

FIT1043 Introduction to Data Science

Module 5: Data Analysis Process

Lecture 8

Monash University

Discussion

In the tutorial you used different tools for data wrangling:

- ▶ DataWrangler
 - ▶ specialised
 - ▶ intuitive Graphical User Interface (GUI)
 - ▶ no coding
- ▶ Python
 - ▶ general purpose, open-source
 - ▶ contains packages (Pandas) for manipulating data
- ▶ SAS (optional to investigate)
 - ▶ general purpose
 - ▶ strange syntax!
 - ▶ very widely used

Note that there are many other tools we could have used

- ▶ R, Matlab, Java, SPSS.

Unit Schedule: Modules

Module	Week	Content
1.	1 2	Overview and look at projects (Job) roles, and the impact
2.	3	Data business models / application areas
3.	4 5	Characterising data and "big" data Data sources and case studies
4.	6 7	Resources and standards Resources case studies
5.	8 9 10	Data analysis theory Regression and decision trees Data analysis process
6.	11 12	Issues in data management Data management frameworks

Introduction to Data Analysis (ePub section 5.1)

motivating examples

Essential Viewing

- ▶ *“The power of emotions: When big data meets emotion data”*, by Rana El Kaliouby
- ▶ *“The wonderful and terrifying implications of computers that can learn”* at TED by Jeremy Howard
- ▶ *“The Unreasonable Effectiveness of Data”* lecture at Univ. of British Columbia by Peter Norvig

Implications of Computers that Learn: Examples

- ▶ checkers (1956)
- ▶ German traffic sign recognition (2011),
- ▶ predicting breast cancer survival rates from images (2011)
- ▶ Microsoft's Chinese text-speech-text (2012)
- ▶ [*IBM Watson at Jeopardy \(2003\)*](#)
Jeopardy! is an American television game show.

Theory of Data Analysis

(ePub section 5.2)

introduction to the intuitions behind theory, but avoiding mathematics

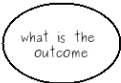
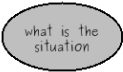


- ▶ Models of data analysis problems
- ▶ Introduction to learning theory

Theory of Data Analysis

Models of data analysis problems

- ▶ simple prediction task
- ▶ more complicated prediction task
- ▶ segmentation (aka clustering) task
- ▶ time series forecasting and sequential learning tasks

Remember: Node Types in Influence Diagrams

CHANCE VARIABLE	KNOWN VARIABLE	DECISION	OBJECTIVE
			

When do we connect an arc to a node?

Chance variable: connect to if it “causes” (is not “procedural”);

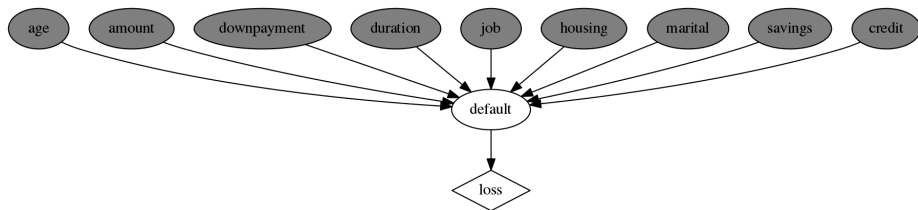
Known variable: no arcs generally, but may show if a related graph has them

Decision: connect to if variable used when making decision;

Objective: connect to if variable used when evaluating;
quality/value/cost of objective

Simple Prediction Task:

Housing Loan Default

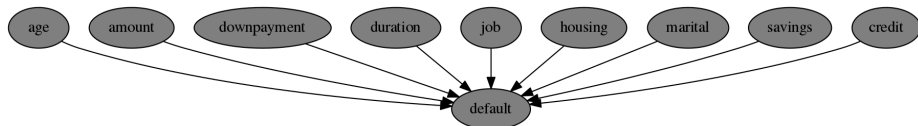


Task: Predict whether an individual will **default** on their loan

- ▶ based on a number of known **feature values**:
 - ▶ age, amount, downpayment, duration, ...
- ▶ the **loss** to the bank is high for a default
 - ▶ not loaning -> loss of business

