START OF DAY 9

Reading: None
Clustering Evaluation
Quality Metrics

• External metrics
  – Computed from a comparison to actual clustering (or classification) labels.

• Internal metrics
  – Computed solely from the composition of a cluster, with no recourse to external information.
External Metrics – Setup (I)

• Let: \( TC = TC_1 \cup TC_2 \cup \ldots \cup TC_n \)
  \( CC = CC_1 \cup CC_2 \cup \ldots \cup CC_m \)

be the target and computed clusterings, respectively.

• \( TC = CC = \) original set of data

• Define the following:
External Metrics – Setup (II)

• \(a\): number of pairs of items that belong to the same cluster in both \(CC\) and \(TC\)
• \(b\): number of pairs of items that belong to different clusters in both \(CC\) and \(TC\)
• \(c\): number of pairs of items that belong to the same cluster in \(CC\) but different clusters in \(TC\)
• \(d\): number of pairs of items that belong to the same cluster in \(TC\) but different clusters in \(CC\)
F-measure

\[ P = \frac{a}{a + c} \]

\[ R = \frac{a}{a + d} \]

\[ F = \frac{2 \times P \times R}{P + R} \]
Rand Index

\[
\frac{a+b}{a+b+c+d}
\]

Measure of clustering agreement: how similar are these two ways of partitioning the data?
Adjusted Rand Index

\[
\frac{2(ab - cd)}{(a + c)(c + b) + (a + d)(d + b)}
\]

Extension of the Rand index that attempts to account for items that may have been clustered by chance
Average Entropy

\[
\text{Entropy}(CC_i) = \sum_{TC_j \in TC} -p(TC_j \mid CC_i) \log p(TC_j \mid CC_i)
\]

\[
\text{AvgEntropy}(CC) = \sum_{i=1}^{m} \frac{|CC_i|}{|CC|} \text{Entropy}(CC_i)
\]

Measure of purity with respect to the target clustering
V-measure

- Entropy-based
- Combines:
  - Homogeneity ($H$): all computed clusters contain only items which are members of the same target cluster/class
  - Completeness ($C$): all data items of a given target cluster/class are in the same computed cluster

$$V = \frac{(1 + \beta) \times H \times C}{(\beta \times H) + C}$$
Internal Metrics

- Assume no target clustering
- Attempt to capture some intrinsic properties of the clustering
  - Distance-based
    - Sum of squares
    - Diameter sum or maximum
    - Sum of minimum distance between clusters, etc.
  - Probability-based
    - E.g., KL divergence

Just a sample presented here
Compactness

• Members of a cluster are all similar and close together
  – One measure of compactness of a cluster is the sum of the squared distances of cluster instances to the cluster centroid

\[
\text{Comp}(C) = \sum_{i=1}^{\|X_c\|} (c - x_i)^2
\]

where \(c\) is the centroid of a cluster \(C\), made up of instances \(X_c\).

• The overall compactness of a particular clustering is just the sum of the compactness of the individual clusters
  – Numeric way to compare different clusterings by seeking clusterings which minimize the compactness metric

• Compactness is usually maximized when there are very many clusters
Separability (I)

• Members of a cluster are sufficiently different from members of another cluster (i.e., cluster dissimilarity)
  – One measure of the separability of two clusters is their squared distance
    \[ dist_{ij} = (c_i - c_j)^2 \]
    where \( c_i \) and \( c_j \) are the cluster centroids
  – Other distance measures are possible (single link, etc.)

• For a clustering which cluster distances should we compare?
Separability (II)

- For each cluster add in the distance to its closest neighbor cluster

\[
\text{Separability} = \sum_{i=1}^{|C|} \min_{j} \text{dist}_{ij}(c_i, c_j)
\]

- Goal: maximize separability

- Separability is usually maximized when there are very few clusters
  - squared distance amplifies larger distances
Davies-Bouldin Index

• Combine compactness and separability into one metric
• Example: Davies-Bouldin $r$ index:

$$scat(c) = \frac{comp(c)}{|X_c|} \quad r_i = \max_{j, j \neq i} \frac{scat(c_i) + scat(c_j)}{dist(c_i, c_j)} \quad r = \frac{1}{|C|} \sum_{i=1}^{|C|} r_i$$

where $|X_c|$ is number of elements in cluster represented by centroid $c$, and $|C|$ is total number of clusters in clustering

• Intuition:
  – Want small scatter – small compactness with many cluster members
  – For each cluster $i$, $r_i$ is maximized for a close neighbor cluster with high scatter (measures worse case close neighbor, hope these are as low as possible)
  – Total clustering score $r$ is just the sum of $r_i$. Lower $r$ better.
  – Davies-Bouldin score $r$ finds a balance of separability (distance) being large and compactness (scatter) being small
KL Divergence (I)

\[ KL(CC_i, CC_j) = \sum_{x \in CC_i \cup CC_j} p(x \mid CC_i) \log \frac{p(x \mid CC_i)}{p(x \mid CC_j)} \]

