Data Preparation

Jingpeng Li

Data Mining is a Lot Like Cooking

- You need to know what temperature to set the oven, and how long to leave it in, but you can get away with a lot by choosing a sensible heat and checking occasionally
- However, if you get the ingredients and the preparation wrong, knowing the correct temperature won't save you

And So It Is With Data

- There are many parameters that you can try to optimise when running a data mining algorithm, but a sensible choice and a bit of trial and error will usually produce a good result
- If, however, your data is not appropriate to the task, no amount of parameter tweaking will help.
- Garbage in, garbage out, as they say

Check Points

- Data quantity and quality: do you have sufficient good data for the task?
  - How many variables are there?
  - How complex is the task?
  - Is the data's distribution appropriate?
    - Outliers
    - Balance
    - Value set size

Distributions

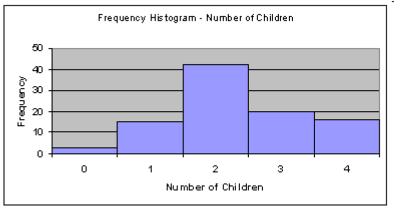
- A frequency distribution is a count of how often each variable contains each value in a data set
- For discrete numbers and categorical values, this is simply a count of each value
- For continuous numbers, the count is of how many values fall into each of a set of sub-ranges

Example Distributions

- Data: 1, 2, 2, 3, 4, 4, 4, 5
- Frequency counts:
  - 1→1, 2→2, 3→1, 4→3, 5→1
- Data: 1.1, 1.2, 2, 3.4, 4.1, 4.2, 4.2, 4.9
- Frequency counts:
  - (1 to <2) → 2
  - (2 to <3) → 1
  - (3 to <4) → 1
  - (4 to <5) → 4

Plotting Distributions

- The easiest way to visualise a distribution is to plot it in a histogram:



- What is the most common number of children represented in the data?

Features of a Distribution to Look For

- Outliers
- Minority values
- Data balance
- Data entry errors

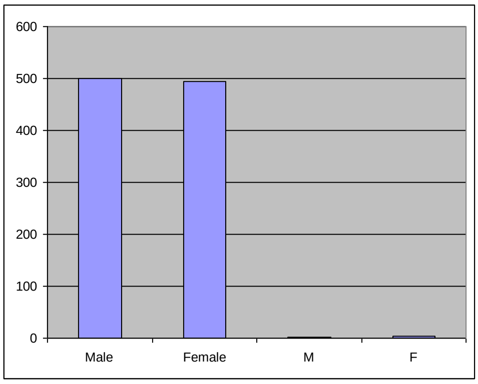
Outliers

- Outliers
  - A small number of values that are much larger or much smaller than all the others
  - Can disrupt the data mining process and give misleading results
  - You should either remove them or, if they are important, collect more data to reflect this aspect of the world you are modelling
  - Could be data entry errors

Minority Values

- Values that only appear infrequently in the data
- Do they appear often enough to contribute to the model?
- Might be worth removing them from the data or collecting more data where they are represented
- Are they needed in the finished system?
- Could they be the result of data entry errors?

Minority Values



What does this chart tell you about the gender variable in a data set?

What should you do before modelling or mining the data?

Flat and Wide Variables

- Variables where all the values are minority values have a flat, wide distribution – one or two of each possible value
- Such variables are of little use in data mining because the goal of DM is to find general patterns from specific data
- No such patterns can exist if each data point is completely different
- Such variables should be excluded from a model

Data Balance

- Imagine I want to predict whether or not a prospective customer will respond to a mailing campaign
- I collect the data, put it into a data mining algorithm, which learns and reports a success rate of 98%
- Sounds good, but when I put a new set of prospects through to see who to mail, what happens?

A Problem

- … the system predicts 'No' for every single prospect.
- With a response rate on a campaign of 2%, then the system is right 98% of the time if it always says 'No'.
- So it never chooses anybody to target in the campaign

A Solution

- One data pre-processing solution is to balance the number of examples of each target class in the output variable
- In our previous example: 50% customers and 50% non-customers
- That way, any gain in accuracy over 50% would certainly be due to patterns in the data, not the prior distribution
- This is not always easy to achieve – you might need to throw away a lot of data to balance the examples, or build several models on balanced subsets
- Not always necessary – if an event is rare because its cause is rare, then the problem won't arise

Data Quantity

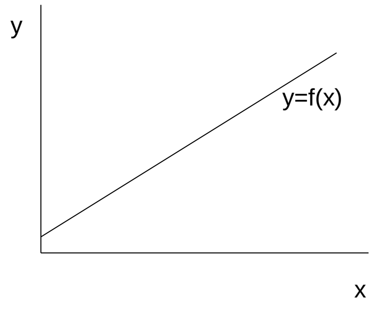
- How much data do you need?
- How long is a piece of string?
- Data must be sufficient to:
  - Represent the dynamics of the system to be modelled
  - Cover all situations likely to be encountered when predictions are needed
  - Compensate for any noise in the data