Simple Prediction Task:

Training Data

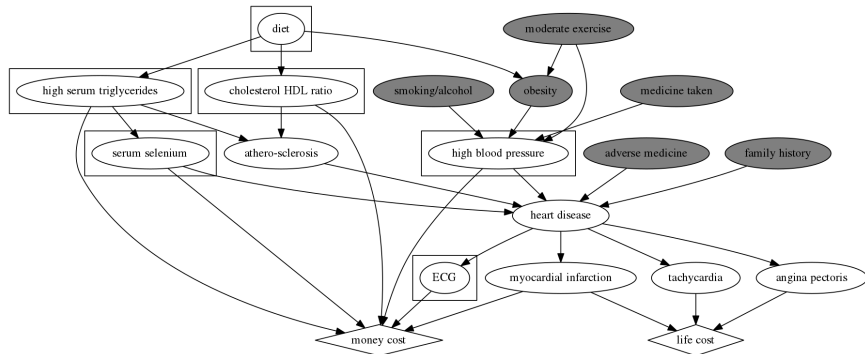


In order to **learn a model**,

- ▶ we're given a database of cases where the true status of **default** is known

Complicated Prediction Task:

Heart Disease Diagnosis

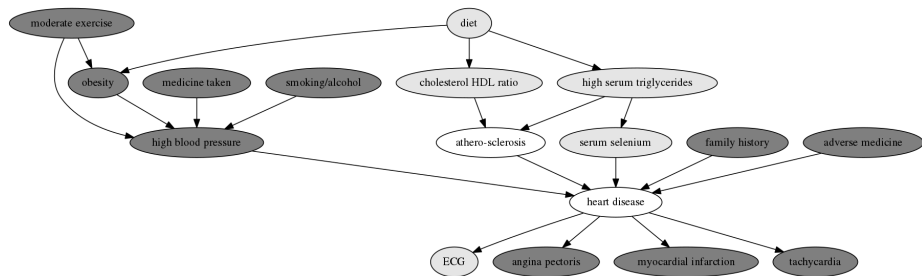


Model contains many variables that link to one another in complicated ways, (called a Bayesian Network)

- ▶ many of the variables are unknown
- ▶ different patients might have **different knowns**

Complicated Prediction Task:

Training Data



- ▶ supplied data may have more complete set of tests done but still have some unknowns

Segmentation Task:

Identifying Customer Segments

- ▶ customers are grouped into **segments**
- ▶ marketing is then specialised to each segment
- ▶ leads to better marketing
- ▶ **but how do you do the grouping?**

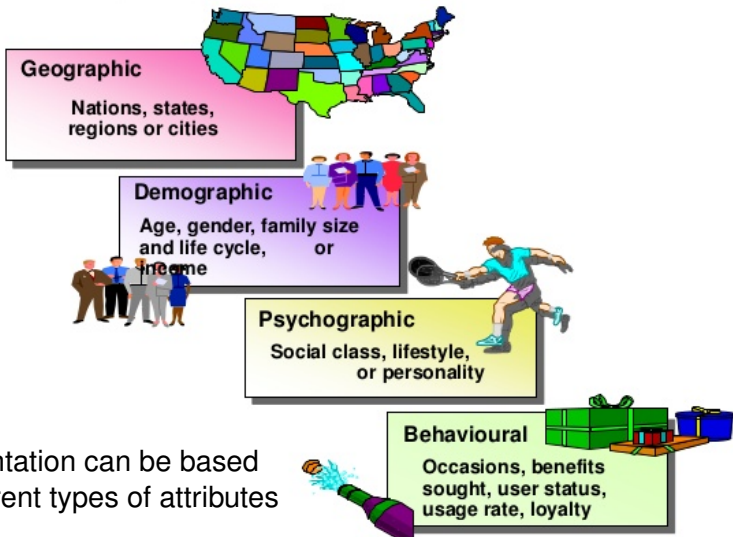


Example segmentation:

- ▶ traditional segmentation in Britain uses class, (from *the Independent*)