Measure of the difference in probability distributions between clusters

How do we compute over a clustering, i.e., how do we obtain \( KL(CC) \)?
KL Divergence (II)

• Compute KL as follows: (one possibility)

1. For each cluster
   a) Randomly split cluster in half
   b) Compute resulting KL
2. Repeat step 1 \( N \) times and compute average
3. Sum averages of all clusters to produce the KL of the clustering
Summary

• Two types of evaluation
  – External
  – Internal

• Various metrics have been proposed

• Handle with care
  – Depends on the application
  – Some element of subjectivity
Association Rule Mining
Association Rules

- Association rule induction is a method for market basket analysis.

- It aims at finding regularities in the shopping behavior of customers of supermarkets, mail-order companies, on-line shops etc.

- More specifically:
  Find sets of products that are frequently bought together.

- Example of an association rule:
  
  If a customer buys bread and wine, 
  then she/he will probably also buy cheese.

- Possible applications of found association rules:
  
  - Improve arrangement of products in shelves, on a catalog’s pages.
  - Support of cross-selling (suggestion of other products), product bundling.
  - Fraud detection, technical dependence analysis.
Association Rule Mining

• Clearly not limited to market-basket analysis
• Associations may be found among any set of attributes
  – If a representative votes Yes on issue A and No on issue C, then he/she votes Yes on issue B
  – People who read poetry and listen to classical music also go to the theater
• May be used in recommender systems
## A Market-Basket Analysis Example

<table>
<thead>
<tr>
<th>Transaction Time</th>
<th>Customer Id</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>June 10 '93</td>
<td>2</td>
<td>10, 20</td>
</tr>
<tr>
<td>June 12 '93</td>
<td>5</td>
<td>90</td>
</tr>
<tr>
<td>June 15 '93</td>
<td>2</td>
<td>30</td>
</tr>
<tr>
<td>June 20 '93</td>
<td>2</td>
<td>40, 60, 70</td>
</tr>
<tr>
<td>June 25 '93</td>
<td>4</td>
<td>30</td>
</tr>
<tr>
<td>June 25 '93</td>
<td>3</td>
<td>30, 50, 70</td>
</tr>
<tr>
<td>June 25 '93</td>
<td>1</td>
<td>30</td>
</tr>
<tr>
<td>June 30 '93</td>
<td>1</td>
<td>90</td>
</tr>
<tr>
<td>June 30 '93</td>
<td>4</td>
<td>40, 70</td>
</tr>
<tr>
<td>July 25 '93</td>
<td>4</td>
<td>90</td>
</tr>
</tbody>
</table>
## Terminology

<table>
<thead>
<tr>
<th>Customer Id</th>
<th>Transaction Time</th>
<th>Items Bought</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>June 25 ’03</td>
<td>30</td>
</tr>
<tr>
<td>1</td>
<td>June 30 ’03</td>
<td>90</td>
</tr>
<tr>
<td>2</td>
<td>June 10 ’03</td>
<td>10, 20</td>
</tr>
<tr>
<td>2</td>
<td>June 15 ’03</td>
<td>30</td>
</tr>
<tr>
<td>2</td>
<td>June 20 ’03</td>
<td>40, 60, 70</td>
</tr>
<tr>
<td>3</td>
<td>June 25 ’03</td>
<td>30, 50, 70</td>
</tr>
<tr>
<td>4</td>
<td>June 25 ’03</td>
<td>30</td>
</tr>
<tr>
<td>4</td>
<td>June 30 ’03</td>
<td>40, 70</td>
</tr>
<tr>
<td>4</td>
<td>July 25 ’03</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>June 12 ’03</td>
<td>90</td>
</tr>
</tbody>
</table>

**Transaction**

**Item**

**Itemset**
Association Rules

• Let $U$ be a set of items
  – Let $X, Y \subseteq U$
  – $X \cap Y = \emptyset$

• An association rule is an expression of the form $X \Rightarrow Y$, whose meaning is:
  – If the elements of $X$ occur in some context, then so do the elements of $Y$
Quality Measures

• Let $T$ be the set of all transactions
• We define:

$$\text{support}(X) = \frac{|\{t \in T : X \subseteq t\}|}{|T|}$$

$$\text{support}(X \Rightarrow Y) = \frac{|\{t \in T : X \cup Y \subseteq t\}|}{|T|}$$

$$\text{confidence}(X \Rightarrow Y) = \frac{|\{t \in T : X \cup Y \subseteq t\}|}{|\{t \in T : X \subseteq t\}|}$$

$$\text{lift}(X \Rightarrow Y) = \frac{\text{confidence}(X \Rightarrow Y)}{\text{support}(Y)}$$
Learning Associations

