Data Analytics

IS5102, Lecture 19

22 November 2021

Alexander Konovalov alexander.konovalov@st-andrews.ac.uk (mailto:alexander.konovalov@st-andrews.ac.uk)

Thanks to: Vinodh Rajan and Susmit Sarkar

NoSQL

- Umbrella term for graph, key-value, document (and other) non-relational DBMS
- Emphasise different query models
- Analytics driving many of these models

Overview

- Introduction
- Steps
- Operations
- Techniques
- Handling multivariate data

Data Mining: What's in a name?

- Data Mining
- KDD (Knowledge Discovery and Data)
- Data Science
- Data Analytics
- Machine Learning

- The difference between these are very fuzzy and there is indeed a lot of overlap
- You can give proper individual text-book definitions for each of these.
- But practically, they more or less aim at the same thing

Data Mining: What is it?

- The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions, (Simoudis,1996)
- Involves the analysis of data and the use of software techniques for finding hidden and unexpected patterns and relationships in sets of data
- Reveals information that is hidden and unexpected, as little value in finding patterns and relationships that are already intuitive
- Patterns and relationships are identified by examining the underlying rules and features in the data

Data Mining: Why we need it?

- Existence of large of amount of data waiting to be analyzed
- The ability to harness past experience/decisions to shape future decisions
- The obsessive need to predict something
- Discover patterns/relationships that are not very obvious
- Ability to model user/market/anything behavior

Types of Data

- Qualitative Data
  - Describes “attributes” ”labels” “properties” “category” etc.
  - For instance Weather descriptors
    - “foggy”, “misty”, “rainy”, “cold”, “sultry”, “dreich” so on and so forth
  - Sometimes numbers can also be used as labels
    - Likert Scale (1 – Strong Disagree to 5 – Strong Agree)
- Quantitative Data
  - Continuous Data
    - These can take fractional value such as temperature, distance etc,
  - Discrete Data
    - Can only take whole numbers such as number of students

Data Mining: Steps

- Data Cleaning
- Data Integration
- Data Transformation/Normalization
- Data Reduction/Selection

Data Mining: Steps: Data Cleaning

- Real data are rarely clean
  - They are noisy: Smoothen your data, remove outliers
  - They are incomplete: Fill in the missing values/ignore them
  - They are inconsistent: Ensure consistency
- You must make sure the data you work with is clean
- Remember: Garbage in, garbage out
  - This is true for data as well.

Data Mining: Steps: Data Integration

- Your data is probably split across multiple databases/files
- They are also most probably heterogeneous
- They must be integrated properly into a coherent entity to which you can apply data mining operations
- This is also means (again) ensuring consistency between different entities

Data Mining: Steps: Data Transformation/Normalization

- More often that not data is not in a directly usable form
- You may have to apply transformation/aggregations to make sensible use of data
  - You cannot compare prices from 1800's to 2000's without accounting for inflation
- You may also want to normalize data if the attributes are scale dependent to transform to the same scale
  - You may have two attributes for rating, where one is scored from 1 to 5 and the other is scored from 1 to 10

Data Mining: Steps: Data Reduction/Selection

- You probably have to deal with a huge number of attributes in your data
- At least some of them are probably useless
- So you need to select the useful attributes and disregard the others
- In some cases, you may want to reduce the number of attributes by combining some of them

Data Mining: Operations

- Predictive Modeling
- Link Analysis
- Deviation Detection

Data Mining: Techniques

- Techniques are specific implementations of the data mining operations
- Each operation has its own strengths and weaknesses

Data Mining: Techniques

- Predictive modeling techniques:
  - classification
  - value prediction
- Link analysis techniques:
  - association discovery
  - sequential pattern discovery
  - similar time sequence discovery
- Deviation detection techniques:
  - statistical analysis
  - visualisation

Predictive Modeling

- Similar to the human learning experience

- uses observations to form a model of the important characteristics of some phenomenon
- Uses generalizations of ‘real world’ and ability to fit new data into a general framework

Predictive Modeling

- Model is developed using a supervised learning approach, which has two phases: training and testing
  - Training builds a model using a large sample of historical data called a training set
  - Testing involves trying out the model on new, previously unseen data to determine its accuracy and physical performance characteristics

