Week 4

Data Mining and Knowledge Discovery

Dr John Evans
j.evans8@herts.ac.uk

University of Hertfordshire UH

Plan for today

Types of Data

- Attributes
- NOIR

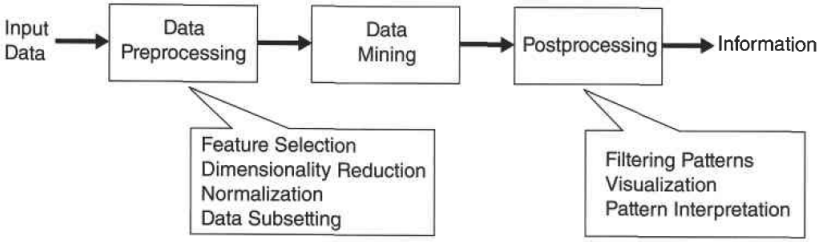
Data Quality

Noise

Data Collection

Recap - Moving on from SQL

In Week 1, we saw the Knowledge Discovery in Databases (KDD) process:



Now that we have an understanding of databases, we can start the actual KDD process. In the lab this will mean looking at ways we can input data while for this lecture, we will start the data preprocessing step. We are concerned with the notion of ‘cleaning up’ our data to make it look nice. To do this, we first need to understand what ‘nice’ means in this context.

Some motivating questions

1. Are we dealing with qualitative or quantitative data? In other words, what kind of mathematical and statistical operations can we perform?
2. Does the data have special characteristics, such as explicit relationships?
3. How was our data collected? Was it for some purpose other than data mining?
4. Are there missing values? Are there inconsistent or duplicate data?

Rough rule of thumb: Better quality data leads to better quality results.

Motivating example

Suppose we want to build a model which will allow us to predict a person's annual credit card spending given their annual income.

Motivating example

Suppose we want to build a model which will allow us to predict a person's annual credit card spending given their annual income.

1. The selection of data is now already done.
2. Next, we ask how the data was collected.
   - Is is accurate?
   - Do we have missing entries? Maybe our individual was out of the country.
   - Are all entries in the same currency?
   - Were there any one-off payments that may skew the data?

Motivating example

Suppose we want to build a model which will allow us to predict a person's annual credit card spending given their annual income.

1. The selection of data is now already done.
2. Next, we ask how the data was collected.
   - Is is accurate?
   - Do we have missing entries? Maybe our individual was out of the country.
   - Are all entries in the same currency?
   - Were there any one-off payments that may skew the data?
3. Once complete, we decide which model to use. A simple approach may be linear regression.
   - We build a predictive model to relate a predictor variable x (annual income) to a response variable y (annual credit card spending).
   - This is expressed through a relationship of the form $y = ax + b$.

Motivating example continued

1. This model may not be perfect, but since spending typically increases with income, hopefully it will be adequate. This is where we need some understanding of what is to be output.
   - If we need more accuracy and detail, perhaps this model will be unsuitable.
   - If we need only a rough idea, this model is fine. Assuming this is the case, linear regression is a good choice since it is simple unless the data set is very large.
   - Simple summaries of the data (sums, sums of squares, sums or products etc.)[1] are sufficient to compute estimates of a, b. This means a single pass through the data will yield results.

[1] Note this means we must be able to add our data, multiply values etc.

Motivating example continued

1. This model may not be perfect, but since spending typically increases with income, hopefully it will be adequate. This is where we need some understanding of what is to be output.
   - If we need more accuracy and detail, perhaps this model will be unsuitable.
   - If we need only a rough idea, this model is fine. Assuming this is the case, linear regression is a good choice since it is simple unless the data set is very large.
   - Simple summaries of the data (sums, sums of squares, sums or products etc.)[1] are sufficient to compute estimates of a, b. This means a single pass through the data will yield results.
2. Next, we need to decide how to quantify and compare how well our model fits the data.
   - The most commonly used 'score function' is the sum of the squared discrepancies between the predicted spending and observed spending. The smaller the better.