• The purpose of association rule learning is to find “interesting” rules, i.e., rules that meet the following two user-defined conditions:
  – support($X \Rightarrow Y$) $\geq$ MinSupport
  – confidence($X \Rightarrow Y$) $\geq$ MinConfidence
Basic Idea

• Generate all frequent itemsets satisfying the condition on minimum support
• Build all possible rules from these itemsets and check them against the condition on minimum confidence
• All the rules above the minimum confidence threshold are returned for further evaluation
Apriori Principle

• Theorem:
  – If an itemset is frequent, then all of its subsets must also be frequent (the proof is straightforward)

• Corollary:
  – If an itemset is not frequent, then none of its superset will be frequent

• In a bottom up approach, we can discard all non-frequent itemsets
AprioriAll

- $L_1 \leftarrow \emptyset$
- For each item $I_j \in I$
  - $count(\{I_j\}) = | \{T_i : I_j \in T_i\} |$
  - If $count(\{I_j\}) \geq MinSupport \times m$
    - $L_1 \leftarrow L_1 \cup \{\{I_j\}, count(\{I_j\})\}$
- $k \leftarrow 2$
- While $L_{k-1} \neq \emptyset$
  - $L_k \leftarrow \emptyset$
  - For each $(l_1, count(l_1)), (l_2, count(l_2)) \in L_{k-1}$
    - If $(l_1 = \{j_1, \ldots, j_{k-2}, x\} \land l_2 = \{j_1, \ldots, j_{k-2}, y\} \land x \neq y)$
      - $l \leftarrow \{j_1, \ldots, j_{k-2}, x, y\}$
      - $count(l) \leftarrow | \{T_i : l \subseteq T_i\} |$
      - If $count(l) \geq MinSupport \times m$
        - $L_k \leftarrow L_k \cup \{(l, count(l))\}$
  - $k \leftarrow k + 1$
- Return $L_1 \cup L_2 \cup \ldots \cup L_{k-1}$
Apriori: Breadth First Search

1: \{a, d, e\}  
2: \{b, c, d\}  
3: \{a, c, e\}  
4: \{a, c, d, e\}  
5: \{a, e\}  
6: \{a, c, d\}  
7: \{b, c\}  
8: \{a, c, d, e\}  
9: \{c, b, e\}  
10: \{a, d, e\}  

<table>
<thead>
<tr>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>3</td>
<td>7</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

- Example transaction database with 5 items and 10 transactions.
- Minimum support: 30%, i.e., at least 3 transactions must contain the item set.
- All one item sets are frequent → full second level is needed.
Apriori: Breadth First Search

1: \{a, d, e\}  
2: \{b, c, d\}  
3: \{a, c, e\}  
4: \{a, c, d, e\}  
5: \{a, e\}  
6: \{a, c, d\}  
7: \{b, c\}  
8: \{a, c, d, e\}  
9: \{c, b, e\}  
10: \{a, d, e\}

- Determining the support of item sets: For each item set traverse the database and count the transactions that contain it (highly inefficient).
Apriori: Breadth First Search

1: \{a, d, e\}
2: \{b, c, d\}
3: \{a, c, e\}
4: \{a, c, d, e\}
5: \{a, e\}
6: \{a, c, d\}
7: \{b, c\}
8: \{a, c, d, e\}
9: \{c, b, e\}
10: \{a, d, e\}

- Minimum support: 30\%, i.e., at least 3 transactions must contain the item set.
- Infrequent item sets: \{a, b\}, \{b, d\}, \{b, e\}.
- The subtrees starting at these item sets can be pruned.
Apriori: Breadth First Search

1: \{a, d, e\}
2: \{b, c, d\}
3: \{a, c, e\}
4: \{a, c, d, e\}
5: \{a, e\}
6: \{a, c, d\}
7: \{b, c\}
8: \{a, c, d, e\}
9: \{c, b, e\}
10: \{a, d, e\}

- Generate candidate item sets with 3 items (parents must be frequent).
- Before counting, check whether the candidates contain an infrequent item set.
  - An item set with \(k\) items has \(k\) subsets of size \(k - 1\).
  - The parent is only one of these subsets.
Apriori: Breadth First Search

1: \{a, d, e\}
2: \{b, c, d\}
3: \{a, c, e\}
4: \{a, c, d, e\}
5: \{a, e\}
6: \{a, c, d\}
7: \{b, c\}
8: \{a, c, d, e\}
9: \{c, b, e\}
10: \{a, d, e\}