Predictive Modeling

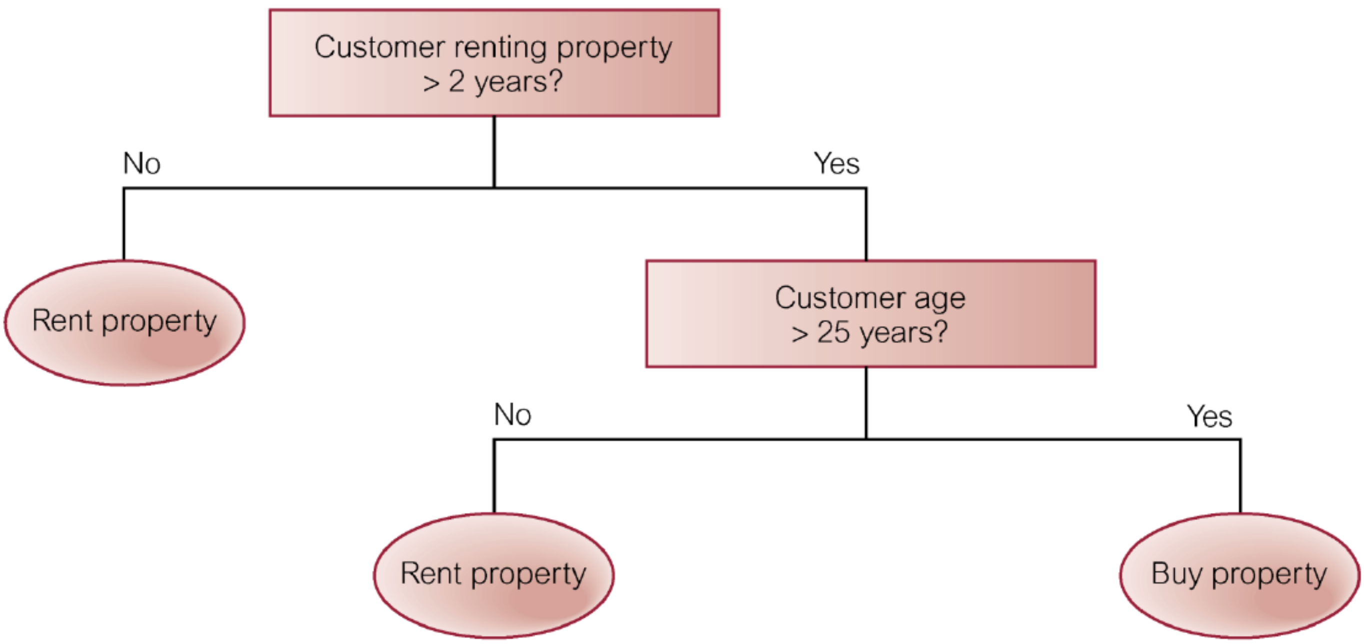
$$X \rightarrow f(X) \rightarrow Y$$

- $f(X)$ is the function approximation that is constructed to map a given X to Y
- We attempt to learn $f(X)$ through a training set of known (X, Y) pairs
- We use this to predict an unseen X, based on the above mapping

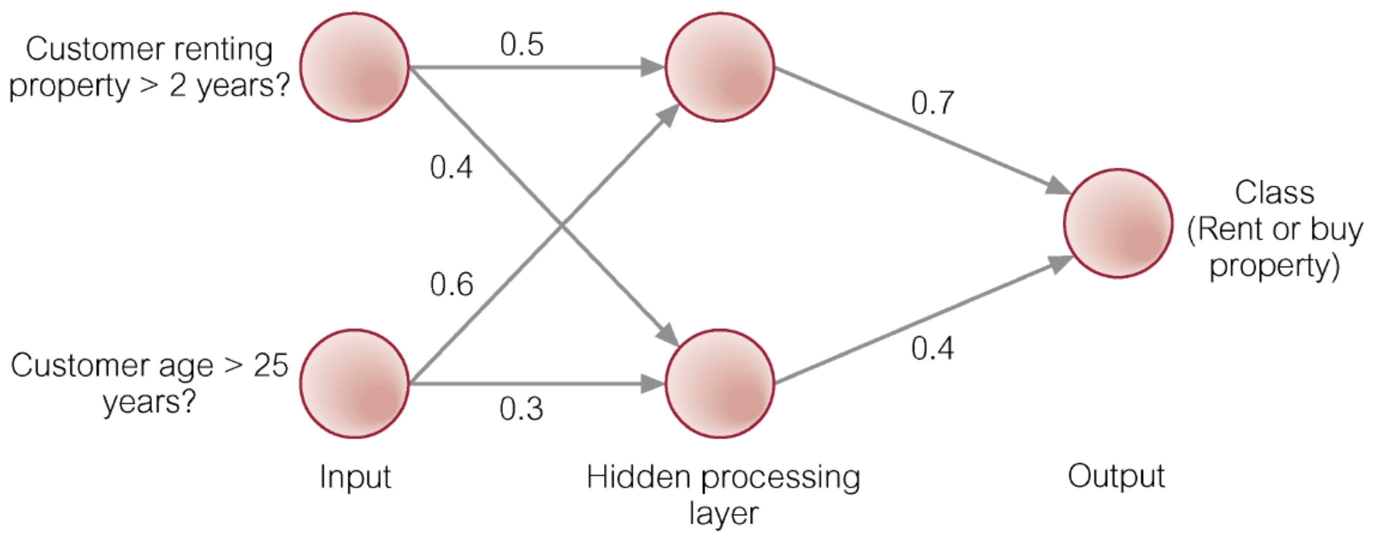
Predictive Modeling: Classification

- Used to establish a specific predetermined class for each record in a database from a finite set of possible, class values
- Two specializations of classification: tree induction and neural induction

