FIT1043 Introduction to Data Science

Module 5: Data Analysis Process

Lecture 9

Monash University

Discussion: Investigating Twitter data in the Shell

Last week we spent another tutorial analysing a **large data file** from Twitter in the shell:

- aim was to understand what data the file contained and how we could reformat the data for further analysis
- file contained many **different types of columns**:
  - text, dates, locations, even code containing data structures
- real data: lots of missing data, errors, ...
- shell commands like *grep* and *cut* simplify the inspection and manipulation of the data

Unit Schedule: This Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5.	**9**	**data analysis theory**
	10	**data analysis process**
6.	11	issues in data management
	12	data management frameworks

Introduction to Data Analysis
(ePub section 5.1)

motivating examples

Essential Viewing

- *"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
- *"Knowledge is Beautiful"* by David McCandless at the RSA
- *"The power of emotions:* When big data meets emotion data", by Rana El Kaliouby
- *"How Predictive Predictive Analytics Is"*, another cartoon intro to a subject from Patricia Florissi of EMC. Look at these parts: accident estimation in cities at 6:06 and aircraft maintenance at 10:10.

Implications of Computers that Learn

From *2014 TED talk* by Jeremy Howard

Examples: checkers (1956), IBM Watson at Jeapody (2003), German traffic sign recognition (2011), predicting breast cancer survival rates from images (2011), Microsoft's Chinese text-speech-text (2012)

Capability: from a picture, generate text explaining it

Need: will never be enough trained doctors for developing world, so use machine learning instead to train up computers

Revolution: computers keep on getting better, exponential improvement, machine learning is a revolution on par with the Industrial Revolution

Theory of Data Analysis
(ePub section 5.2)

introduction to the intuitions behind theory, but avoiding mathematics

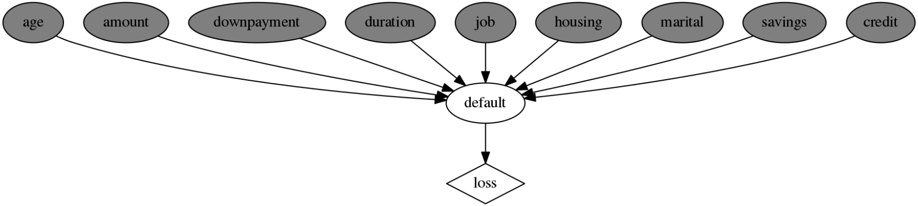
- graphical models
  - structural models of data analysis problems
  - characterising learning problems
- introduction to learning theory
  - key ideas from theory

Theory of Data Analysis
Graphical Models

structural models of data analysis problems:

- simple prediction (aka classification/regression) task
- more complicated prediction task
- segmentation (aka clustering) task
- time series forecasting and sequential learning tasks
- causal inference task

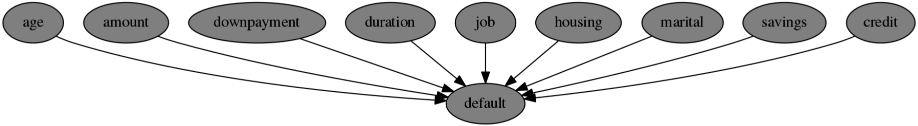
Simple Prediction Task:
Housing Loan Default



Task is to predict whether an unknown value:

- whether or not 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
  - but not loaning results in loss of business
  - would need a decision node (lend?) to define this loss.

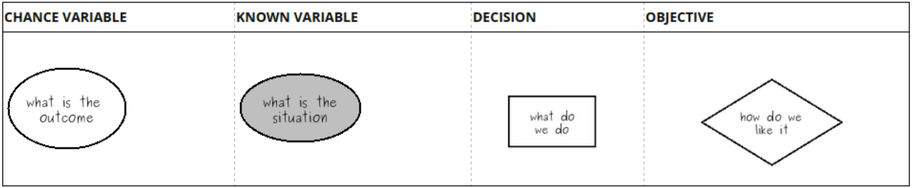
Simple Prediction Task:
Training Data



In order to **learn a model**,

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

Node Types



When do we connect an arc to a node?