[1] Note this means we must be able to add our data, multiply values etc.

Motivating example continued

1. This model may not be perfect, but since spending typically increases with income, hopefully it will be adequate. This is where we need some understanding of what is to be output.
   - If we need more accuracy and detail, perhaps this model will be unsuitable.
   - If we need only a rough idea, this model is fine. Assuming this is the case, linear regression is a good choice since it is simple unless the data set is very large.
   - Simple summaries of the data (sums, sums of squares, sums or products etc.)[1] are sufficient to compute estimates of a, b. This means a single pass through the data will yield results.
2. Next, we need to decide how to quantify and compare how well our model fits the data.
   - The most commonly used 'score function' is the sum of the squared discrepancies between the predicted spending and observed spending. The smaller the better.
3. Finally, we decide what information we output and what form it takes. For something like linear regression, a good choice might be a simple graph.

[1] Note this means we must be able to add our data, multiply values etc.

Types of Data

- A data set is a set of measurements taken from some environment or process.
- A data set can be viewed as a collection of data objects.
  - Data objects are typically described by their attributes. These capture some characteristic of the object, such as mass of a physical object or the time when an event occurred.

Types of Data

- A data set is a set of measurements taken from some environment or process.
- A data set can be viewed as a collection of data objects.
  - Data objects are typically described by their attributes. These capture some characteristic of the object, such as mass of a physical object or the time when an event occurred.
- If the data objects are stored in a database, they are *data tuples*; that is, the rows of a database correspond to the data objects, and the columns to the attributes.

Example: A database could consist of the entire student body at UH. Each row would then correspond to a student, and each column to an attribute such as *Student ID, Year of Study, Course Enrolled, Modules enrolled, Grades* etc.

Attributes

- Attributes are also called *variables, characteristics, fields, features* or *dimensions*.
- Attributes can take many different forms.

Attributes

- Attributes are also called *variables, characteristics, fields, features* or *dimensions*.
- Attributes can take many different forms.
  - {*Blue, Brown, Green, Hazel*} is a set of possible eye colours.
  - Temperature is an attribute of an object that varies over time.
  - {*Student ID, Name, Home Address*} is a set of attributes of a student at UH.
  - *Height, weight, sex* are all attributes of a patient at a hospital.

Attributes

- Attributes are also called *variables, characteristics, fields, features* or *dimensions*.
- Attributes can take many different forms.
  - $\{$*Blue, Brown, Green, Hazel*$\}$ is a set of possible eye colours.
  - Temperature is an attribute of an object that varies over time.
  - $\{$*Student ID, Name, Home Address*$\}$ is a set of attributes of a student at UH.
  - *Height, weight, sex* are all attributes of a patient at a hospital.

N.B. Properties of an attribute need not be the same as the properties of the values used to measure it.

e.g. Consider student ID numbers. These are most likely represented by integers. So too is age. However, it does not make meaningful sense to compute the average student ID number, whereas it does make meaningful sense to compute the average age of a group of students.

Attribute type

We classify attributes into one of four categories (NOIR):

- Nominal
- Ordinal
- Interval
- Ratio

Attribute type

We classify attributes into one of four categories (NOIR):

- Nominal
- Ordinal
- Interval
- Ratio

These four classifications caputre one or more of the following properties:

- Distinctness, = and $\neq$
- Order, $<$, $\leq$, $>$ and $\geq$
- Addition, +, $-$
- Multiplication, $\times$, $/$

Cumulative description of attribute types: Qualitative

Attribute Type	Description	Examples	Operations
Nominal	These are just different names, i.e. nominal values provide only enough information to distinguish one object from another ($=, \neq$).	Postcode, employee ID, eye colour, sex	Mode, entropy, contingency, correlation, χ^2 test
Ordinal	These values can distinguish one object from another *and* provide an order ($<, >$).	hardness of minerals, {good, bad}, grades, street numbers	Median, percentiles, rank correlation, run tests, sign tests