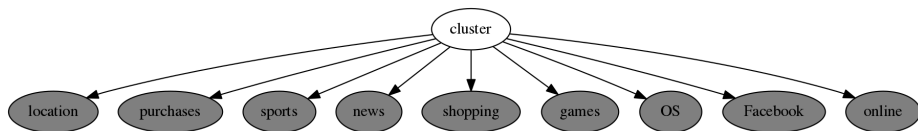
Market Segmentation

Bases for Segmenting Consumer Markets



Segmentation can be based on different types of attributes

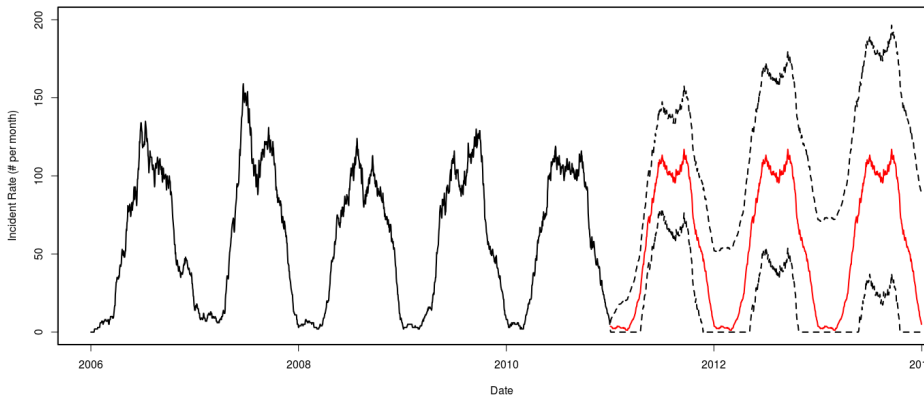
Segmentation (cont.)



- ▶ the *cluster* variable is unknown
- ▶ the cluster variable identifies the segments

Time Series Forecasting

Projected bicycle collision rates in Montreal



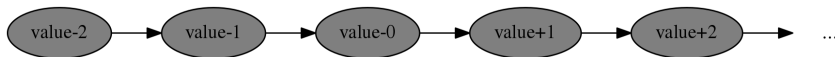
from [*bayesianbiologist*](#)

Time Series: 1st Order

Task: predict the next value in a series based on the previous value from the same series:



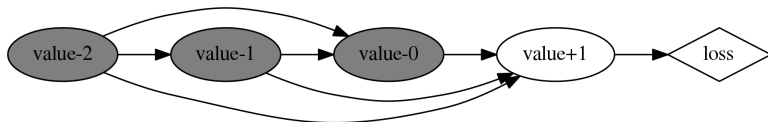
Training data consists of one or more series of values:



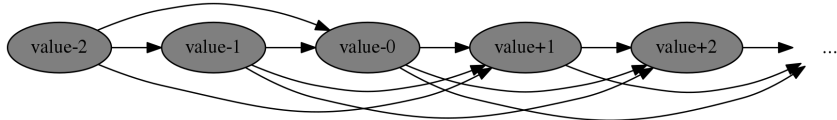
Time Series: 3rd Order

Higher order models predict the next value in a series based on more than just the previous value:

- ▶ in this case the last 3 values

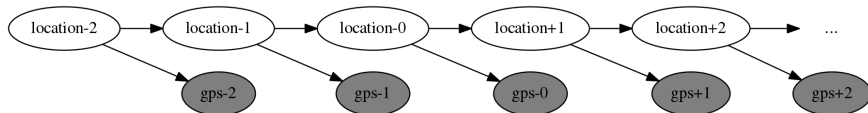


Training data is again just sequences of data:



Sequential Learning Task:

GPS Tracking



In the case of GPS tracking:

- ▶ the “true” location is never actually known
- ▶ but can be inferred approximately from observed GPS signal, coupled with knowledge of signal noise and speed considerations

Theory of Data Analysis Introduction to Learning Theory

What is Learning Theory

Wikipedia definition: (Computational) learning theory is a subfield of Artificial Intelligence devoted to studying the design and analysis of machine learning.

