COMP20008 Elements of Data Processing

Semester 2 2018

Lecture 6: Data Visualisation

What we've covered so far!

Finished:

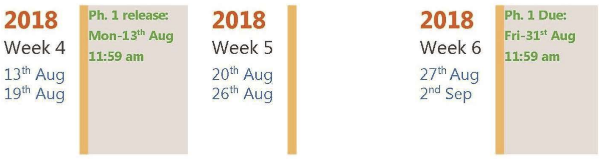
- Lecture 1: Introduction
- Lectures 2-3: Data formats: structured, unstructured and semi-structured
- Lectures 4-5: Data preprocessing and cleaning: missing values, outlier detection and recommender systems

Next:

- Lecture 6: Some basic visualisation methods
  - Scatter plots, heat maps, parallel co-ordinates

Announcements

- Key dates: Week 4-6



2018
Week 4
13th Aug
19th Aug

Ph. 1 release:
Mon-13th Aug
11:59 am

2018
Week 5
20th Aug
26th Aug

2018
Week 6
27th Aug
2nd Sep

Ph. 1 Due:
Fri-31st Aug
11:59 am

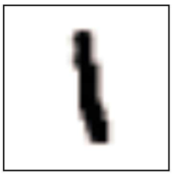
- Assignment consultation sessions will be released end of next week

Plan today

- Complete section of collaborative filtering
  - Item item similarity
  - Matrix factorisation
- Some basic visualisation methods
  - Scatter plots, heat maps, parallel co-ordinates

Motivation for visualisation

- Converting data into a visual format
  - Reveals characteristics of the data, relationships between objects or relationships between features
  - Simplifies the data
- Humans are very good at analysing information in a visual format
  - Spot trends, patterns, outliers
  - Visualisation can help show data quality
- Visualisation helps tell a story ….



$$\simeq \begin{bmatrix} 0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&.6&.8&0&0&0&0&0&0\\ 0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&.7&1&0&0&0&0&0&0\\ 0&0&0&0&0&0&.5&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.4&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&.7&0&0&0&0&0\\ 0&0&0&0&0&0&0&1&1&0&0&0&0&0\\ 0&0&0&0&0&0&0&.9&1&.1&0&0&0&0\\ 0&0&0&0&0&0&0&.3&1&.1&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0\\ 0&0&0&0&0&0&0&0&0&0&0&0&0&0 \end{bmatrix}$$

Image is taken from:
https://www.tensorflow.org/versions/r1.0/get_started/mnist/beginners

Visualisations we have already encountered

- Boxplots
  - Median, quartiles, outliers

- Scatter plots
  - Plotting points in 2D or 3D space, using colours to indicate classes/segments

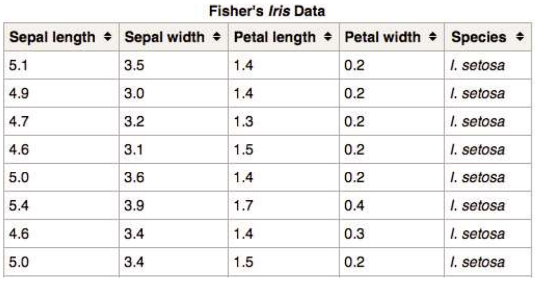
Running Example: Iris Flower Dataset

- Well known dataset introduced by statistician Ronald Fisher with 150 objects
  - https://en.wikipedia.org/wiki/Iris_flower_data_set
- Three flower types (classes):
  - Setosa
  - Virginica
  - Versicolour
- Four features
  - Sepal width and length
  - Petal width and length

Virginica. Robert H. Mohlenbrock. USDA NRCS. 1995. Northeast wetland flora: Field office guide to plant species. Northeast National Technical Center, Chester, PA. Courtesy of USDA NRCS Wetland Science Institute.