Cumulative description of attribute types: Quantitative

Attribute Type	Description	Examples	Operations
Interval	We can now also measure the differences between values in a meaningful way, i.e. a unit of measurement exists $(+, -)$	Calendar dates, temperature (in Celsius or Fahrenheit)	Mean, standard deviation, Pearson's correlation, *t* and F tests.
Ratio	Now, both differences and ratios are meaningful, i.e. statements such as 'x is twice the size of y' now make sense $(\times, /)$	Temperature (in Kelvin), monetary quantities, counts, age, mass, length, electrical current.	Geometric mean, harmonic mean, percent variation

Example

- Suppose we measure the temperature of this room each day for a month. The data object here is the day and we can order these values. In this way, we get a ranking of which days were hottest/coldest.
- We can also quantify the swing in temperature from one day to the next.

Example

- Suppose we measure the temperature of this room each day for a month. The data object here is the day and we can order these values. In this way, we get a ranking of which days were hottest/coldest.
- We can also quantify the swing in temperature from one day to the next.
- This means temperature is certainly an interval attribute. Is it ratio?

Example

- Suppose we measure the temperature of this room each day for a month. The data object here is the day and we can order these values. In this way, we get a ranking of which days were hottest/coldest.
- We can also quantify the swing in temperature from one day to the next.
- This means temperature is certainly an interval attribute. Is it ratio?
- This depends on how we measure the temperature. If Celsius/Fahrenheit, then the answer is no. For ratio attributes, we need some absolute concept of 0 and $0^\circ C/F$ does not mean 'no temperature'.
- However, if we measure temperature in Kelvin, then this *does* have an absolute zero. This means we can make statements such as '$10^\circ K$ is half the temperature of $20^\circ K$. In this case, temperature is a ratio attribute.

Alternative ways to organise attributes

- A *discrete* attribute has a finite, or countably infinite, set of values.

Alternative ways to organise attributes

- A *discrete* attribute has a finite, or countably infinite, set of values.
  - Such attributes can be categorical (such as ID numbers) or numeric (such as counts).
  - Often represented using integer variables.
  - A special case is *binary*, which takes one of two values: 0 or 1.
  - e.g. *hair colour* is discrete, *smoker* is binary, *medical test* is discrete and can be binary, *drink size* is discrete, *student ID* is discrete and countably infinite.

Alternative ways to organise attributes

- A *discrete* attribute has a finite, or countably infinite, set of values.
  - Such attributes can be categorical (such as ID numbers) or numeric (such as counts).
  - Often represented using integer variables.
  - A special case is *binary*, which takes one of two values: 0 or 1.
  - e.g. *hair colour* is discrete, *smoker* is binary, *medical test* is discrete and can be binary, *drink size* is discrete, *student ID* is discrete and countably infinite.
- Those attributes which are not discrete are *continuous*. These generally mean the real numbers.
  - *Temperature, height, weight* are all continuous attributes, typically represented as floating-point variables.

Symmetry and asymmetry

- For *asymmetric* attributes, only presence is regarded as important, i.e. we are interested in non-zero values only.

Symmetry and asymmetry

- For *asymmetric* attributes, only presence is regarded as important, i.e. we are interested in non-zero values only.
  - Consider a data set where each student is a data object and each attribute (apart from perhaps an identifier such as *Student ID*) records whether a student takes a particular module at a university.
  - For a specific student, an attribute has a value of 1 if they take the module and 0 otherwise.

Symmetry and asymmetry

- For *asymmetric* attributes, only presence is regarded as important, i.e. we are interested in non-zero values only.
  - Consider a data set where each student is a data object and each attribute (apart from perhaps an identifier such as *Student ID*) records whether a student takes a particular module at a university.
  - For a specific student, an attribute has a value of 1 if they take the module and 0 otherwise.
  - As students take only a small fraction of all available modules, most of the values in the data set will be 0. Thus, it is more meaningful (and efficient) to focus on non-zero values.
  - Medical tests are usually asymmetric (binary) attributes.