Truth

Heart Disease Diagnosis

- ▶ For a single patient the “truth” can be measured directly
- ▶ How can you measure the “true” model?
 - ▶ collect infinite data
 - ▶ but: a dynamic problem
- ▶ We assume some underlying “truth” is out there

Quality

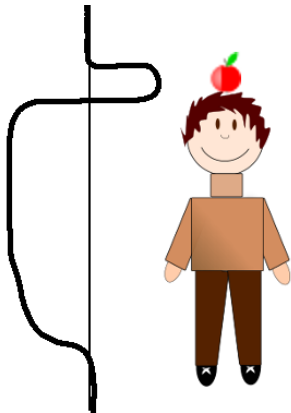
- ▶ to evaluate the quality of results derived from learning, we need notions of value
- ▶ so we will review quality and value

William Tell's Apple Shot



- ▶ William Tell forced to shoot the apple on his son's head
- ▶ if he strikes it, he gets both their freedoms

William Tell's Apple Shot, cont.



- ▶ this shows “value” as a function of height
- ▶ loss varies depending on where it strikes
- ▶ how do you compare loss of life versus gain of freedom?

the boy is smiling! its hard to find a cartoon with an apple on a boy's head

Quality

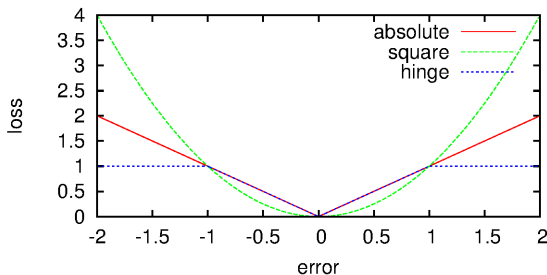
- ▶ may be the quality of your prediction
- ▶ may be the consequence of your actions
(making a prediction is a kind of action)
- ▶ can be measured on a positive or negative scale

loss: positive when things are bad, negative (or zero) when they're good

gain: positive when things are good, negative when they're not

error: measure of “miss”, sometimes a distance, but **not** a measure of quality

Quality is a Function of Error



error measures the distance between the prediction and the actual value

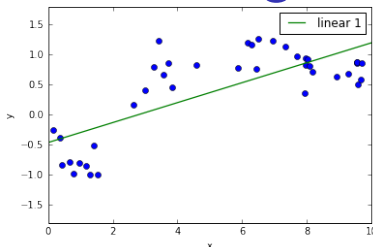
- ▶ “0” means no error, prediction was exactly right
- ▶ we can convert error to a measure of quality using a loss function, e.g.:

$$\text{absolute-error}(x) = |x|$$

$$\text{square-error}(x) = x * x$$

$$\text{hinge-error}(x) = \begin{cases} |x| & \text{if } |x| \leq 1 \\ 1 & \text{otherwise} \end{cases}$$

Linear Regression



data is shown with blue dots, green line is the **linear** “fitted model”

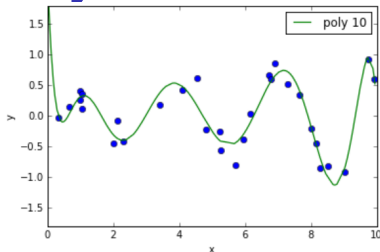
regression fits a very simple equation to the data:

$$\hat{y}(x; \vec{a}) = a_0 + a_1 x$$

- ▶ Here $\hat{y}(x; \vec{a})$ is the prediction for y at the point x using the model parameters $\vec{a} = (a_0, a_1)$, i.e. the intercept and slope terms.
- ▶ Given some data pairs $(x_1, y_1), \dots, (x_N, y_N)$, we fit a model by finding the vector \vec{a} that minimises the loss function:

$$\text{mean square error} = MSE_{train} = \frac{1}{N} \sum_{i=1}^N (\hat{y}(x_i; \vec{a}) - y_i)^2$$

Polynomial Regression



data is shown with blue dots, green curve is the polynomial “fit”