Induction tree



Neural network



Predictive Modeling: Value Prediction

- Used to estimate a continuous numeric value that is associated with a database record
- Uses the traditional statistical techniques
- Relatively easy-to-use and understand

Regression

- Explaining one variable in terms of another
- It models the relationship between two variables
- y the dependent variable is modelled in terms of x, which is the independent variable
- It is used to predict/forecast or interpolate y given x

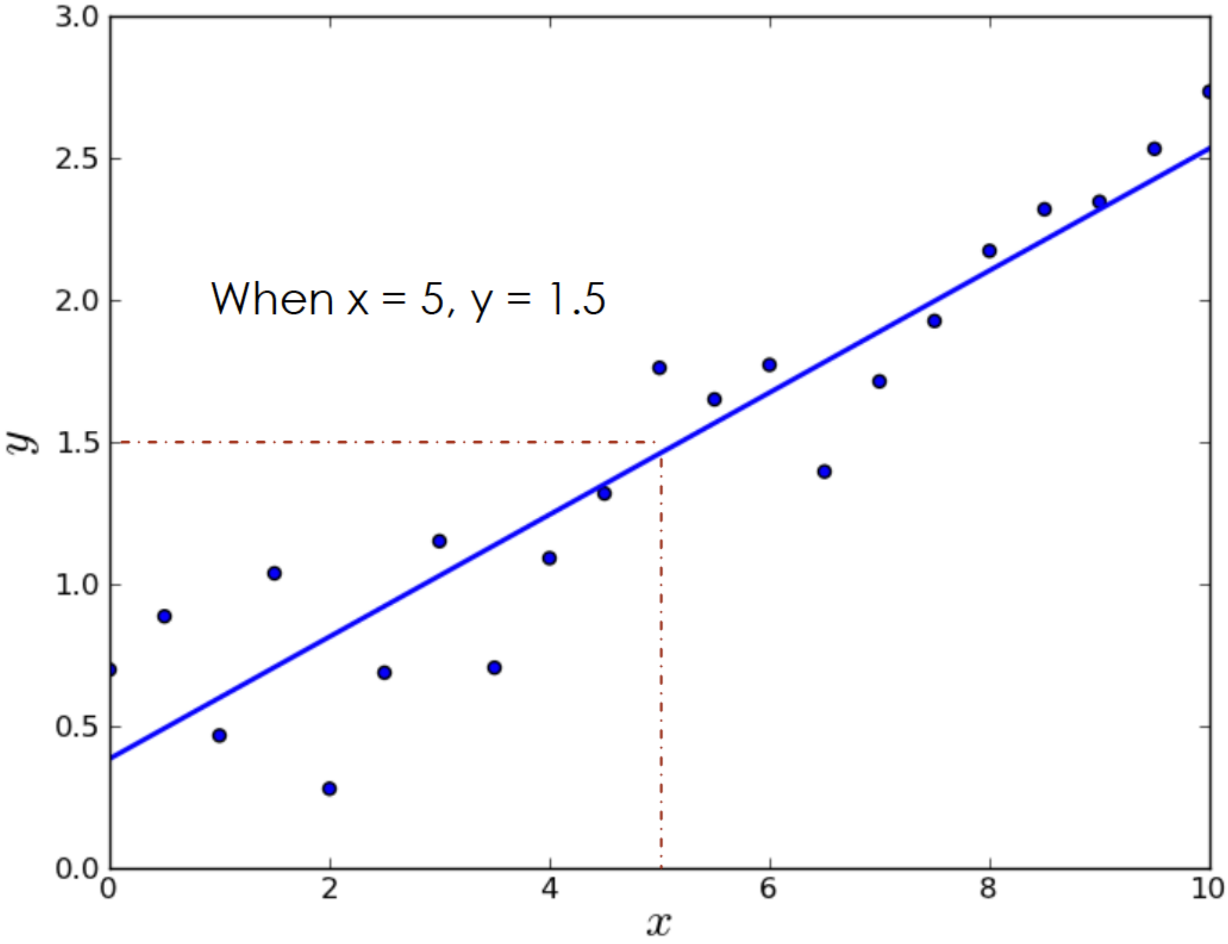
Simple Linear Regression

- Linear regression fits the data in the form of a straight line:
  - $Y = aX + b$
  - b = intercept on the y-axis
  - a = slope
  - The fit is measured in terms of R^2
  - Called the co-efficient of determination
  - $R^2 = 1$ means a perfect fit
  - Shows how much of variation in Y is explained in terms of X