Symmetry and asymmetry

- For *asymmetric* attributes, only presence is regarded as important, i.e. we are interested in non-zero values only.
  - Consider a data set where each student is a data object and each attribute (apart from perhaps an identifier such as *Student ID*) records whether a student takes a particular module at a university.
  - For a specific student, an attribute has a value of 1 if they take the module and 0 otherwise.
  - As students take only a small fraction of all available modules, most of the values in the data set will be 0. Thus, it is more meaningful (and efficient) to focus on non-zero values.
  - Medical tests are usually asymmetric (binary) attributes.
- A binary attribute in which both states are equally valuable and carry the same weight (e.g. *sex* having *male/female*) are called *symmetric attributes.*

Exercise: Read pp. 21-24 of the notes to learn more about the kinds of data we will encounter (specifically, record-data, graph-based data and ordered data).

Data quality

- The point of this is so ensure we actually consider what we can and cannot do with data.
- In the lab, we will see how to input data into Python. One of the first things we then do is compute some summary statistics, create basic visualisations and start to get an idea of what the data is telling us.
- What statistics we compute and what visualisations we produce is partly down to the type of data we have. A simple example would be the aforementioned mean computation of *Student ID*. Python would not complain if we asked it to do this, no error message is produced and we will get an output. However, the output is meaningless.

Data quality

- The next issue is to ensure our data is as ‘nice’ as possible. In order to extract insightful information, we need our data to be as tailored as possible to our purposes.
- This is where a problem arises: *Data is usually collected for some other purpose than data mining.*

[2]This step is often called *data cleaning.*

Data quality

- The next issue is to ensure our data is as ‘nice’ as possible. In order to extract insightful information, we need our data to be as tailored as possible to our purposes.
- This is where a problem arises: *Data is usually collected for some other purpose than data mining.*

Consequence: Preventing data quality problems is not usually an option. Instead, our focus becomes one of

- detection and correction[2]; and/or
- using algorithms that tolerate poor quality.

[2]This step is often called *data cleaning.*

What issues can we encounter?

- Human error.
- Limitations of measuring devices.
- Flaws in the data collection process.

What issues can we encounter?

- Human error.
- Limitations of measuring devices.
- Flaws in the data collection process.

This can result in values/data objects being missing, inaccurate or duplicated.

- e.g. two different records for the same person who has lived at two different addresses.
- e.g. inconsistencies in health records such as height being 2m but weight only 2kg.
- e.g. a similar species in a data collection region could be confused with the target species.

Data measurement

Definition: *Measurement error* refers to any problem resulting from the measurement process.

A common problem is that the value recorded differs from the true value. The extent of this difference determines how crucial the measurement error is.

Definition: For continuous attributes, the *error* is the numerical difference of the measured and true value.

The term *data collection error* refers to errors arising specifically from the data collection process.

Noise

Noise is a random error or variance in a measured variable that typically involves the distortion of a value or the addition of spurious objects.

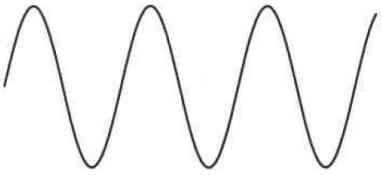
Noise

Noise is a random error or variance in a measured variable that typically involves the distortion of a value or the addition of spurious objects.

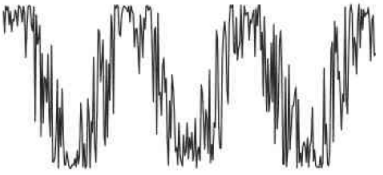
Noise has two main sources:

1. Implicit errors introduced by measurement tools (e.g. different types of sensors);
2. Random errors introduced by batch processes or experts when the data are gathered (e.g. in a document digitisation process).