- The item sets \{b, c, d\} and \{b, c, e\} can be pruned, because
  - \{b, c, d\} contains the infrequent item set \{b, d\} and
  - \{b, c, e\} contains the infrequent item set \{b, e\}. 
Apriori: Breadth First Search

1: \{a, d, e\}
2: \{b, c, d\}
3: \{a, c, e\}
4: \{a, c, d, e\}
5: \{a, e\}
6: \{a, c, d\}
7: \{b, c\}
8: \{a, c, d, e\}
9: \{c, b, e\}
10: \{a, d, e\}

- Only the remaining four item sets of size 3 are evaluated.
Apriori: Breadth First Search

1: \(\{a, d, e\}\)
2: \(\{b, c, d\}\)
3: \(\{a, c, e\}\)
4: \(\{a, c, d, e\}\)
5: \(\{a, e\}\)
6: \(\{a, c, d\}\)
7: \(\{b, c\}\)
8: \(\{a, c, d, e\}\)
9: \(\{c, b, e\}\)
10: \(\{a, d, e\}\)

- Minimum support: 30%, i.e., at least 3 transactions must contain the item set.
- Infrequent item set: \(\{c, d, e\}\).
Apriori: Breadth First Search

1: \{a, d, e\}
2: \{b, c, d\}
3: \{a, c, e\}
4: \{a, c, d, e\}
5: \{a, e\}
6: \{a, c, d\}
7: \{b, c\}
8: \{a, c, d, e\}
9: \{c, b, e\}
10: \{a, d, e\}

- Generate candidate item sets with 4 items (parents must be frequent).
- Before counting, check whether the candidates contain an infrequent item set.
Apriori: Breadth First Search

1: \(\{a, d, e\}\)
2: \(\{b, c, d\}\)
3: \(\{a, c, e\}\)
4: \(\{a, c, d, e\}\)
5: \(\{a, e\}\)
6: \(\{a, c, d\}\)
7: \(\{b, c\}\)
8: \(\{a, c, d, e\}\)
9: \(\{c, b, e\}\)
10: \(\{a, d, e\}\)

- The item set \(\{a, c, d, e\}\) can be pruned, because it contains the infrequent item set \(\{c, d, e\}\).
- Consequence: No candidate item sets with four items.
- Fourth access to the transaction database is not necessary.
## Illustrative Training Set

### Risk Assessment for Loan Applications

<table>
<thead>
<tr>
<th>Client #</th>
<th>Credit History</th>
<th>Debt Level</th>
<th>Collateral</th>
<th>Income Level</th>
<th>RISK LEVEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bad</td>
<td>High</td>
<td>None</td>
<td>Low</td>
<td>HIGH</td>
</tr>
<tr>
<td>2</td>
<td>Unknown</td>
<td>High</td>
<td>None</td>
<td>Medium</td>
<td>HIGH</td>
</tr>
<tr>
<td>3</td>
<td>Unknown</td>
<td>Low</td>
<td>None</td>
<td>Medium</td>
<td>MODERATE</td>
</tr>
<tr>
<td>4</td>
<td>Unknown</td>
<td>Low</td>
<td>None</td>
<td>Low</td>
<td>HIGH</td>
</tr>
<tr>
<td>5</td>
<td>Unknown</td>
<td>Low</td>
<td>None</td>
<td>High</td>
<td>LOW</td>
</tr>
<tr>
<td>6</td>
<td>Unknown</td>
<td>Low</td>
<td>Adequate</td>
<td>High</td>
<td>LOW</td>
</tr>
<tr>
<td>7</td>
<td>Bad</td>
<td>Low</td>
<td>None</td>
<td>Low</td>
<td>HIGH</td>
</tr>
<tr>
<td>8</td>
<td>Bad</td>
<td>Low</td>
<td>Adequate</td>
<td>High</td>
<td>MODERATE</td>
</tr>
<tr>
<td>9</td>
<td>Good</td>
<td>Low</td>
<td>None</td>
<td>High</td>
<td>LOW</td>
</tr>
<tr>
<td>10</td>
<td>Good</td>
<td>High</td>
<td>Adequate</td>
<td>High</td>
<td>LOW</td>
</tr>
<tr>
<td>11</td>
<td>Good</td>
<td>High</td>
<td>None</td>
<td>Low</td>
<td>HIGH</td>
</tr>
<tr>
<td>12</td>
<td>Good</td>
<td>High</td>
<td>None</td>
<td>Medium</td>
<td>MODERATE</td>
</tr>
<tr>
<td>13</td>
<td>Good</td>
<td>High</td>
<td>None</td>
<td>High</td>
<td>LOW</td>
</tr>
<tr>
<td>14</td>
<td>Bad</td>
<td>High</td>
<td>None</td>
<td>Medium</td>
<td>HIGH</td>
</tr>
</tbody>
</table>
Running Apriori (I)

