Part I

Graphical exploratory data analysis

Graphical summaries of data

Exploratory data analysis
- A preliminary exploration of the data to better understand its characteristics.
- Motivations:
  - Helping to select the right tool for preprocessing or analysis.
  - Making use of humans’ abilities to recognize patterns.
- Pioneered by John Tukey, one of the giants of 20th century statistics.
- Our focus
  - Visual summary statistics
  - Quantitative summary statistics
  - Extraction of data slices

Visualization
- Visualization is the conversion of data into a visual or tabular format so that the characteristics of the data and the relationships among data items or attributes can be analyzed or reported.
- Visualization of data is one of the most powerful and appealing techniques for data exploration.
- Humans have a well developed ability to analyze large amounts of information that is presented visually.
- Goals: to detect general patterns and trends and/or to detect outliers and unusual patterns.
Stem and leaf plot

The decimal point is at the |
1  | 012233334444444444
2  |
3  | 033
3  | 5567899
4  | 000011122234444
4  | 55555566677778889999
5  | 000011111123344
5  | 55566666677788899
6  | 0011134
6  | 6779

Graphical summaries of data

Histogram

- Usually shows the distribution of values of a single variable.
- Divide the values into bins and show a bar plot of the number of objects in each bin.
- The height of each bar indicates the number of objects if all bins are of same width.
- If bins are of different width, then often it is the *area* of the bar that indicates the number of objects in that bin.
- Shape of histogram depends on the number of bins.
Graphical summaries of data

Density

- A smoothed / infinitesimal version of histogram.
- The height of the curve, \( f(x) \) is proportional to probability a value falls in interval \([x, x + dx]\).
- A density is non-negative, \( f(x) \geq 0 \) and integrates to 1

\[
\int_{-\infty}^{\infty} f(x) \, dx = 1 = 100\%
\]

- The integral over any interval \([a, b]\) is the frequency / probability a value falls in this interval.

Estimated density of petal.length within each iris type

The standard normal density

The area between \( z = -0.7 \) and \( z = 0.7 \) is 51.61%

Graphical summaries of data

Density estimate

- Simplest way to estimate a density of a sample \( \{x_1, \ldots, x_n\} \) is to use a kernel density estimator
- Definition

\[
\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} \cdot \phi((x - x_i)/h)
\]

where \( h \) is the width of each bump attached to each sample point.
- R chooses \( h \) automatically.
**Graphical summaries of data**

**Distribution function**
- Associated to a density $f$ is its distribution function

$$F(x) = \int_{-\infty}^{x} f(u) \, du$$

with $0 \leq F(x) \leq 1$.
- Given a sample $\{x_1, \ldots, x_n\}$ we define its ECDF (Empirical Cumulative Distribution Function) as

$$\hat{F}(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{(-\infty,x]}(x_i)$$

where

$$1_{(-\infty,x]}(x_i) = \begin{cases} 1 & -\infty < x_i \leq x \\ 0 & x_i > x. \end{cases}$$

**ECDF of petal.length for each iris type**
Graphical summaries of data

Quantile function

- The inverse of a distribution function

\[ F(Q(p)) = p \]

with \( 0 \leq F(x) \leq 1 \).

- For a sample \( \{x_1, \ldots, x_n\} \) an estimated quantile \( \hat{Q}(p) \) is chosen so that approximately \( np \) of the data points are less than \( \hat{Q}(p) \).

Boxplot

Quantile plots of petal.length for each iris type

Boxplot of all variables across iris types
Boxplot of petal.length for each iris type

Contour of 2D density estimate

Heatmap of 2D density estimate

Correlation (case × case) matrix of iris data
Part II

Multidimensional scaling

Classical multidimensional scaling

- Given a similarity matrix $A$, recall the relationship between a similarity and “distance”
  \[ D_{ij} = (A_{ii} - 2A_{ij} + A_{jj})^{1/2} \]
- Now, consider the matrix $B_{ij}$ with entries
  \[ B_{ij} = -\frac{1}{2}D_{ij}^2. \]
- Finally, consider the matrix
  \[ C = HBH. \]

A visual tool

- Recall the PCA scores were
  \[ \tilde{X}V = U\Delta \]
  where $\tilde{X} = HXS^{-1/2} = U\Delta V^T$.
- Above, $U$ are eigenvectors of
  \[ \tilde{X}\tilde{X}^T = HXS^{-1}X^TH \]
- Also, $XS^{-1}X^T$ is a measure of similarity between cases.