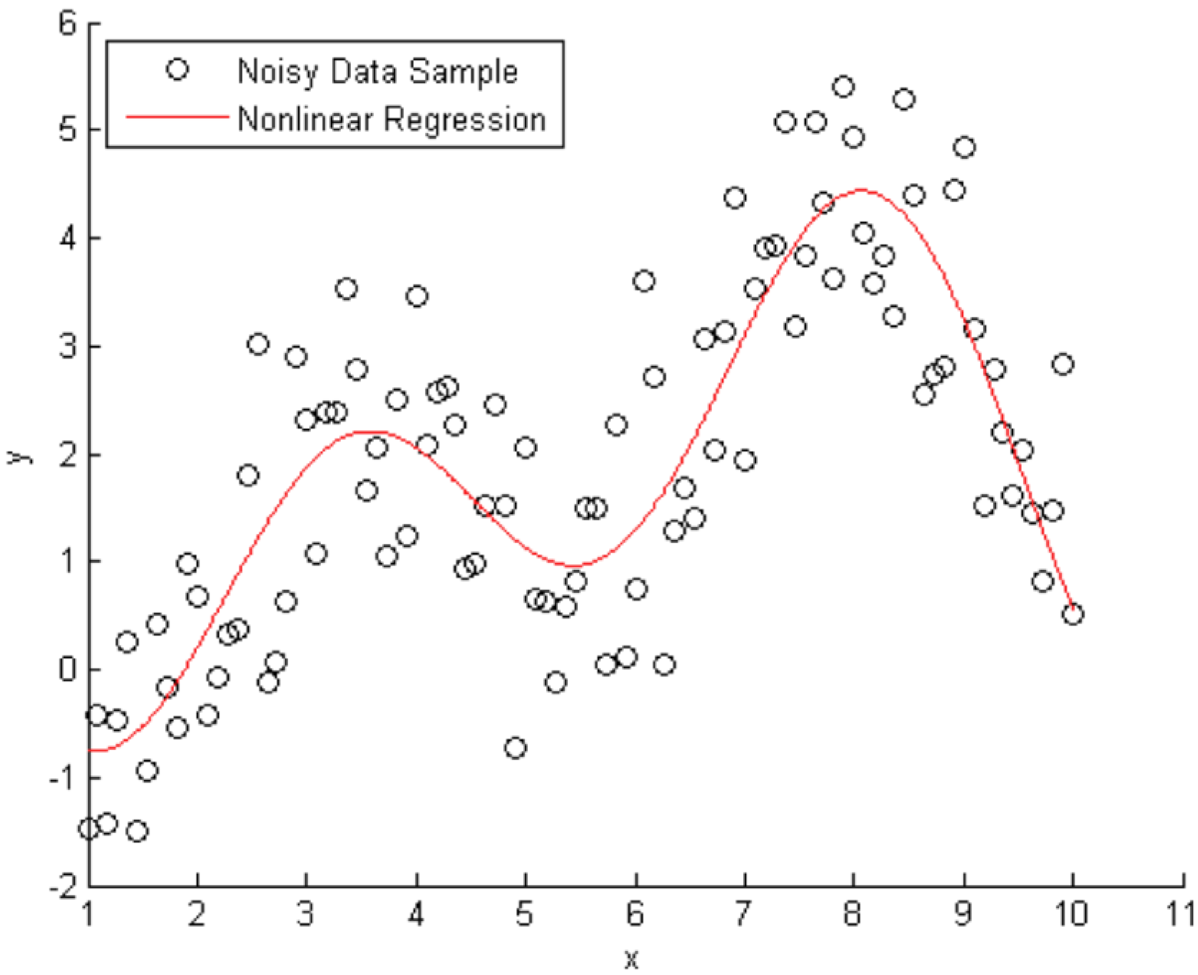
Simple Linear Regression

- The relationship is not always linear
- If a linear model doesn't fit properly, try using a nonlinear model to fit the data
- Do not extrapolate/predict what would happen outside the range of the data. We just don't know