• Items:
  - (CH=Bad, .29) (CH=Unknown, .36) (CH=Good, .36)
  - (DL=Low, .5) (DL=High, .5)
  - (C=None, .79) (C=Adequate, .21)
  - (IL=Low, .29) (IL=Medium, .29) (IL=High, .43)
  - (RL=High, .43) (RL=Moderate, .21) (RL=Low, .36)

• Choose MinSupport=.4 and MinConfidence=.8
Running Apriori (II)

- $L_1 = \{(DL=\text{Low}, .5); (DL=\text{High}, .5); (C=\text{None}, .79); (IL=\text{High}, .43); (RL=\text{High}, .43)\}$

- $L_2 = \{(DL=\text{High} + C=\text{None}, .43)\}$

- $L_3 = \{\}$
Running Apriori (III)

• Two possible rules:
  – DL = High $\Rightarrow$ C = None (A)
  – C = None $\Rightarrow$ DL = High (B)

• Confidences:
  – Conf(A) = .86  Retain
  – Conf(B) = .54  Ignore
Summary

• Note the following about Apriori:
  – A “true” data mining algorithm
  – Easy to implement with a sparse matrix and simple sums
  – Computationally expensive
    • Actual run-time depends on \( \text{MinSupport} \)
    • In the worst-case, time complexity is \( O(2^n) \)
    • Efficient implementations exist (e.g., FP-Growth)
  – Multiple supports
  – Other interestingness measures (about 61!)
Logistic Regression
Regression

• A form of statistical modeling that attempts to evaluate the relationship between one variable (termed the dependent variable) and one or more other variables (termed the independent variables). It is a form of global analysis as it only produces a single equation for the relationship.

• A model for predicting one variable from another.
Linear Regression

- Regression used to fit a linear model to data where the dependent variable is continuous:

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n + \varepsilon \]

- Given a set of points \((X_i, Y_i)\), we wish to find a linear function (or line in 2 dimensions) that “goes through” these points.

- In general, the points are not exactly aligned:
  - Find line that best fits the points
• Error or residue:
  – Observed value - Predicted value

Chart Title

- Observed
- Linear (Observed)
Sum-squared Error (SSE)

\[ SSE = \sum_y (y_{\text{observed}} - y_{\text{predicted}})^2 \]

\[ TSS = \sum_y (y_{\text{observed}} - \bar{y}_{\text{observed}})^2 \]

\[ R^2 = 1 - \frac{SSE}{TSS} \]
What is Best Fit?

• The smaller the SSE, the better the fit
• Hence,
  – Linear regression attempts to minimize SSE (or similarly to maximize R2)
• Assume 2 dimensions

\[ Y = \beta_0 + \beta_1 X \]
Analytical Solution

Take partial derivatives of SSE with respect to the coefficients, set to 0, and solve

\[
\beta_0 = \frac{\sum y - \beta_1 \sum x}{n}
\]

\[
\beta_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - \left(\sum x\right)^2}
\]
Example (I)

<table>
<thead>
<tr>
<th>x</th>
<th>y</th>
<th>x^2</th>
<th>xy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>4.00</td>
<td>1.44</td>
<td>4.80</td>
</tr>
<tr>
<td>2.30</td>
<td>5.60</td>
<td>5.29</td>
<td>12.88</td>
</tr>
<tr>
<td>3.10</td>
<td>7.90</td>
<td>9.61</td>
<td>24.49</td>
</tr>
<tr>
<td>3.40</td>
<td>8.00</td>
<td>11.56</td>
<td>27.20</td>
</tr>
<tr>
<td>4.00</td>
<td>10.10</td>
<td>16.00</td>
<td>40.40</td>
</tr>
<tr>
<td>4.60</td>
<td>10.40</td>
<td>21.16</td>
<td>47.84</td>
</tr>
<tr>
<td>5.50</td>
<td>12.00</td>
<td>30.25</td>
<td>66.00</td>
</tr>
<tr>
<td><strong>24.10</strong></td>
<td><strong>58.00</strong></td>
<td><strong>95.31</strong></td>
<td><strong>223.61</strong></td>
</tr>
</tbody>
</table>

