Feature Learning I
Features vs. Labels

Features: describing your data objects
- color, texture, taste…

Labels: discriminating your data objects
- apples, pears
Unsupervised Setting

Fruits

Cluster 1

Cluster 2
Supervised Setting

Training

Pears

Apples

Testing

Apple!
Semi-Supervised Setting

Transduction

Induction
Projection

$y_i = w^\top x_i$
Principle Component Analysis

Find a direction where the data have the largest variance

\[
\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i
\]

\[
\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^\top (x_i - \bar{x})
\]

\[
\frac{1}{n} \sum_{i=1}^{n} (w^\top x_i - w^\top \bar{x})^2
\]

\[
\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i = \frac{1}{n} \sum_{i=1}^{n} w^\top x_i = w^\top \left( \frac{1}{n} \sum_{i=1}^{n} x_i \right) = w^\top \bar{x}
\]
Principle Component Analysis

1. Compute data covariance matrix
2. Do eigenvalue decomposition on the covariance matrix
3. Sort the eigenvalues from large to small

$$C = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})^\top$$
Whitening Transform

\[ C = U \Lambda U^T \]

\[ Y = U \Lambda^{-1/2} U^T X \]
Two Dimensional Principle Component Analysis

\[ J(\mathbf{X}) = \text{trace}\left\{ E\left[ (\mathbf{Y} - E\mathbf{Y})(\mathbf{Y} - E\mathbf{Y})^T \right] \right\} = \text{trace}\left\{ E\left[ (\mathbf{AX} - E(\mathbf{AX}))(\mathbf{AX} - E(\mathbf{AX}))^T \right] \right\} = \text{trace}\left\{ \mathbf{X}^T E\left[ (\mathbf{A} - E\mathbf{A})^T (\mathbf{A} - E\mathbf{A}) \right] \mathbf{X} \right\} \]

\[ E\mathbf{Y} = \bar{\mathbf{Y}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{AX}_i \]

\[ \mathbf{G} = E\left[ (\mathbf{A} - E\mathbf{A})^T (\mathbf{A} - E\mathbf{A}) \right] \]

\[ \mathbf{G} = \frac{1}{M} \sum_{k=1}^{M} (\mathbf{A}_k - \bar{\mathbf{A}})^T (\mathbf{A}_k - \bar{\mathbf{A}}) \]

\[ \mathbf{Y} = \mathbf{AX} \]
With Label Information
Linear Discriminant Analysis

\[ y = w^T x \]

Within class compactness:
\[ s_i^2 = \sum_{y \in \omega_i} (y - \mu_i)^2 = \sum_{x \in \omega_i} (w^T x - w^T \mu_i)^2 = \sum_{x \in \omega_i} w^T (x - \mu_i)(x - \mu_i)^T w = w^T S_i w \]

Between class Scatterness:
\[ (\mu_1 - \mu_2)^2 = (w^T \mu_1 - w^T \mu_2)^2 = w^T (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T S_B w \]

\[ J(w) = \frac{w^T S_B w}{w^T S_w w} \]
Relevant Component Analysis

- Chunklet: data have same but unknown class labels

  [Diagrams of unlabeled data, chunklet data, labeled data]

- Sum of in-chunklet covariance matrices for $p$ points in $k$ chunklets:

  $$
  C = \frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{n_j} (\hat{x}_{ji} - \hat{m}_j)(\hat{x}_{ji} - \hat{m}_j)^T, \quad \text{chunklet } j : \{\hat{x}_{ji}\}_{i=1}^{n_j}, \text{ with mean } \hat{m}_j
  $$

- Apply linear transformation

  $$
  y = C^{-\frac{1}{2}} x
  $$
Relevant Component Analysis

- (a) The fully labeled data set with 3 classes.
- (b) Same data unlabeled; classes' structure is less evident.
- (c) The set of chunklets
- (d) The centered chunklets, and their empirical covariance.
- (e) The RCA transformation applied to the chunklets. (centered)
- (f) The original data after applying the RCA transformation.
Transactions vs. Sequences

- A sequence database consists of ordered elements or events
- Transaction databases vs. sequence databases