Classical multidimensional scaling

- If $A = XS^{-1}X^T$ is a similarity matrix for a (scaled) data matrix $X$.
- Then,
  \[ B = A - \frac{1}{2}(\nu^T + \nu^T) \]
  \[ = XS^{-1}X^T - \frac{1}{2}(\nu^T + \nu^T) \]
- Above, $\nu = \text{diag}(A)$.
- Therefore,
  \[ C = HAH = \tilde{X}\tilde{X}^T = U\Delta^2U. \]
- In short, the eigenvectors of $C$ are the PCA scores.
Olympic PCA scores

Olympic MDS scores

MDS vs. PCA

MDS vs. PCA
Multidimensional scaling

Classical multidimensional scaling

- We can form $B, C$ for any dissimilarity matrix $D$.
- The matrix $C$ will be symmetric, so it will be diagonalizable as $C = WW^T$ with $W_{n \times k}$ and $\operatorname{rank}(C) = k$. In the general case the diagonal entries $\Lambda_{ii}$ are not necessarily non-negative.
- This leads to a Euclidean representation

$$\left( W^{1/2} \right)_{n \times k}$$

with each row of $W^{1/2}$ being a point in Euclidean space.
- Taking the first two columns of $W^{1/2}$ gives a two-dimensional representation.

Distances between US capitals

If the dissimilarity is Euclidean and the points lie in a 2-dimensional plane in $\mathbb{R}^n$, then the interpoint distances will be identical...
Map of U.S. capitals

Iris PCA scores

MDS of Iris data, $\ell_2$

MDS of Iris data, $\ell_1$
MDS of Iris data, $\ell_\infty$ / sup norm

MDS of Iris data, $\ell_{20}$

MDS of Iris data, $\ell_{0.2}$

Other applications

**Manifold learning**
- Goal of manifold learning is to find “low-dimensional” representation $s$ of $X$ that are not necessarily linear.
- Examples of techniques:
  - ISOMAP
  - Laplacian eigenmaps / diffusion geometry
  - Local linear embedding
Manifold learning

Graph distance

- In the “bottleneck”, we already have a metric (Euclidean distance), so we can use MDS.
- But we might want to emphasize the fact that the groups are “barely connected”...
- How can we emphasize this bottleneck in terms of a distance?
- ISOMAP and Diffusion Geometry method does this by creating a graph and creating a new metric based on the graph.

Graph distances

- In our “bottleneck” picture, let’s connect $k$ mutual nearest neighbours to form a graph $G_k$
  - That is, insert an edge $(i,j)$ if either
    - $i$ is within $j$’s $k$ nearest neighbours;
    - $j$ is within $i$’s $k$ nearest neighbours.
- Let
  
  $$D_{G_k}(i,j) = \text{length of shortest path between vertices } i \text{ and } j \text{ in } G_k.$$ 
- This is a metric on $G_k$.

ISOMAP

- Shove $D_{G_k}$ through classical MDS.
- ISOMAP can “flatten” data if there is a low-dimensional Euclidean configuration where the Euclidean distances well approximate the graph distances $D_{G_k}$.
- This means that the data has to be “flat” if ISOMAP is to recover it.
Part III

Multidimensional arrays (data cubes)

Multidimensional arrays

It is sometimes useful to summarize data into a multidimensional array, with one axis per “category.”

In the unemployment data, we might summarize the results by tuples (state, period, variable)

In the Iris data, we might summarize by (petal.width, petal.length, iris.type)

In order to summarize by petal.width, petal.length we might form categories: low, medium, high.

Iris grouped by petal.width, petal.length, iris.type

<table>
<thead>
<tr>
<th>Petal Length</th>
<th>Petal Width</th>
<th>Species Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low</td>
<td>Setosa</td>
<td>46</td>
</tr>
<tr>
<td>low</td>
<td>medium</td>
<td>Setosa</td>
<td>2</td>
</tr>
<tr>
<td>medium</td>
<td>low</td>
<td>Setosa</td>
<td>2</td>
</tr>
<tr>
<td>medium</td>
<td>medium</td>
<td>Versicolour</td>
<td>43</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>Versicolour</td>
<td>3</td>
</tr>
<tr>
<td>medium</td>
<td>high</td>
<td>Virginica</td>
<td>3</td>
</tr>
<tr>
<td>high</td>
<td>medium</td>
<td>Versicolour</td>
<td>2</td>
</tr>
<tr>
<td>high</td>
<td>medium</td>
<td>Virginica</td>
<td>3</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>Versicolour</td>
<td>2</td>
</tr>
<tr>
<td>high</td>
<td>high</td>
<td>Virginica</td>
<td>44</td>
</tr>
</tbody>
</table>
Iris grouped by petal.width, petal.length, iris.type

Drilling down

Resolution of a data cube

- In the process of forming a data cube for unemployment across states, we had already aggregated over county.
- We could make another cube with county-level resolution. This would be *drilling down*.
- The operation of going from county-level to state level is *rolling up*.
- Such operations are often best handled by database tools rather than *R*, but *R* does support multidimensional arrays.