Iris dataset running example

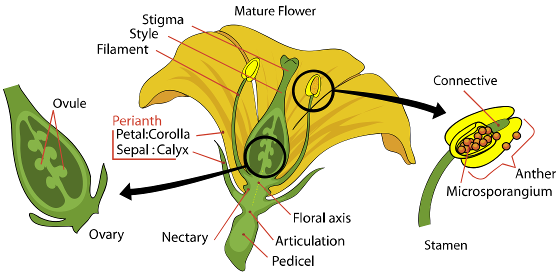
- Extract of Iris data from Wikipedia



Fisher's *Iris* Data

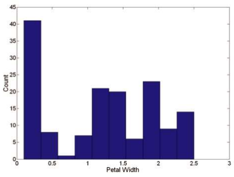
Sepal length ⬍	Sepal width ⬍	Petal length ⬍	Petal width ⬍	Species ⬍
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

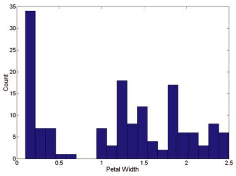
Flower diagram: https://en.wikipedia.org/wiki/Sepal



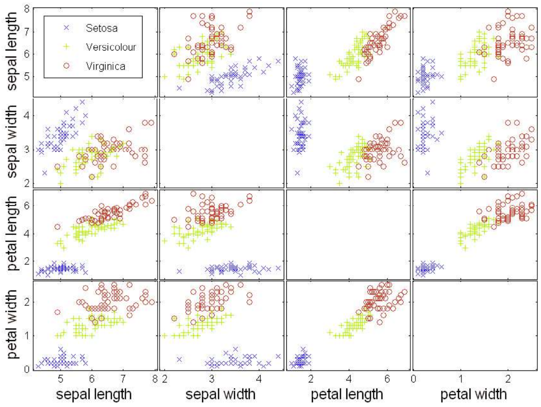
Basic Visualisations: Histograms

- 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
  - Shape of histogram depends on the number of bins
- Example: Petal Width (10 and 20 bins, respectively)





Basic Visualisations: Scatter plots

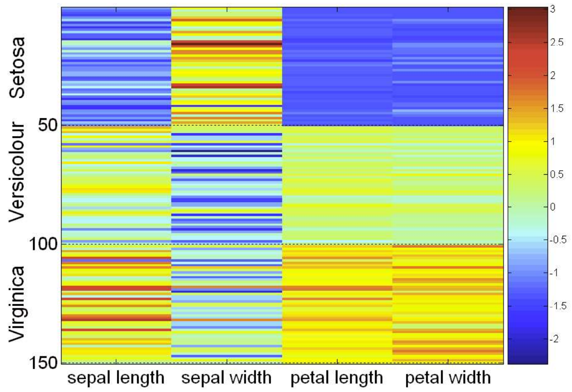


Scatter plots for iris dataset

Heat maps

- Heat maps
  - Plot the data matrix
  - This can be useful when objects are sorted according to class
  - Typically, features are normalized to prevent one attribute from dominating the plot

Visualization of the (normalised) Iris Data Matrix



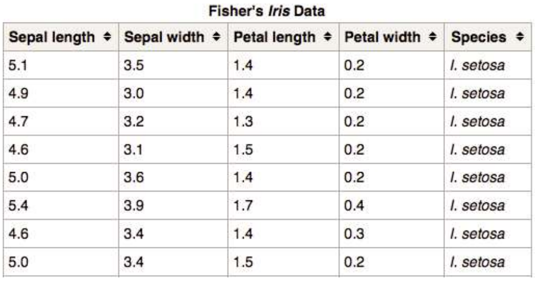
[Columns have been standardized to have a mean of zero and standard deviation of 1]

Parallel Coordinates

- Parallel Coordinates
  - Used to plot the feature values of high-dimensional data
  - Instead of using perpendicular axes, use a set of parallel axes
  - The feature values of each object are plotted as a point on each corresponding coordinate axis and the points are connected by a line
  - Thus, each object is represented as a line
  - Often, the lines representing a distinct class of objects group together, at least for some features
  - Ordering of attributes is important in seeing such groupings