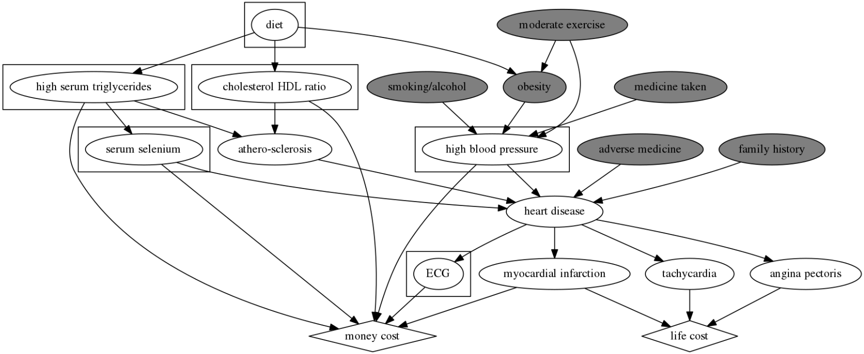
Chance variable: connect to if it “causes” (is not “procedural”);

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

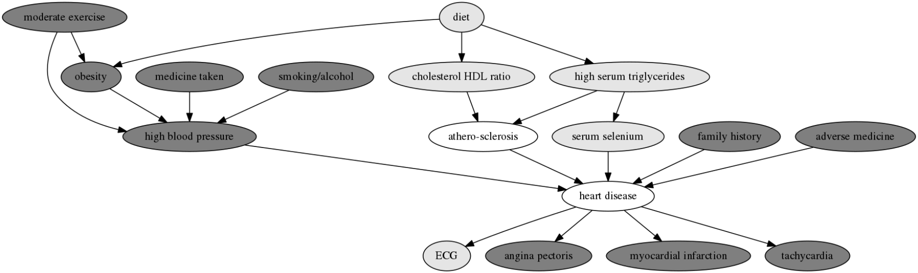
Complicated Prediction Task:
Heart Disease Diagnosis



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:

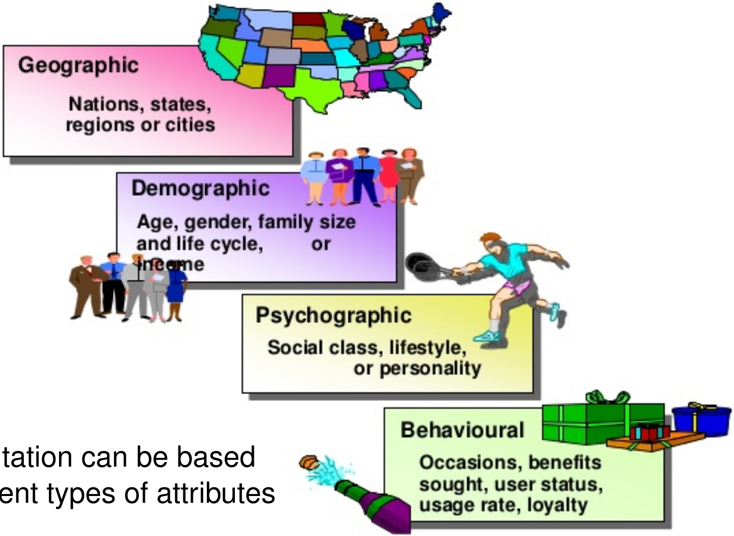
- traditional segmentation in Britain uses class, (from *the Independent*)

Another example segmentation ;-):

- used to develop specialised political marketing, (from *graphicgranola blog*)

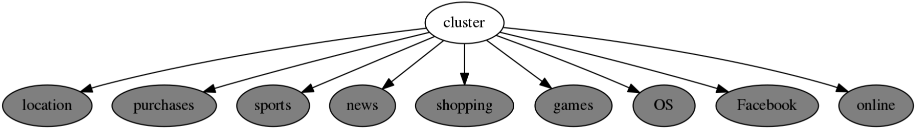
Market Segmentation

Bases for Segmenting Consumer Markets



Segmentation can be based on different types of attributes

Segmentation (cont.)



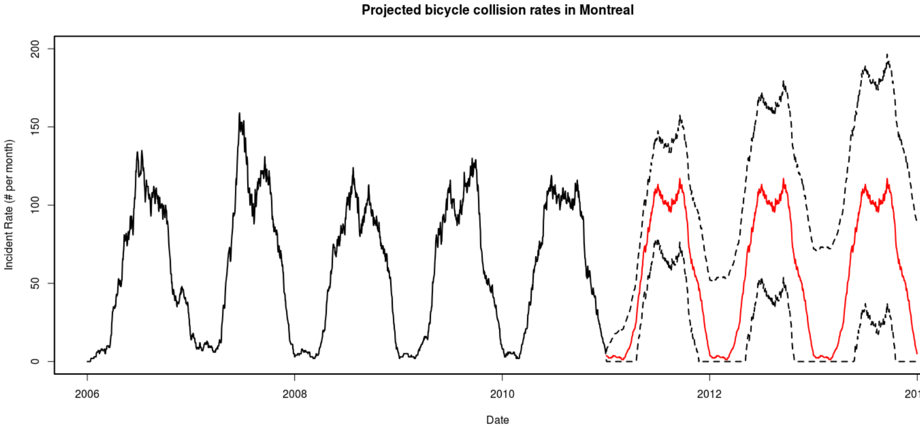
A segmentation model is a graphical model where