Target: \( y = 2x + 1.5 \)

\[
\beta_1 = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2} = \frac{7 \times 223.61 - 24.10 \times 58.00}{7 \times 95.31 - 24.10^2} = \frac{1565.27 - 1397.80}{667.17 - 580.81} = \frac{167.47}{86.36} = 1.94
\]

\[
\beta_0 = \frac{\sum y - \beta_1 \sum x}{n} = \frac{58.00 - 1.94 \times 24.10}{7} = \frac{11.27}{7} = 1.61
\]
Example (II)
Example (III)

\[
R^2 = 1 - \frac{SSE}{TSS} = 1 - \frac{0.975}{47.369} = 0.98
\]

<table>
<thead>
<tr>
<th>x</th>
<th>y (obs)</th>
<th>y (pred)</th>
<th>SSE</th>
<th>TSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.20</td>
<td>4.00</td>
<td>3.94</td>
<td>0.004</td>
<td>18.367</td>
</tr>
<tr>
<td>2.30</td>
<td>5.60</td>
<td>6.07</td>
<td>0.221</td>
<td>7.213</td>
</tr>
<tr>
<td>3.10</td>
<td>7.90</td>
<td>7.62</td>
<td>0.078</td>
<td>0.149</td>
</tr>
<tr>
<td>3.40</td>
<td>8.00</td>
<td>8.21</td>
<td>0.044</td>
<td>0.082</td>
</tr>
<tr>
<td>4.00</td>
<td>10.10</td>
<td>9.37</td>
<td>0.533</td>
<td>3.292</td>
</tr>
<tr>
<td>4.60</td>
<td>10.40</td>
<td>10.53</td>
<td>0.017</td>
<td>4.470</td>
</tr>
<tr>
<td>5.50</td>
<td>12.00</td>
<td>12.28</td>
<td>0.078</td>
<td>13.796</td>
</tr>
</tbody>
</table>

\[
\text{SSE} = 0.975, \quad \text{TSS} = 47.369
\]
Multiple Linear Regression

\[ Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n \]

- There is a closed form for finding multiple linear regression weights which requires matrix inversion, etc.
- There are also iterative techniques to find weights
  - E.g., delta rule: use an output node which is not thresholded (just does a linear sum) and iteratively apply the delta rule
    \[ \Delta w_i = \eta \sum_{d \in D} (t_d - net_d) x_{id} \]
- Delta rule will update towards the objective of minimizing the SSE, thus solving multiple linear regression
- There are many other regression approaches that give different results by trying to better handle outliers and other statistical anomalies
Intelligibility

• One advantage of linear regression models is the potential to look at the coefficients to give insight into which input variables are most important in predicting the output
• The variables with the largest magnitude have the highest correlation with the output
  – A large positive coefficient implies that the output will increase when this input is increased (positively correlated)
  – A large negative coefficient implies that the output will decrease when this input is increased (negatively correlated)
  – A small or 0 coefficient suggests that the input is uncorrelated with the output
• Linear regression can be used to find best "indicators"
• However, careful not to confuse correlation with causality
Non-Linear Tasks

• Linear Regression will not generalize well to the task below: needs a non-linear surface
• Could pre-process features as before
  – E.g., use an arbitrary polynomial in $x$
  – Thus it is still \textbf{linear in the coefficients}, and can be solved with delta rule, etc.

\[ Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n \]

Instead we will focus on classification
Logistic Regression

• Regression used to fit a curve to data in which the dependent variable is binary, or dichotomous (i.e., classification)

• Typical application: Medicine
  – We might want to predict response to treatment, where we might code survivors as 1 and those who don’t survive as 0
**Example**

**Observations:**
For each value of SurvRate, the number of dots is the number of patients with that value of NewOut.

**Regression:**
Standard linear regression

**Problem:** extending the regression line a few units left or right along the X axis produces predicted probabilities that fall outside of [0, 1]
A Better Solution

Regression Curve: Sigmoid function!
(bounded by asymptotes $y=0$ and $y=1$)
Odds

• Given some event with probability $p$ of being 1, the odds of that event are given by:

$$\text{odds} = \frac{p}{1-p}$$

• Consider the following data

<table>
<thead>
<tr>
<th>Testosterone</th>
<th>Delinquent</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
<td>Total</td>
<td></td>
</tr>
<tr>
<td>Normal</td>
<td>402</td>
<td>3614</td>
<td>4016</td>
<td></td>
</tr>
<tr>
<td>High</td>
<td>101</td>
<td>345</td>
<td>446</td>
<td></td>
</tr>
<tr>
<td></td>
<td>503</td>
<td>3959</td>
<td>4462</td>
<td></td>
</tr>
</tbody>
</table>

• The odds of being delinquent if you are in the Normal group are:

$$\text{pdelinquent}/(1-\text{pdelinquent}) = (402/4016) / (1 - (402/4016)) = 0.1001 / 0.8889 = 0.111$$
Odds Ratio