Iris dataset

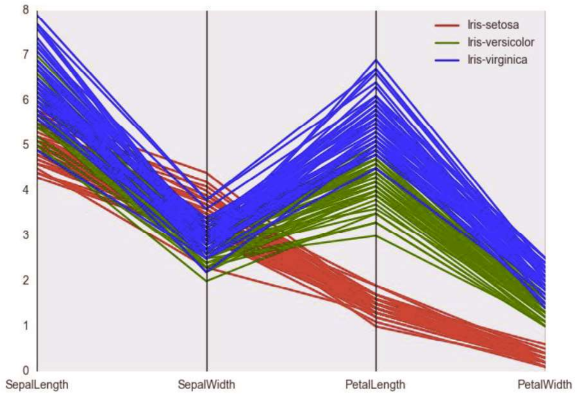
- Extract of Iris data from Wikipedia



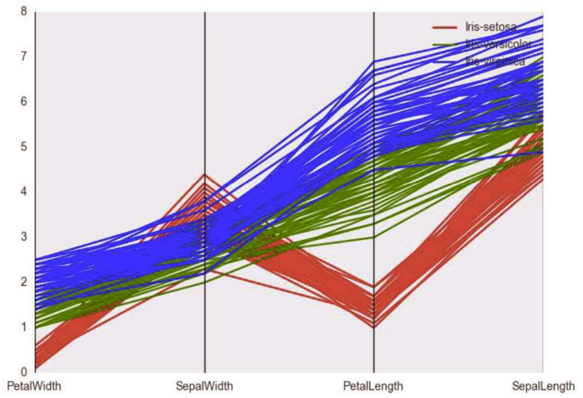
Fisher's *Iris* Data

Sepal length ⬍	Sepal width ⬍	Petal length ⬍	Petal width ⬍	Species ⬍
5.1	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
5.0	3.6	1.4	0.2	*I. setosa*
5.4	3.9	1.7	0.4	*I. setosa*
4.6	3.4	1.4	0.3	*I. setosa*
5.0	3.4	1.5	0.2	*I. setosa*

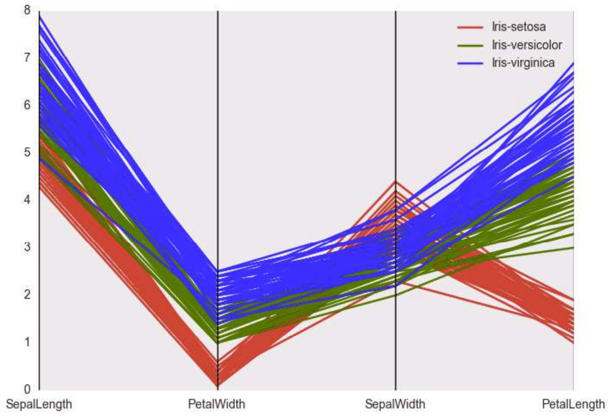
Parallel coordinates: ordering 1



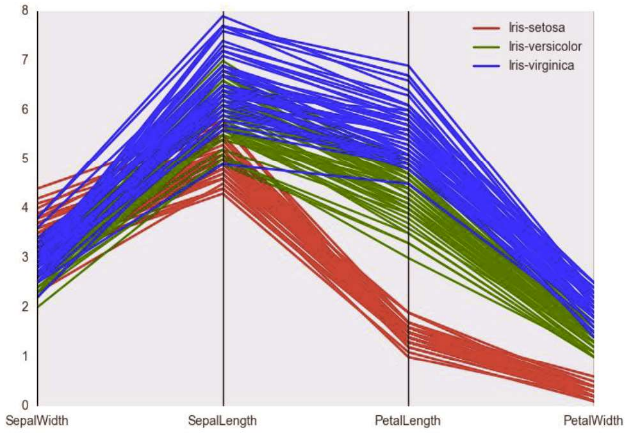
Parallel coordinates: ordering 2



Parallel coordinates: ordering 3



Parallel coordinates: ordering 4



Key issues in using parallel coordinates

- Scaling axes
  - Affects the visualisation. May choose to scale all features into the range [0,1] via a pre-processing step
- Ordering of axes
  - Influences the relationships that can be seen. Correlations between pairs of features may only be visible in certain orderings

Parallel co-ordinates code

- Python code
  - *parallel_coordinates* in *pandas.tools.plotting*
  - Will practice in workshop

Acknowledgements

- Material partly adapted from
  - “Data Mining Concepts and Techniques”, Han et al, 2nd edition 2006.
  - “Introduction to Data Mining”, Tan et al 2005.