- the *cluster* variable is unknown, called “latent”
- the cluster variable identifies the segments
- latent means the variable is never observed in the data

For examples of the use of clustering, watch:

- *“How Predictive Predictive Analytics Is”* starting at 1:30

Time Series Forecasting

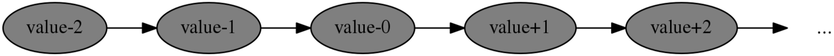


from *bayesianbiologist*

Time Series: 1st Order

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

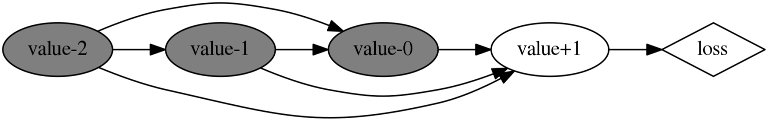
Training data consists of one or more series of values:



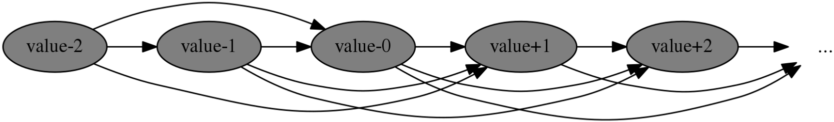
Time Series: 3rd Order

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

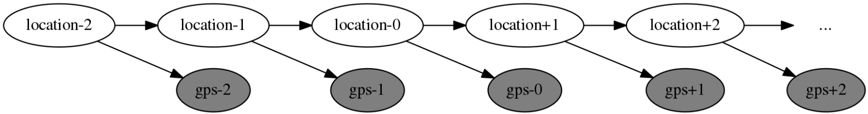
- in this case the last 3 values



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

Causal Models: Obesity

Example of a really big causal model for obesity:

- *"causal loop diagram"*

Raises more questions than answers:

- does this degree of complexity help?
- can it be practically used?
- could it ever be tested on real data?
- is it more a conceptual artifact to support researchers?

Theory of Data Analysis
Introduction to Learning Theory

key ideas from theory

Truth

For variables for an individual data case (e.g. a single loan application or a single heart disease patient), the “truth” can be measured directly

- Across examples, the “true” model is harder to define:
  - What is a “true” model of physics? – Newtonian physics, String Theory?
- How can you measure the “true” model for the heart disease problem?
  - collect infinite data and infer statistically
  - but its a dynamic problem and general population characteristics always changing
- regardless, we assume some underlying “truth” is out there

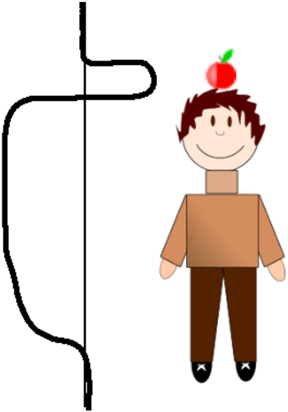
Quality

- 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

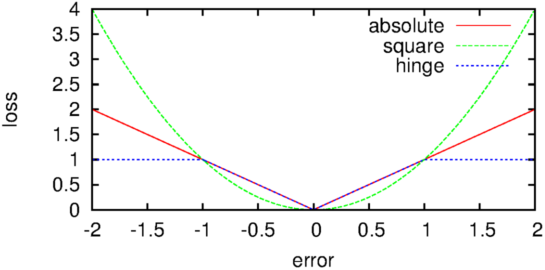
- may be the quality of your prediction
- may be the consequence of your actions (making a prediction is a kind of action)
- can be measured on a positive or negative scale

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

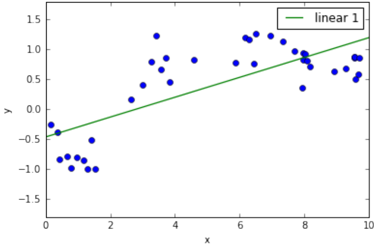


error measures the distance between the prediction and the actual value

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

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

Linear Regression



data is shown with blue dots, green line is the linear "fitted model"