Linear regression



Non-linear regression

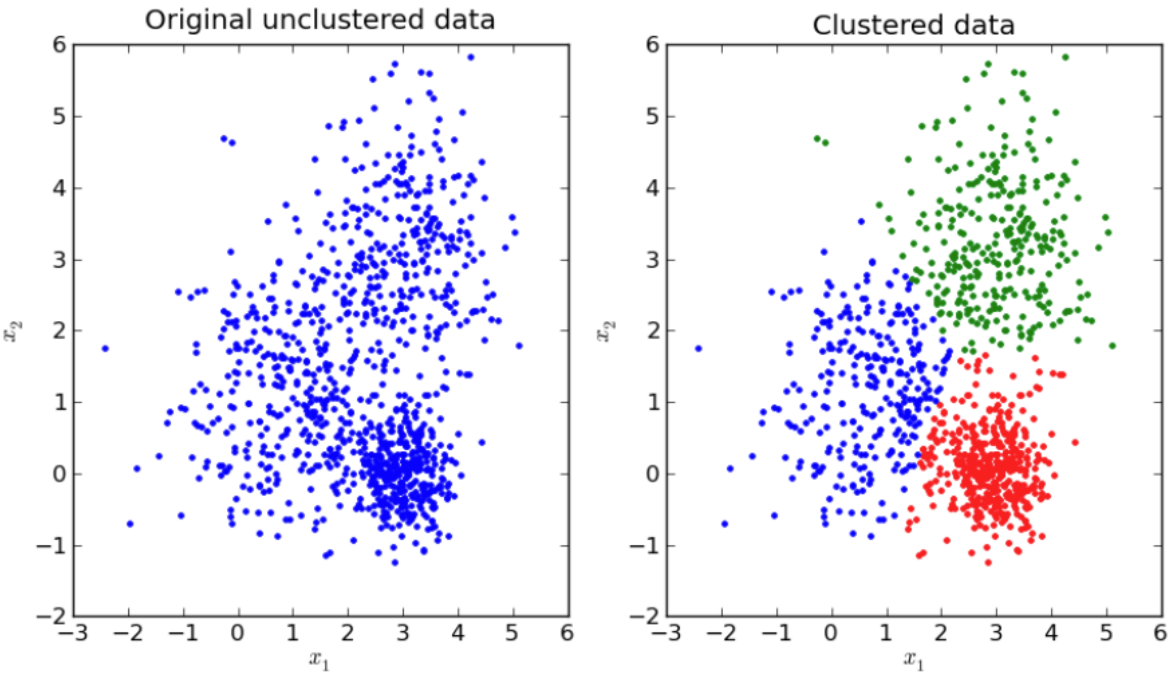


Clustering

- The earlier techniques we saw (predictive modeling) are part of supervised learning
- This requires the existence of a training set. However, very frequently we come across data that we have no prior information (unsupervised learning)
- Clustering allows us to group related data without any prior information on how to do it
- Group related people based on their demographic attributes

Clustering: *kMeans*

- *kMeans* is a very commonly used technique to perform clustering
- Divides n observations into k clusters ($k \leq n$) clustered around a mean



Link Analysis

- Aims to establish links (associations) between records, or sets of records, in a database
- There are three specializations
  - Associations discovery
  - Sequential pattern discovery
  - Similar time sequence discovery
- Applications include product affinity analysis, direct marketing, and stock price movement

Link Analysis: Association Discovery

- Finds items that imply the presence of other items in the same event
- Affinities between items are represented by association rules
  - e.g. 'When a customer rents property for more than 2 years and is more than 25 years old, in 40% of cases, the customer will buy a property. This association happens in 35% of all customers who rent properties'

Link Analysis: Sequential Pattern Discovery

- Finds patterns between events such that the presence of one set of items is followed by another set of items in a database of events over a period of time
- e.g. used to understand long term customer buying behavior

Link Analysis: Similar Time Sequence Discovery

- Finds links between two sets of data that are timedependent, and is based on the degree of similarity between the patterns that both time series demonstrate

- e.g. Within three months of buying property, new home owners will purchase goods such as cookers, freezers, and washing machines

