CMPSC382
Lecture 36: High-Dimensional Data

Prof. John Wenskovitch
11/30/2015
(adapted from Harvard CS171, Alexander Lex)
Last Time

• Methods for Comparing Data Values
  – Bar Charts
  – Line/Dot Plots
High-Dimensional Data

• Tabular data, containing:
  – **Records** – rows, individual data points
  – **Dimensions** – columns, variables for each data point
  – Frequently more rows than columns

<table>
<thead>
<tr>
<th></th>
<th>Age</th>
<th>Gender</th>
<th>Height</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bob</td>
<td>25</td>
<td>M</td>
<td>181</td>
</tr>
<tr>
<td>Alice</td>
<td>22</td>
<td>F</td>
<td>185</td>
</tr>
<tr>
<td>Chris</td>
<td>19</td>
<td>M</td>
<td>175</td>
</tr>
</tbody>
</table>
Parallel Coordinate Plot

• Each vertical axis represents one dimension
• Lines connecting axes represent records
• Good for:
  – All tabular data types
  – Heterogeneous data – records of a different nature in the same dataset; records that can be easily categorized
Parallel Coordinate Plot

Parallel Coordinate Plot

- Primarily shows relationships between adjacent axes
- Limited scalability -- ~50 dimensions, ~1-5000 records
  - Can prevent overdraw issues with line transparency
- Interaction is important!
  - Axis reordering, brushing, filtering
Star Plot

• Similar to parallel coordinates
  – Dimensions radiate from a common origin, rather than being parallel
Small Multiples

- Use multiple views to show different partitions of a dataset
Scatterplot Matrix (SPLOM)

- Matrix of size $d \times d$
- Each row/column represents one dimension
- Each cell plots a scatterplot of row dimension vs. column dimension
- Limited scalability – $\sim 20$ dimensions, $\sim 500-1K$ records
- Brushing is important
Scatterplot Matrix

Combining PCs and SPLOMs

[Claessen & van Wijk 2012]
Connected Charts

[Viau and McGuffin 2012]
Pixel-Based Displays

• Each cell is a “pixel,” value is encoded in color/value
• Meaning derived from ordering
• If no ordering is inherent, clustering is used
• Very scalable – 1 pixel per item
• Good for **homogeneous data** – records of the same nature in the same dataset
Pixel-Based Displays

[Image: Gehlenborg & Wong 2012]
Glyphs

- **Glyph** – A simple image that encodes multiple dimensions
  - One image per record
Chernoff Faces

- Encode size of face, curvature of face, position of eyes, length of nose, position of mouth
Complex Glyphs for Bio Workflows

[Maguire 2012]
Techniques for Reducing Records

• **Sampling**
  – Don’t show every element, show a (random) subset
  – Efficient for large datasets
  – Apply only for display purposes
  – Need to consider the affects of outliers

• **Filtering**
  – Define criteria to remove data (minimum variability, greater/less than a specific value)
  – Can be interactive, combined with sampling
Techniques for Reducing Dimensions

- **Dimensionality reduction** – reduce high-dimensional data to lower-dimensional space
- Preserve as much variation as possible
- Plot into a lower-dimension space
- **Principal Component Analysis** – linear mapping, by order of variance
Any Questions?