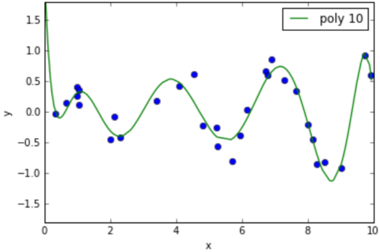
regression fits a very simple equation to the data:

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

- 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 $(x_1, y_1), ..., (x_N, y_N)$, we fit a model by finding the vector $\vec{a}$ that minimises the loss function:

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

Polynomial Regression



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

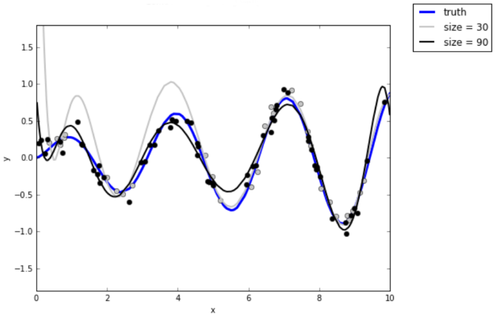
polynomial regression uses the same linear regression infrastructure to fit a higher order polynomial.
In this case we fit a 10-th order polynomial:

$$\hat{y}(x;\vec{a}) = a_0 + a_1 x + a_2 x^2 + \ldots a_9 x^9 + a_{10} x^{10} = \sum_{i=0}^{10} a_i x^i$$

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

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

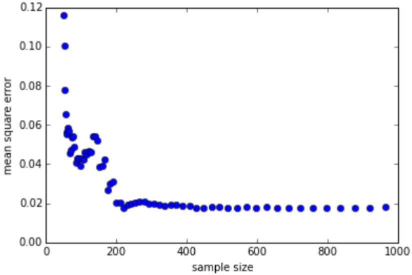
More Data Improves the Fit

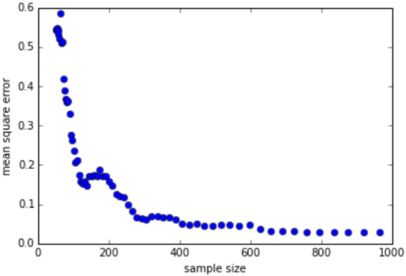


- blue line is true model that generated the data (before noise was added)
- grey curve is model fit to 30 data points
- black curve is model fit to 90 data points

In general, more data means better fit (most of the time)

Loss decreases with Training Data

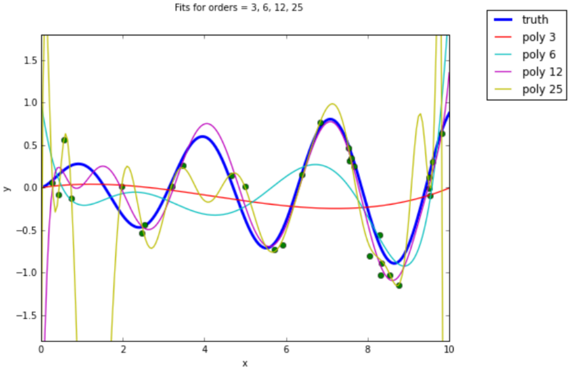




MSE decreases as the amount of training data grows

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

Overfitting



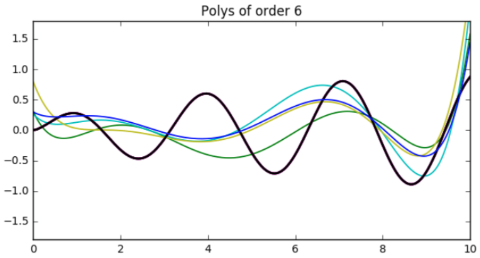
The more parameters a model has, the more complicated a curve it can fit.