• The odds of being not delinquent in the Normal group is the reciprocal of this:
  – \( \frac{0.8999}{0.1001} = 8.99 \)

• Now, for the High testosterone group
  – odds(delinquent) = \( \frac{101}{345} = 0.293 \)
  – odds(not delinquent) = \( \frac{345}{101} = 3.416 \)

• When we go from Normal to High, the odds of being delinquent nearly triple:
  – Odds ratio: \( \frac{0.293}{0.111} = 2.64 \)
  – 2.64 times more likely to be delinquent with high testosterone levels
Logit Transform

- The logit is the natural log of the odds

\[ \text{logit}(p) = \ln(\text{odds}) = \ln \left( \frac{p}{1-p} \right) \]
Logistic Regression

- In logistic regression, we seek a model:

\[ \text{logit}(p) = \beta_0 + \beta_1 X \]

- That is, the log odds (logit) is assumed to be linearly related to the independent variable X

- So, now we can focus on solving an ordinary (linear) regression!
Recovering Probabilities

\[
\ln\left( \frac{p}{1 - p} \right) = \beta_0 + \beta_1 X
\]

\[\Leftrightarrow \frac{p}{1 - p} = e^{\beta_0 + \beta_1 X}\]

\[\Leftrightarrow p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}}\]

which gives \( p \) as a sigmoid function!
Logistic Response Function

- When the response variable is binary, the shape of the response function is often sigmoidal:
Interpretation of $\beta_1$

• Let:
  – odds1 = odds for value X ($p/(1-p)$)
  – odds2 = odds for value X + 1 unit

• Then:

\[
\frac{odds2}{odds1} = \frac{e^{\beta_0 + \beta_1(X + 1)}}{e^{\beta_0 + \beta_1 X}}
\]

\[
= \frac{e^{(\beta_0 + \beta_1 X) + \beta_1}}{e^{\beta_0 + \beta_1 X}} = \frac{e^{(\beta_0 + \beta_1 X)} e^{\beta_1}}{e^{\beta_0 + \beta_1 X}} = e^{\beta_1}
\]

• Hence, the exponent of the slope describes the proportionate rate at which the predicted odds ratio changes with each successive unit of X
Sample Calculations

• Suppose a cancer study yields:
  – log odds = $-2.6837 + 0.0812 \text{ SurvRate}$

• Consider a patient with SurvRate = 40
  – log odds = $-2.6837 + 0.0812(40) = 0.5643$
  – odds = $e^{0.5643} = 1.758$
  – patient is 1.758 times more likely to be improved than not

• Consider another patient with SurvRate = 41
  – log odds = $-2.6837 + 0.0812(41) = 0.6455$
  – odds = $e^{0.6455} = 1.907$
  – patient’s odds are $1.907/1.758 = 1.0846$ times (or 8.5%) better than those of the previous patient

• Using probabilities
  – $p_{40} = 0.6374$ and $p_{41} = 0.6560$
  – Improvements appear different with odds and with $p$
Example 1 (I)

- A systems analyst studied the effect of computer programming experience on ability to complete a task within a specified time.
- Twenty-five persons selected for the study, with varying amounts of computer experience (in months).
- Results are coded in binary fashion: $Y = 1$ if task completed successfully; $Y = 0$, otherwise.

Scatter Plot and Loess Curve—Programming Task Example.

Loess: form of local regression.
Example 1 (II)

• Results from a standard package give:
  \( \beta_0 = -3.0597 \) and \( \beta_1 = 0.1615 \)

• Estimated logistic regression function:

\[
p = \frac{1}{1 + e^{3.0597 - 0.1615X}}
\]

• For example, the fitted value for \( X = 14 \) is:

\[
p = \frac{1}{1 + e^{3.0597 - 0.1615 \times 14}} = 0.31
\]

(Estimated probability that a person with 14 months experience will successfully complete the task)
Example 1 (III)

• We know that the probability of success increases sharply with experience
  – Odds ratio: \( \exp(\beta_1) = e^{0.1615} = 1.175 \)
  – Odds increase by 17.5% with each additional month of experience

• A unit increase of one month is quite small, and we might want to know the change in odds for a longer difference in time
  – For \( c \) units of \( X \): \( \exp(c\beta_1) \)
Example 1 (IV)

- Suppose we want to compare individuals with relatively little experience to those with extensive experience, say 10 months versus 25 months (c = 15)
  - Odds ratio: $e^{15x0.1615} = 11.3$
  - Odds of completing the task increase 11-fold!
Example 2 (I)

- In a study of the effectiveness of coupons offering a price reduction, 1,000 homes were selected and coupons mailed.
- Coupon price reductions: 5, 10, 15, 20, and 30 dollars.
- 200 homes assigned at random to each coupon value.
- $X$: amount of price reduction.
- $Y$: binary variable indicating whether or not coupon was redeemed.
Example 2 (II)