Deviation Detection

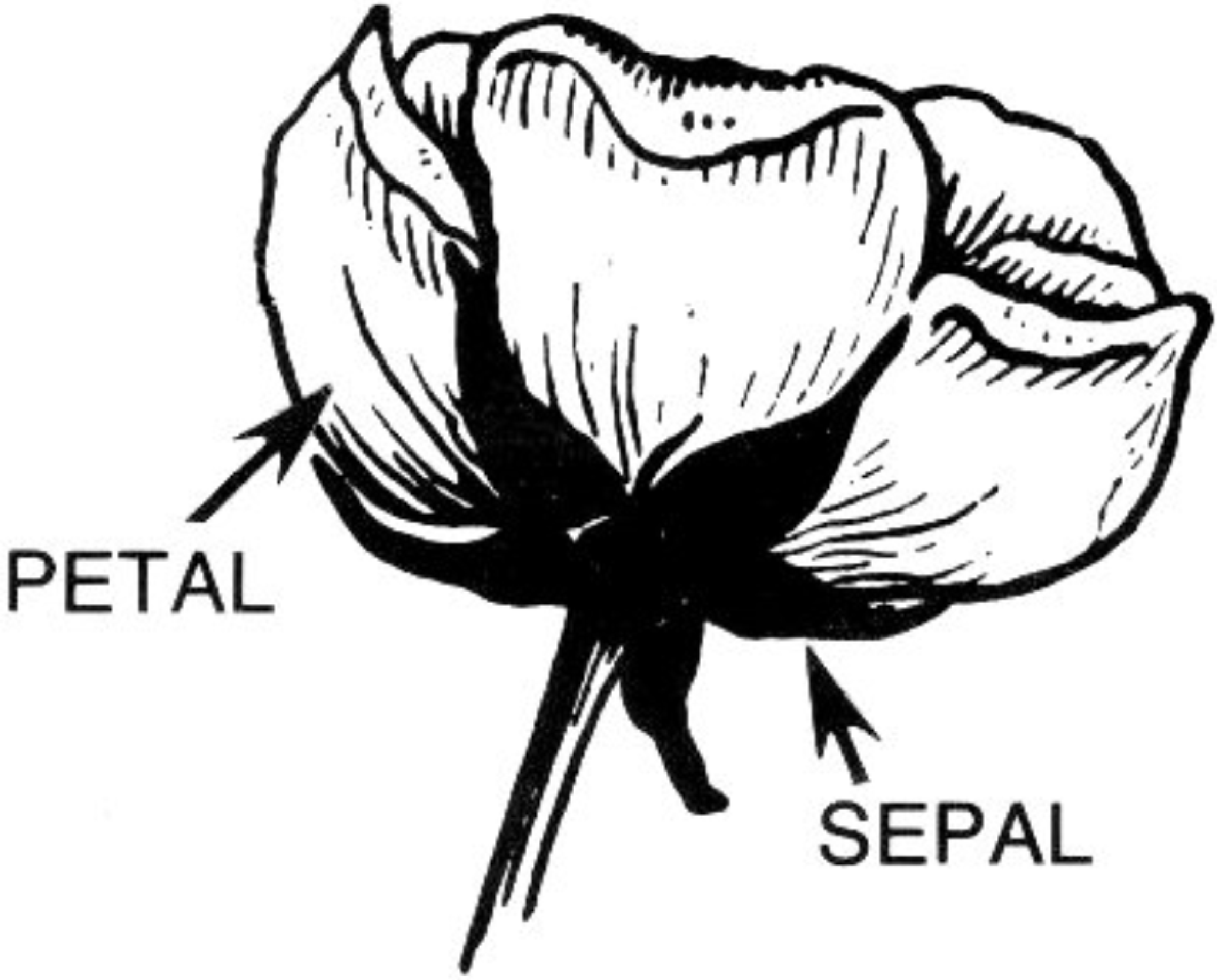
- Can be performed using statistics and visualization techniques or as a by-product of data mining
- Often a source of true discovery because it identifies outliers, which express deviation
- Applications include fraud detection, quality control, and defects tracing

Multivariate Data

- Real datasets consists of multiple variables/attributes in tens or even hundreds
- For instance, house prices may depends on city, neighborhood, area, number of bedrooms and other factors
- The attributes are usually related to each other
- Visualizing and analyzing these data sets is comparatively complex

