</td>
<td>25</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>53000</td>
<td>30</td>
<td>5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>15000</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>30000</td>
<td>15000</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>2000</td>
<td>13000</td>
<td>28000</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>28000</td>
<td>13000</td>
<td>2000</td>
<td>26000</td>
<td>0</td>
</tr>
</tbody>
</table>
Solution: Normalization

The goal of normalization is to make an entire set of values have a particular property.
Data Transformation: Normalization

- Normalization is often performed on data to remove amplitude variation and only focus on the underlying distribution shape.

- Makes training less sensitive to the scale of features

- Sometimes used in order to speed up the convergence.
Data Transformation: Normalization

- Two Common Techniques:
  - Min-max normalization
  - Z-score Standardization
Min-Max Normalization

• Transform the data from measured units to a new interval from \(new_{min}_F\) to \(new_{max}_F\) for feature \(F\)

• Follows the following formula where \(v\) is the current value of feature \(F\).

\[
v' = \frac{v - min_f}{max_f - min_f} (new_{max}_f - new_{min}_f) + new_{min}_f
\]
Min-Max Normalization: Example

• Suppose that the max and min values for the feature income are $120,000 and $98,000, respectively. We would like to map income to the range $[0.0,1.0]$. By min-max normalization, a value of $103,600$ for income is transformed to:

\[
v' = \frac{v - \text{min}_f}{\text{max}_f - \text{min}_f} (\text{newmax}_f - \text{newmin}_f) + \text{newmin}_f
\]

\[
v' = \frac{103000 - 98000}{120000 - 98000} (1 - 0) + 0 = 0.227
\]
Z-score (zero-mean) Standardization

Transform the data by converting the values to a common scale with an average of zero and a standard deviation of one. A value, $v$, of $A$ is normalized to $v'$ by computing:

$$v' = \frac{v - F}{\sigma_F}$$

where $F$ and $\sigma_F$ are the mean and standard deviation of feature $F$, respectively.
Z-score (zero-mean) Standardization

- The normalized value of $X_i$ is calculated as:

$$Z_i = \frac{X_i - \bar{X}}{s}$$

$$s = \sqrt{\frac{(35 - 51)^2 + (36 - 51)^2 + (46 - 51)^2 + (68 - 51)^2 + (70 - 51)^2}{5 - 1}}$$

$$= \frac{1}{2} \sqrt{(-16)^2 + (-15)^2 + (-5)^2 + 17^2 + 19^2}$$

$$= 17.$$

$$y = \begin{bmatrix} 35 \\ 36 \\ 46 \\ 68 \\ 70 \end{bmatrix}$$

$$z = \begin{bmatrix} \frac{35-51}{17} \\ \frac{36-51}{17} \\ \frac{46-51}{17} \\ \frac{68-51}{17} \\ \frac{70-51}{17} \end{bmatrix} = \begin{bmatrix} -\frac{16}{17} \\ -\frac{15}{17} \\ -\frac{5}{17} \\ \frac{17}{17} \\ \frac{19}{17} \end{bmatrix} = \begin{bmatrix} -0.9412 \\ -0.8824 \\ -0.2941 \\ 1.0000 \\ 1.1176 \end{bmatrix}$$

vs. Min-Max Normalization:

$[0, 1/35, 11/35, 33/35, 1] = [0, 0.0286, 0.3143, 0.9429, 1.0]$
Additional Use

• We will see this again in a few weeks in the modeling portion of this course with regard to:
  – Interpretation of models
  – Stability and convergence
  – Generalizability
Data Transformation

- Distributional Transform
- Normalization / Standardization
- Discretization
- Feature Construction
Discretization

Transformation of continuous data into discrete counterparts.

- Some algorithms require categorical or binary features
- Sometimes exact precision is not required by the research question
- Can improve visualization
- Can reduce categories for feature with many values
Discretization

Unsupervised Discretization
• Binning
• Histogram Analysis

Supervised Discretization
• Cluster
• Decision Tree
• Correlation
Discretization

• Top Down:
  – The process starts by first finding one or a few points (called *split points* or *cut points*) to split the entire attribute range, and then repeats this recursively on the resulting intervals, *splitting*.

• Bottom Up
  – Starts by considering all of the continuous values as potential split-points, removes some by merging neighborhood values to form intervals, and then recursively applies this process to the resulting intervals.
Most Common Methods

• Binning
  – First sort data and partition into bins to get a more general grouping
  – Then one can smooth by bin means, smooth by bin median, smooth by bin boundaries, etc.