- Fitted response function
  - $\beta_0 = -2.04$ and $\beta_1 = 0.097$

- Odds ratio: $\exp(\beta_1) = e^{0.097} = 1.102$

- Odds of a coupon being redeemed are estimated to increase by 10.2% with each $1$ increase in the coupon value (i.e., $1$ in price reduction)
Putting it to Work

• For each value of X, you may not have probability but rather a number of $<x,y>$ pairs from which you can extract frequencies and hence probabilities
  – Raw data: $<12,0>, <12,1>, <14,0>, <12,1>, <14,1>, <14,1>, <12,0>, <12,0>$
  – Probability data (p=1, 3rd entry is number of occurrences in raw data): $<12, 0.4, 5>, <14, 0.66, 3>$
  – Odds ratio data...
## Coronary Heart Disease (I)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>Coronary Heart Disease</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>1</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>13</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>9</td>
<td>3</td>
</tr>
<tr>
<td>4</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>5</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>6</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>13</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>8</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>57</td>
<td>43</td>
</tr>
</tbody>
</table>
## Coronary Heart Disease (II)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>p(CHD)(=1)</th>
<th>odds</th>
<th>log odds</th>
<th>#occ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1000</td>
<td>0.1111</td>
<td>-2.1972</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>0.1333</td>
<td>0.1538</td>
<td>-1.8718</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>0.2500</td>
<td>0.3333</td>
<td>-1.0986</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>0.3333</td>
<td>0.5000</td>
<td>-0.6931</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>0.4615</td>
<td>0.8571</td>
<td>-0.1542</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>0.6250</td>
<td>1.6667</td>
<td>0.5108</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>0.7647</td>
<td>3.2500</td>
<td>1.1787</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>0.8000</td>
<td>4.0000</td>
<td>1.3863</td>
<td>10</td>
</tr>
</tbody>
</table>
Coronary Heart Disease (III)

<table>
<thead>
<tr>
<th>X (AG)</th>
<th>Y (log odds)</th>
<th>X^2</th>
<th>XY</th>
<th>#occ</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-2.1972</td>
<td>1.0000</td>
<td>-2.1972</td>
<td>10</td>
</tr>
<tr>
<td>2</td>
<td>-1.8718</td>
<td>4.0000</td>
<td>-3.7436</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>-1.0986</td>
<td>9.0000</td>
<td>-3.2958</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>-0.6931</td>
<td>16.0000</td>
<td>-2.7726</td>
<td>15</td>
</tr>
<tr>
<td>5</td>
<td>-0.1542</td>
<td>25.0000</td>
<td>-0.7708</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>0.5108</td>
<td>36.0000</td>
<td>3.0650</td>
<td>8</td>
</tr>
<tr>
<td>7</td>
<td>1.1787</td>
<td>49.0000</td>
<td>8.2506</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>1.3863</td>
<td>64.0000</td>
<td>11.0904</td>
<td>10</td>
</tr>
<tr>
<td>448</td>
<td>-37.6471</td>
<td>2504.0000</td>
<td>106.3981</td>
<td>100</td>
</tr>
</tbody>
</table>

Note: the sums reflect the number of occurrences
(Sum(X) = X1.#occ(X1)+...+X8.#occ(X8), etc.)
Coronary Heart Disease (IV)

- Results from regression:
  - \( \beta_0 = -2.856 \) and \( \beta_1 = 0.5535 \)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>p(CHD)=1</th>
<th>est. p</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1000</td>
<td>0.0909</td>
</tr>
<tr>
<td>2</td>
<td>0.1333</td>
<td>0.1482</td>
</tr>
<tr>
<td>3</td>
<td>0.2500</td>
<td>0.2323</td>
</tr>
<tr>
<td>4</td>
<td>0.3333</td>
<td>0.3448</td>
</tr>
<tr>
<td>5</td>
<td>0.4615</td>
<td>0.4778</td>
</tr>
<tr>
<td>6</td>
<td>0.6250</td>
<td>0.6142</td>
</tr>
<tr>
<td>7</td>
<td>0.7647</td>
<td>0.7346</td>
</tr>
<tr>
<td>8</td>
<td>0.8000</td>
<td>0.8280</td>
</tr>
</tbody>
</table>

\( \text{SSE} \quad 0.0028 \)
\( \text{TSS} \quad 0.5265 \)
\( \text{R2} \quad 0.9946 \)
Summary

• Regression is a powerful machine learning technique
  – It provides prediction
  – It offers insight on the relative power of each variable
Homework: Backpropagation Learning

END OF DAY 9