<table>
<thead>
<tr>
<th>TID</th>
<th>itemsets</th>
<th>SID</th>
<th>sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>a, b, d</td>
<td>10</td>
<td>&lt;a(abc)(ac)d(cf)&gt;</td>
</tr>
<tr>
<td>20</td>
<td>a, c, d</td>
<td>20</td>
<td>&lt;(ad)c(bc)(ae)&gt;</td>
</tr>
<tr>
<td>30</td>
<td>a, d, e</td>
<td>30</td>
<td>&lt;(ef)(ab)(df)cb&gt;</td>
</tr>
<tr>
<td>40</td>
<td>b, e, f</td>
<td>40</td>
<td>&lt;eg(af)cbc&gt;</td>
</tr>
</tbody>
</table>
Subsequence vs. Supersequence

- A sequence is an ordered list of events, denoted $<e_1 e_2 \ldots e_i>$
- Given two sequences $\alpha=<a_1 a_2 \ldots a_n>$ and $\beta=<b_1 b_2 \ldots b_m>$
- $\alpha$ is called a subsequence of $\beta$, denoted as $\alpha \subseteq \beta$, if there exist integers $1 \leq j_1 < j_2 < \ldots < j_n \leq m$ such that $a_1 \subseteq b_{j_1}$, $a_2 \subseteq b_{j_2}$, $\ldots$, $a_n \subseteq b_{j_n}$
- $\beta$ is a super sequence of $\alpha$
  - E.g. $\alpha=<(ab), d>$ and $\beta=<(abc), (de)>
Apriori Property

- A basic property: Apriori (Agrawal & Sirkant’ 94)
  - If a sequence $S$ is not frequent, then none of the super-sequences of $S$ is frequent
  - E.g, $<hb>$ is infrequent $\rightarrow$ so do $<hab>$ and $<(ah)b>$

<table>
<thead>
<tr>
<th>Seq. ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>$&lt;(bd)cb(ac)&gt;$</td>
</tr>
<tr>
<td>20</td>
<td>$&lt;(bf)(ce)b(fg)&gt;$</td>
</tr>
<tr>
<td>30</td>
<td>$&lt;(ah)(bf)abf&gt;$</td>
</tr>
<tr>
<td>40</td>
<td>$&lt;(be)(ce)d&gt;$</td>
</tr>
<tr>
<td>50</td>
<td>$&lt;a(bd)bcb(ade)&gt;$</td>
</tr>
</tbody>
</table>

Given **support threshold**

$min\_sup = 2$
Frequent Itemset Mining

- Initial candidates:
  - <a>, <b>, <c>, <d>, <e>, <f>, <g>, <h>
- Scan database once, count support for candidates

\[ \text{min\_sup} = 2 \]

<table>
<thead>
<tr>
<th>Seq. ID</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>(bd)cb(ac)</td>
</tr>
<tr>
<td>20</td>
<td>(bf)(ce)b(fg)</td>
</tr>
<tr>
<td>30</td>
<td>(ah)(bf)abf</td>
</tr>
<tr>
<td>40</td>
<td>(be)(ce)d</td>
</tr>
<tr>
<td>50</td>
<td>a(bd)bcb(ade)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cand</th>
<th>Sup</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;a&gt;</td>
<td>3</td>
</tr>
<tr>
<td>&lt;b&gt;</td>
<td>5</td>
</tr>
<tr>
<td>&lt;c&gt;</td>
<td>4</td>
</tr>
<tr>
<td>&lt;d&gt;</td>
<td>3</td>
</tr>
<tr>
<td>&lt;e&gt;</td>
<td>3</td>
</tr>
<tr>
<td>&lt;f&gt;</td>
<td>2</td>
</tr>
<tr>
<td>&lt;g&gt;</td>
<td>1</td>
</tr>
<tr>
<td>&lt;h&gt;</td>
<td>1</td>
</tr>
</tbody>
</table>
## Frequent Itemset Mining