• Percentile Thresholds
  – I.e. top and bottom 25\textsuperscript{th} percentile (quartiles)
  – Helps to reduce noise from moderate value instances as well
Discretization: Equal Width Binning

- Divides the range into $N$ intervals of equal size
- If $A$ and $B$ are the lowest and highest values of the attribute, the width of intervals with be:

\[ W = \frac{(B - A)}{N} \]

- The most straight-forward
- But outliers may dominate presentation
- Skewed data is not handled well.
Discretization: Equal Frequency Binning

- It divides the range of size N into K intervals, each containing approximately the same number of samples.
- Good data scaling.
- Managing categorical features can be tricky.
Binning Example

Data = 0, 4, 12, 16, 16, 18, 24, 28

- **Equal Width**
  - Bin 1: 0, 4 \([-\infty, 10)\)
  - Bin 2: 12, 16, 16, 18 \([10, 20)\)
  - Bin 3: 24, 26, 28 \([20, \infty)\)

- **Equal Frequency**
  - Bin 1: 0, 4, 12 \([-\infty, 14)\)
  - Bin 2: 16, 16, 18 \([14, 21)\)
  - Bin 3: 24, 26, 28 \([21, \infty)\)
Binning Smoothing Example

Data = 4, 8, 15, 21, 21, 24, 25, 28, 34

- **Equal Frequency**
  - Bin 1: 4, 8, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 28, 34

- **Smoothing by Bin Mean**
  - Bin 1: 9, 9, 9
  - Bin 2: 22, 22, 22
  - Bin 3: 29, 29, 29

- **Smoothing by Boundaries**
  - Bin 1: 4, 4, 15
  - Bin 2: 21, 21, 24
  - Bin 3: 25, 25, 34
Picking The Right Method
Aggregation / Generalization

- Natural groupings or hierarchies can be leveraged to reduce the feature space. i.e. “rolling up” ICD-9 Codes
  - Internal like the code hierarchy
  - External: SICU and CCU are both surgery units.

ICD-9-CM Code Format

- Category
- Etiology
- Anatomic Site
- Manifestation
- Category
Data Transformation

- Distributional Transform
- Normalization / Standardization
- Discretization
- Feature Construction
Feature Construction

• Creation of novel features for original feature data.
  – Original features may not be suitable for algorithms
  – Sometimes more useful features can be engineered

• Can be:
  – Simple as a numeric Age, derived from individuals DOB
  – Or complex as the beat-to-beat intervals (R-R interval) of an ECG
Wrapping up...

1. Data Cleaning
2. Data Integration
3. Data Transformation
4. Data Reduction
Data Reduction

Reduce the number of features
(Feature Selection / Dimensionality Reduction)

\[
\begin{array}{cccccc}
  & x_1 & x_2 & \ldots & x_m & y \\
1 & & & & & \\
2 & & & & & \\
. & & & & & \\
. & & & & & \\
n & & & & & \\
\end{array}
\]

Reduce the number of instances
(Sampling)

We will cover this in 2 weeks, during modeling.
Sampling
Sampling

• Sampling: obtaining a small sample $s$ to represent the whole data set $N$
• Key principle: Choose a representative subset of the data
  – Simple random sampling may have very poor performance in the presence of skew

**Simple random sampling:**
Equal probability of selecting any particular item

**Sampling without replacement:**
Once an object is selected, it is removed from the population

**Sampling with replacement:**
A selected object is not removed from the population
Sampling

Commonly used for selecting subset of data.

• Typically used because too expensive or time-consuming to process all of the data.

• **Key idea:** obtain a representative sample of the data
Sampling With(out) Replacement

Simple Random Sampling without Replacement

Simple Random Sampling with Replacement
Systematic Sampling

Select instances from an ordered sampling window.

- **Equal-probability method:**
  1. Randomly select element from list
  2. Select every $k^{th}$ element in window, where $k = \frac{N}{n}$.

- Risks interaction with regularities in the data
Simple Random Sampling

Shuffle the data and then select examples

• Avoids regularities.
• What if the dataset is imbalanced?
What if the data is imbalanced?
What if the data is imbalanced?

Random Sample n=8
Stratified Sampling

Partition (or cluster) the data set, and draw samples from each partition (proportionally, i.e., approximately the same percentage of the data)
Stratified Sampling

Stratified Sample n=8
Stratified/Clustered Sampling

Raw Data

Stratified/Clustered Samples
Next Class – Statistics

You're the clinician, tell me what it means.
You're the computational scientist, tell me what's important.

STOP! Put down the data and back away before someone gets hurt.

Big Data Stalemate

http://redpenblackpen.com