Multivariate Data: Iris Flower Data Set

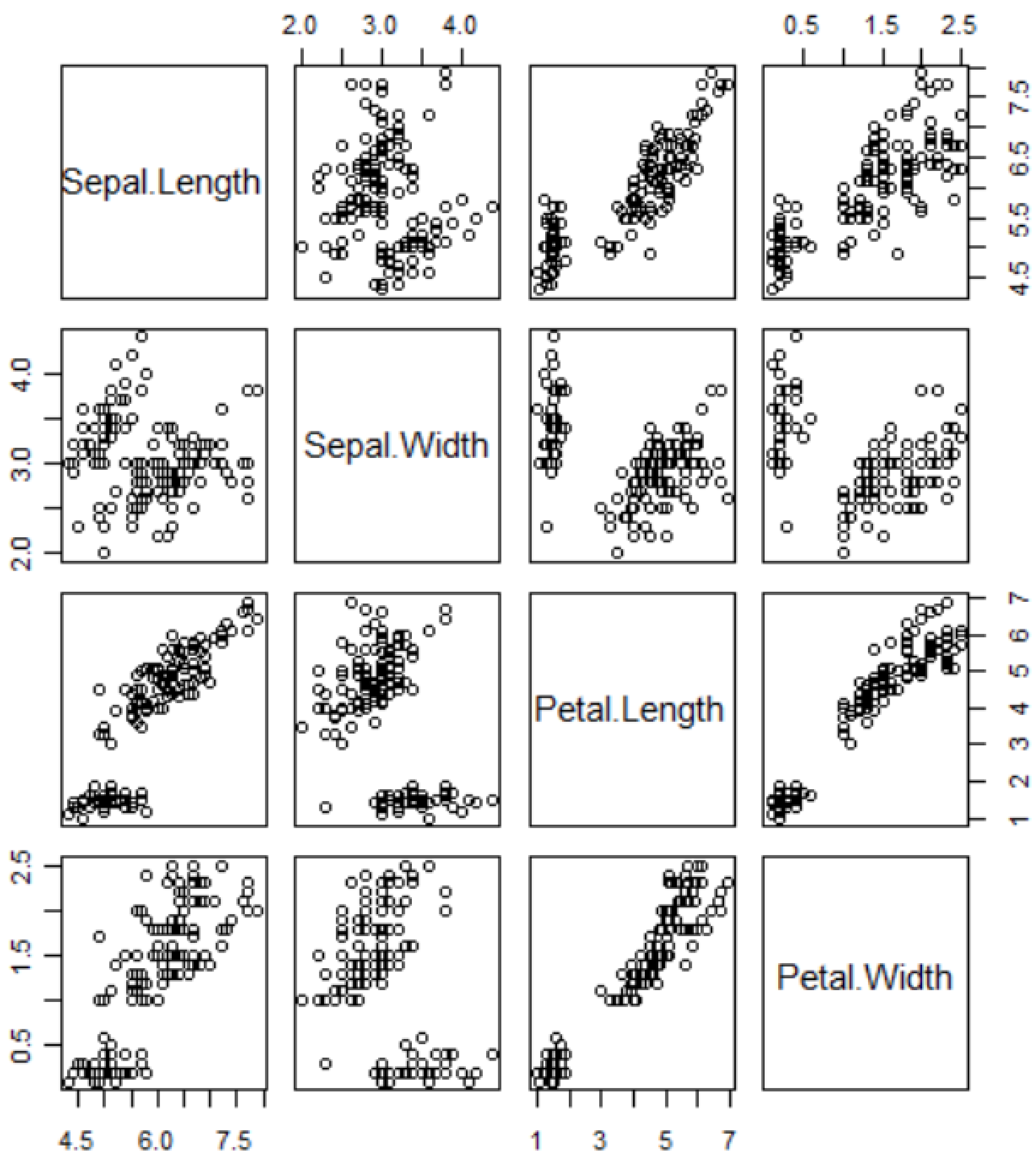
- The iris data set consists of 50 samples from each of three species of Iris (Iris setosa, Iris virginica and Iris versicolor).
- Has 150 samples in total
- Iris dataset has the attributes: Sepal Length (S_L), Sepal Width(S_W), Petal Length (P_L), Petal Width (P_W)



Visualizing Multivariate Data: Scatterplot Matrices

- 3D visualization is difficult to interpret by humans
- Visualizing more than 3 dimensions at the same becomes exponentially very difficult
- The easiest approach is to do a pairwise scatterplot and arrange them in a matrix to see the pairwise interaction between variables