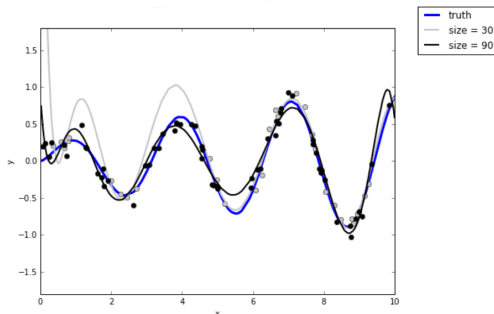
polynomial regression uses the same linear regression infrastructure to fit a higher order polynomial. In this case we fit a 10-th order polynomial:

$$\hat{y}(x; \vec{a}) = a_0 + a_1x + a_2x^2 + \dots a_9x^9 + a_{10}x^{10} = \sum_{i=0}^{10} a_i x^i$$

By finding the vector \vec{a} that for a given set of data pairs $(x_1, y_1), \dots, (x_N, y_N)$ minimises the loss function:

$$\text{mean square error} = MSE_{\text{train}} = \frac{1}{N} \sum_{i=1}^N (\hat{y}(x_i; \vec{a}) - y_i)^2$$

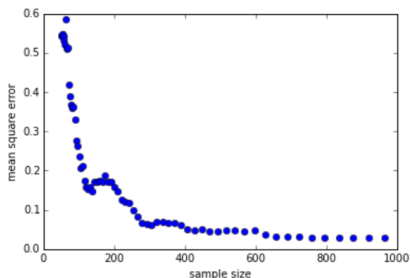
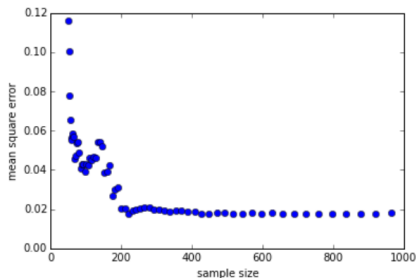
More Data Improves the Fit



- ▶ blue line is true model that generated the data (before noise was added)
- ▶ grey curve is model fit to 30 data points
- ▶ black curve is model fit to 90 data points

In general, more data means better fit

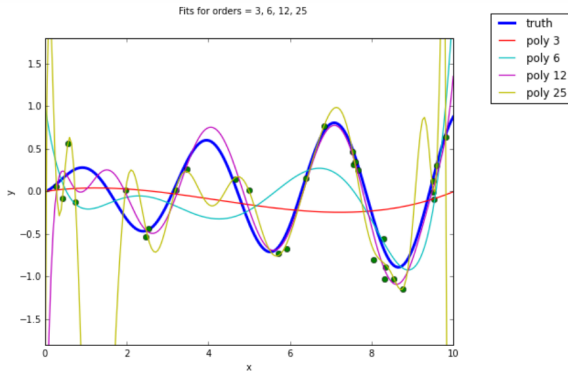
Loss decreases with Training Data



MSE decreases as the amount of training data grows

- ▶ these plots are called **learning curves**
- ▶ different learning algorithms exhibit different behaviour (rate of decay)

Overfitting



The more parameters a model has, the more complicated a curve it can fit.

- ▶ If we don't have very much data and we try to fit a complicated model to it, the model will make wild predictions.
- ▶ This phenomenon is referred to as **overfitting**

Overfitting, cont.

- ▶ small polynomial; cannot fit the data well; said to have **high bias**
- ▶ large polynomial; can fit the data well; fits the data too well; said to have **small bias**
- ▶ if there is known error in the data, then a close fit is wasted:
 - ▶ 25-th degree polynomial does all sorts of wild contortions!
- ▶ poor fit due to high bias called **underfitting**
- ▶ poor fit due to low bias called **overfitting**

Unit Schedule: Next Week

Module	Week	Content
1.	1	Overview and look at projects (Job) roles, and the impact
	2	
2.	3	Data business models / application areas
3.	4	Characterising data and "big" data Data sources and case studies
	5	
4.	6	Resources and standards Resources case studies
	7	
5.	8	Data analysis theory Regression and decision trees Data analysis process
	9	
	10	
6.	11	Issues in data management Data management frameworks
	12	