Linearity

- Two variables have a linear relationship if plotting one against the other on a scatter plot produces a straight line
- Put another way, if a constant change in x leads to a constant change in y, for all values of x and y, then x and y are linearly related

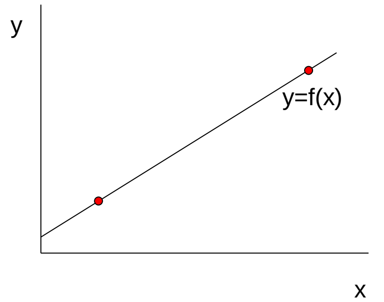
Linearity and Data Quantity

- If you know that x and y are linearly related, how many data points do you need to build a model of that relationship?



Linearity and Data Quantity

- Clearly, two examples are enough if you know the relationship is linear
- This is only true if there is no sampling error

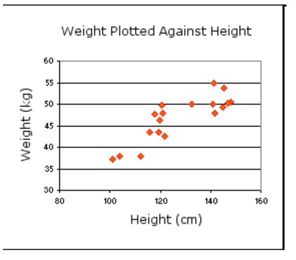


Sampling Theory

- The data that you will use for a data mining project will almost always be a sample taken from a much larger population
- There will be data you couldn't collect, so the true nature of the world that you are trying to capture is represented by the snapshot of data that you have

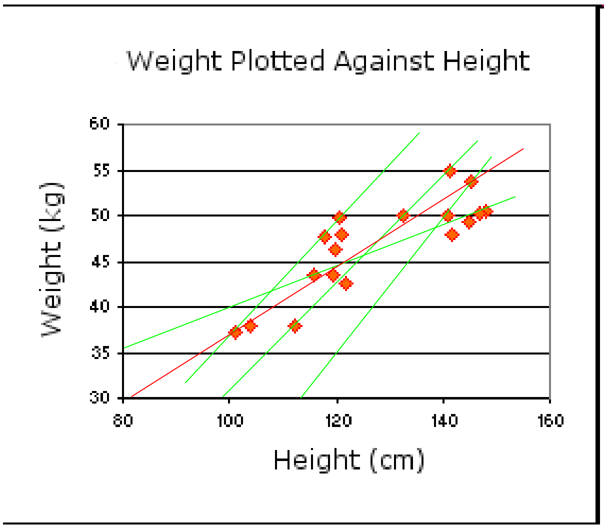
Noise and Variability

- Two variables with a linear relationship might not produce a set of data that lie perfectly on a straight line
- They could lie in a long thin cloud around a straight line:



Variability

- The spread around the line (which is what correlation measures) could be due to either:
  - Imperfect measurements or noise
  - Variability caused by other factors not being measured
  - Simple randomness



A sample of people's height and weight plotted as a scatter plot. The green lines show how modelling from two points can produce widely differing models. The red line is the correct regression line for the data.

The Need For More Data

- So two points are no longer enough, even for a linear relationship if noise or other variability is present
- The green lines on the previous slide show some potential models of the data if only two points are used
- The red line is the correct model

Finding The Right Line

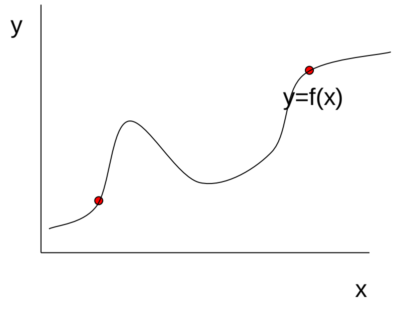
- The correct way to draw a straight line through this data is to find one that minimises the distance between all the points and the line
- This distance is known as the ‘error’ of the model and is usually calculated as the average of the squared errors
- Known as MSE – mean squared error

Learning

- The process of learning is the process of minimising the MSE
- This can be done in a number of ways:
  - Linear regression equation solving
  - Iterative search
    - Some form of gradient descent to minimise the MSE

Non-Linear Relationship

- Now we need more data points to capture the nature of the function y=f(x) from data
- These two points are no longer enough



Non-Linear Relationships

- More data is needed for learning non-linear relationships as it is hard to tell the difference between random variation from a line and a curve when you don't have much data

Summary

- Data quality and quantity rely on:
  - The shape of the data's distribution
  - The number of variables in the data
  - The degree of linearity in the relationship to be captured
  - The amount of noise and unaccounted for variability in the data