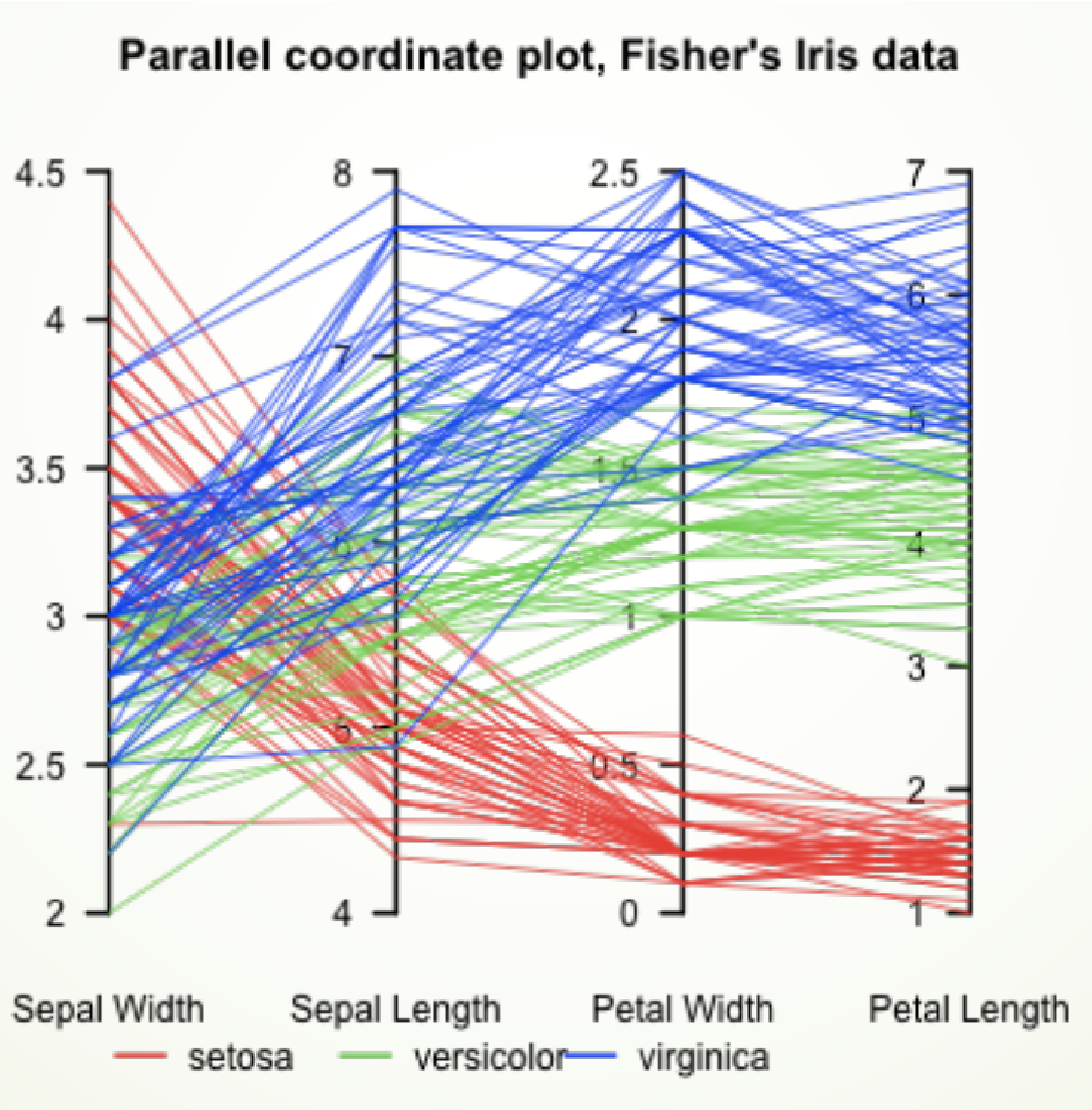
Scatterplot matrix



Visualizing Multivariate Data: Parallel Coordinates Plots

- Allows to visualize all data attributes at the same time
- Consists of a number of axes arranged in parallel
- A data item is a line that moves across all the axes
- Can visualize interactions between multiple attributes at the same time
- Can detect patterns easily

Parallel coordinates plot



Analyzing Multivariate Data: Dimensionality Reduction

- Assuming, your data has 30 attributes – essentially it means an item has 30 dimensions.
  - You'd need 435 scatterplots!
- The analysis can be simplified if we are to reduce the dimensions (a.k.a attributes) to say 5
  - Now, you'd need only 5C2 = 10 scatterplots
- This is done by projecting data items from a higher dimensional space to a lower dimensional space

Dimensionality Reduction: Principal Component Analysis (PCA)

- When there are numerous attributes present, they can be linearly combined to form aggregate attributes called Principal Components (PC)

- PCs helps us to bring out the structure of the data and relationships between the attributes
- Suppose, we have a dataset consisting of attributes:
  - a_1 to a_n
- It's possible to create new attributes PC_1 to PC_m where m is significantly less than n:

$$
\begin{gathered}
PC_1 = c_{11}a_1 + c_{12}a_2 + .. + c_{1n}a_n \\
PC_2 = c_{21}a_1 + c_{22}a_2 + .. + c_{2n}a_n \\
\ldots \\
PC_m = c_{m1}a_1 + c_{m2}a_2 + .. + c_{mn}a_n
\end{gathered}
$$

Dimensionality Reduction: Principal Component Analysis (PCA)

- Example:
  - For instance, for the iris data set we can possibly derive the following Principal Components reducing the set to 2 Dimensions that might look like this:
  - $PC_1 = 0.4S_W + 0.6P_W$ – We could call this new feature (Flower Width)
  - $PC_2 = 0.5S_L + 0.5P_L$ – We could call this new feature (Flower Length)
- Essentially, we have have created two new aggregate attributes that can replace the original 4 attributes

Multivariate Linear Regression

- Similar to simple (bivariate) linear regression, but now the regression is dependent on multiple variables
  - $Y = a + b_1X_1 + b_2X_2 + b_3X_3 ... + b_nX_n$
- You have single dependent variable regressed upon several independent variables
- Can be useful when several factors affect the dependent variable

Multivariate Linear Regression

- Assume the example of the house price:
- Given a relevant dataset it can be possible to create a multiple linear regression model which might look like

$$Price = a + b \times area + c \times bedrooms + d \times floors$$

Conclusions

- Significant investment in collecting/storing relevant data now pays off
- Data analyst jobs are now widespread
- Tools and how to learn them
- Ethical data science

Obligatory XKCD

- Correlation: [https://xkcd.com/552/ (https://xkcd.com/552/)](https://xkcd.com/552/)
- Linear regression: [https://www.xkcd.com/1725/ (https://www.xkcd.com/1725/)](https://www.xkcd.com/1725/)
- Curve fitting: [https://xkcd.com/2048/ (https://xkcd.com/2048/)](https://xkcd.com/2048/)