Noise in a time series context

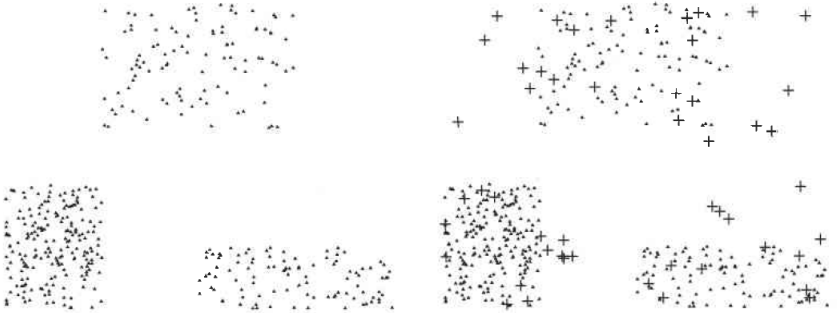


(a) Time series.



(b) Time series with noise.

Noise in a spatial context



(a) Three groups of points.

(b) With noise points (+) added.

Strategies to ‘smooth’ out data: Binning

Given the importance of classification, and the difficulty presented by noise, there are lots of resources/strategies aimed at overcoming this. One such strategy is *binning*.

Strategies to ‘smooth’ out data: Binning

Given the importance of classification, and the difficulty presented by noise, there are lots of resources/strategies aimed at overcoming this. One such strategy is *binning*.

Binning strategies smooth a sorted data value by consulting its ‘neighbourhood’, i.e. the values around it. The sorted values are then distributed into a number of ‘bins’.

As these methods consult neighbourhoods, they perform what is called *local* smoothing.

Example

Suppose we have the data for a price attribute which has been sorted:
4, 8, 15, 21, 21, 24, 25, 28, 34.

[3] We can also smooth by bin medians, or by bin boundaries.

Example

Suppose we have the data for a price attribute which has been sorted:
4, 8, 15, 21, 21, 24, 25, 28, 34. Next, we partition into *equal-frequency* bins of size 3:

Bin 1	4, 8, 15
Bin 2	21, 21, 24
Bin 3	25, 28, 34.

[3] We can also smooth by bin medians, or by bin boundaries.

Example

Suppose we have the data for a price attribute which has been sorted:
4, 8, 15, 21, 21, 24, 25, 28, 34. Next, we partition into *equal-frequency* bins of size 3:

Bin 1	4, 8, 15
Bin 2	21, 21, 24
Bin 3	25, 28, 34.

In a *smoothing by bin means*[3] method, each value in a bin is replaced by the mean value of the bin. We end up with the following:

Bin 1	9, 9, 9
Bin 2	22, 22, 22
Bin 3	29, 29, 39.

In general, the larger the width, the greater the effect of the smoothing.

[3] We can also smooth by bin medians, or by bin boundaries.

Strategies to ‘smooth’ out data: Regression

- Data smoothing can be done by regression (conforms data values to a function).
- Linear regression involves finding the ‘best’ line to fit two attributes (or variables) so that one attribute can be used to predict the other. We can even attempt to fit data to a multidimensional surface.

Strategies to ‘smooth’ out data: Outliers

- Outliers may be detected by clustering, for example, where similar values are organised into groups (or ‘clusters’).
- Intuitively, values that fall outside of the set of clusters may be considered outliers.
- We can then remove such outliers. This would work if we are trying to find dense areas of data that we suspect contain the information we seek.

Precision, bias and accuracy

In statistics and experimental science, the quality of the measurement process and the resulting data are measured by precision and bias.

Definition: *Precision* is the closeness of repeated measurements (of the same quantity) to one another.

Definition: *Bias* refers to the systematic variation of measurements from the quantity being measured.

Definition: *Accuracy* is the closeness of measurements to the true value of the quantity being measured.

Measuring precision, bias and accuracy

- A typical way to measure precision is using the standard deviation of a set of values.

Measuring precision, bias and accuracy

- A typical way to measure precision is using the standard deviation of a set of values.
- Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.
  - Given this, bias can only be determined for objects whose measured quantity is known by means external to the current situation.

Measuring precision, bias and accuracy

- A typical way to measure precision is using the standard deviation of a set of values.
- Bias is measured by taking the difference between the mean of the set of values and the known value of the quantity being measured.
  - Given this, bias can only be determined for objects whose measured quantity is known by means external to the current situation.
- Accuracy depends on precision and bias, but there is no specific formula for accuracy in terms of these quantities. Instead, it is more of a general term that captures both ideas.