### 51 length-2 Candidates

<table>
<thead>
<tr>
<th></th>
<th>(&lt;a&gt;)</th>
<th>(&lt;b&gt;)</th>
<th>(&lt;c&gt;)</th>
<th>(&lt;d&gt;)</th>
<th>(&lt;e&gt;)</th>
<th>(&lt;f&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt;a&gt;)</td>
<td>(&lt;aa&gt;)</td>
<td>(&lt;ab&gt;)</td>
<td>(&lt;ac&gt;)</td>
<td>(&lt;ad&gt;)</td>
<td>(&lt;ae&gt;)</td>
<td>(&lt;af&gt;)</td>
</tr>
<tr>
<td>(&lt;b&gt;)</td>
<td>(&lt;ba&gt;)</td>
<td>(&lt;bb&gt;)</td>
<td>(&lt;bc&gt;)</td>
<td>(&lt;bd&gt;)</td>
<td>(&lt;be&gt;)</td>
<td>(&lt;bf&gt;)</td>
</tr>
<tr>
<td>(&lt;c&gt;)</td>
<td>(&lt;ca&gt;)</td>
<td>(&lt;cb&gt;)</td>
<td>(&lt;cc&gt;)</td>
<td>(&lt;cd&gt;)</td>
<td>(&lt;ce&gt;)</td>
<td>(&lt;cf&gt;)</td>
</tr>
<tr>
<td>(&lt;d&gt;)</td>
<td>(&lt;da&gt;)</td>
<td>(&lt;db&gt;)</td>
<td>(&lt;dc&gt;)</td>
<td>(&lt;dd&gt;)</td>
<td>(&lt;de&gt;)</td>
<td>(&lt;df&gt;)</td>
</tr>
<tr>
<td>(&lt;e&gt;)</td>
<td>(&lt;ea&gt;)</td>
<td>(&lt;eb&gt;)</td>
<td>(&lt;ec&gt;)</td>
<td>(&lt;ed&gt;)</td>
<td>(&lt;ee&gt;)</td>
<td>(&lt;ef&gt;)</td>
</tr>
<tr>
<td>(&lt;f&gt;)</td>
<td>(&lt;fa&gt;)</td>
<td>(&lt;fb&gt;)</td>
<td>(&lt;fc&gt;)</td>
<td>(&lt;fd&gt;)</td>
<td>(&lt;fe&gt;)</td>
<td>(&lt;ff&gt;)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>(&lt;a&gt;)</th>
<th>(&lt;b&gt;)</th>
<th>(&lt;c&gt;)</th>
<th>(&lt;d&gt;)</th>
<th>(&lt;e&gt;)</th>
<th>(&lt;f&gt;)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(&lt;a&gt;)</td>
<td>(&lt;ab&gt;)</td>
<td>(&lt;ac&gt;)</td>
<td>(&lt;ad&gt;)</td>
<td>(&lt;ae&gt;)</td>
<td>(&lt;af&gt;)</td>
<td></td>
</tr>
<tr>
<td>(&lt;b&gt;)</td>
<td>(&lt;bc&gt;)</td>
<td>(&lt;bd&gt;)</td>
<td>(&lt;be&gt;)</td>
<td>(&lt;bf&gt;)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;c&gt;)</td>
<td>(&lt;cd&gt;)</td>
<td>(&lt;ce&gt;)</td>
<td>(&lt;cf&gt;)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;d&gt;)</td>
<td>(&lt;de&gt;)</td>
<td>(&lt;df&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(&lt;e&gt;)</td>
<td>(&lt;ef&gt;)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Sequential Pattern Extraction
Sequential Pattern Extraction

Diagram showing a sequence of patterns and corresponding matrix representation.
Sequential Pattern Extraction

S-extension

\[
\begin{array}{c|c|c|c}
A & T(A) & B & A \rightarrow B \\
1 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 \\
0 & 1 & 1 & 1 \\
0 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
1 & 0 & 1 & 0 \\
0 & 1 & 1 & 1 \\
\end{array}
\]

& =

\[
\begin{array}{c|c|c|c}
A & B & A \rightarrow B \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
\end{array}
\]

I-extension

\[
\begin{array}{c|c|c|c}
A & B & A \rightarrow B \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
\end{array}
\]

& =

\[
\begin{array}{c|c|c|c}
A & B & A \rightarrow B \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
0 & 1 & 0 & 0 \\
1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 \\
0 & 1 & 0 & 0 \\
\end{array}
\]
Bag-of-Pattern Representation

Let $n$ be the size of the pattern dictionary, then each event sequence is represented as an $n$ dimensional vector, with the $i$-th element indicating the number of times of the $i$-th pattern appeared in the sequence.

Maximum interval between pairwise consecutive events: 30 days
Pattern duration: within 90 days

Very similar to the bag-of-words representation of documents.