Example

Suppose we have a standard laboratory weight with a mass of 1g and we want to assess the precision and bias of a new laboratory scale. We weigh the mass five times and obtain the following five values: $\{1.015, 0.990, 1.013, 1.001, 0.986\}$.

Example

Suppose we have a standard laboratory weight with a mass of 1g and we want to assess the precision and bias of a new laboratory scale. We weigh the mass five times and obtain the following five values: $\{1.015, 0.990, 1.013, 1.001, 0.986\}$.
Keeping the notation from the notes, we have the following:

$$
\begin{aligned}
N &= 5 \\
\bar{x} &= 1.001 \\
\sigma^2 &= \tfrac{1}{4}[(1.015 - 1.001)^2 + (0.990 - 1.001)^2 + (1.013 - 1.001)^2 + \\
&\qquad +(1.001 - 1.001)^2 + (0.986 - 1.001)^2] \\
&= 0.0001715 \\
\sigma &= 0.013
\end{aligned}
$$

So, the bias is $1.001 - 1.000 = 0.001$ and the precision is 0.013.

Data collection: Missing values

It is not unusual for an object to be missing one or more attribute values. This could be due to any of the following:

- The information not being collected (e.g. a person declining to give their age or weight).

Data collection: Missing values

It is not unusual for an object to be missing one or more attribute values. This could be due to any of the following:

- The information not being collected (e.g. a person declining to give their age or weight).
- An attribute not being applicable to all objects (e.g. a form which has conditional parts that are filled out only if a previous question is answered in a specific way).

Data collection: Missing values

It is not unusual for an object to be missing one or more attribute values. This could be due to any of the following:

- The information not being collected (e.g. a person declining to give their age or weight).
- An attribute not being applicable to all objects (e.g. a form which has conditional parts that are filled out only if a previous question is answered in a specific way).
- Data being lost (e.g. when transferring data from one database to another).

There are several strategies for dealing with missing data, each of which is appropriate in certain circumstances.

Data Collection: Missing values

- In practice, we can utilise several strategies simultaneously.
- Ideally, data sets are accompanied by documentation that describes different aspects of the data, e.g. the documentation may identify several attributes that are important and/or strongly related.
  - If the former, then we may not want to eliminate such attributes.
  - If the latter, then perhaps we do not need all of the attributes since one is known to imply the other. A typical example of this is sales tax and purchase price.

Data Collection: Missing values

- In practice, we can utilise several strategies simultaneously.
- Ideally, data sets are accompanied by documentation that describes different aspects of the data, e.g. the documentation may identify several attributes that are important and/or strongly related.
  - If the former, then we may not want to eliminate such attributes.
  - If the latter, then perhaps we do not need all of the attributes since one is known to imply the other. A typical example of this is sales tax and purchase price.

Of course, the documentation may be poor (e.g. it may fail to tell us that missing values are indicated with a -9999). In which case, our analysis may be faulty.

Strategies for dealing with missing values

- Eliminate data objects or attributes

Strategies for dealing with missing values

- Eliminate data objects or attributes
- Estimate missing values, e.g. interpolation, average/mode, nearest neighbour etc.

Strategies for dealing with missing values

- Eliminate data objects or attributes
- Estimate missing values, e.g. interpolation, average/mode, nearest neighbour etc.
- Ignore the missing values during analysis.

Data collection: Inconsistent values

- Data can often contain inconsistent values, e.g. when entering your address, it is not uncommon to enter errors.
  - This could be due to a postcode that technically does not belong to a given city, but locally is thought to (this could be due to boundary changes, for example); or
  - due to a typo being entered (O instead of 0, for example). Such errors could be due to an individual or if the information is being scanned from a handwritten form.

Data collection: Inconsistent values

- Data can often contain inconsistent values, e.g. when entering your address, it is not uncommon to enter errors.
  - This could be due to a postcode that technically does not belong to a given city, but locally is thought to (this could be due to boundary changes, for example); or
  - due to a typo being entered (O instead of 0, for example). Such errors could be due to an individual or if the information is being scanned from a handwritten form.
- Regardless of the cause, such inconsistent values are important to detect and, if possible, correct. Depending on the inconsistency, the difference may be easy to correct.
  - For example, if a postcode/city pair are often entered even if they are not technically correct, then this would be an easy fix to implement.
  - Likewise if a height is entered as a negative, this is clearly an inconsistency.
  - However, if the error is a typo (1 instead of 2, say) then this may be more difficult to detect and/or correct.

Example

This example illustrates an inconsistency in actual time series data that measures the **sea surface temperature (SST)** at various points on the ocean.

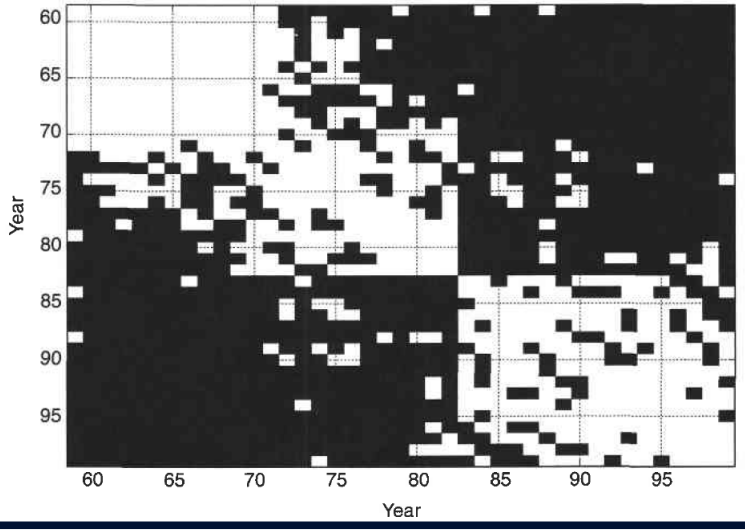
- SST data was originally collected using ocean-based measurements from ships or buoys, but more recently using satellites.
- For example, see NOAA SST Data.

Example

This example illustrates an inconsistency in actual time series data that measures the **sea surface temperature (SST)** at various points on the ocean.

- SST data was originally collected using ocean-based measurements from ships or buoys, but more recently using satellites.
- For example, see NOAA SST Data.
- To create a long-term data set, both sources of data must be used. However, as the data comes from different sources, the two parts of the data are subtly different. This discrepancy is visually displayed on the next slide and shows the correlation of SST values between pairs of years.

Example continued



Example concluded

- If a pair of years has a positive correlation, then the location corresponding to the pairs of years is coloured white; otherwise, it is coloured black.[4]

[4] Seasonal variations were removed, otherwise all years would be highly correlated.

Example concluded

- If a pair of years has a positive correlation, then the location corresponding to the pairs of years is coloured white; otherwise, it is coloured black.[4]
- We note that there is a distinct change in behaviour where the data has been put together in 1983.
  - Years within each of the two groups (1958-1982, 1983-1999) tend to have a positive correlation with one another, but a negative correlation with years in the other group.
  - This does not mean that this data should not be used, only that the analyst should consider the potential impact of such discrepancies on the data mining analysis.

[4] Seasonal variations were removed, otherwise all years would be highly correlated.

Summary

- We can classify attributes as either qualitative or quantitative.
- We can further classify attributes as Nominal, Ordinal, Interval, Ratio (NOIR).
- We can also classify attributes as continuous or discrete (binary).
- For asymmetric attributes, only presence is regarded as important (i.e. we are interested in non-zero values).
- A binary attribute in which both states are equally valuable and carry the same weight are called symmetric.
- We have discussed noise and ways of ‘smoothing out’ data.
- We have seen ways to define and computed precision and bias (discussed accuracy also).
- We have seen how to deal with missing/inconsistent values.