- 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

Bias and Variance

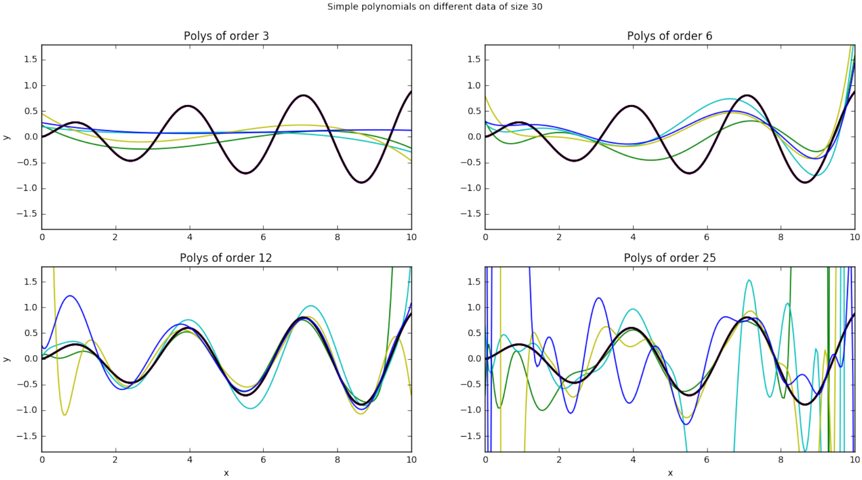


On the plot: different data sets of size 30, showing there fit.

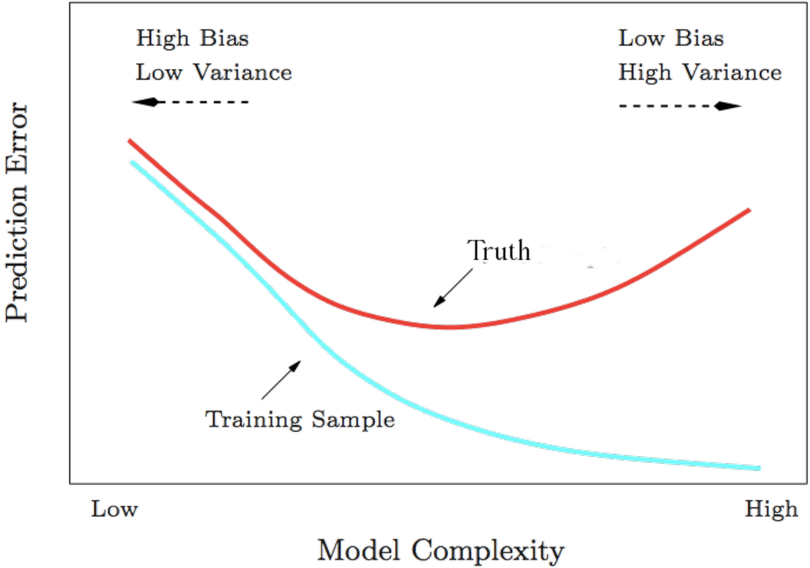
Bias: what is the *least error* one can get when fitting any possible model to the data (impracticle to achieve).

Variance: what is the *average error* one gets for different data sets *over and above the minimum error*.

Bias-Variance Examples



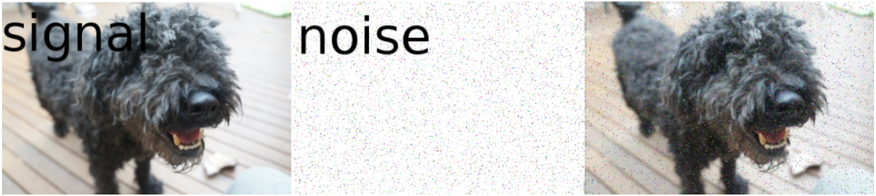
Bias-Variance Tradeoff



Training Set and Test Set

- split up the data we have into two non-overlapping parts, a training set and a test set
- do your learning, run your algorithm, build your model using the training set
- run evaluation using the test set
- don't run evaluation on the training set
- how big to make the test set?

Signal versus Noise

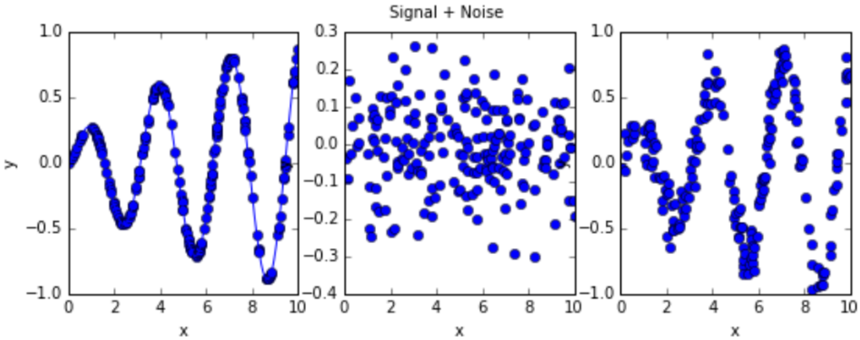


Signal: "truth" usually unknown

Noise: difference between "truth" and the data

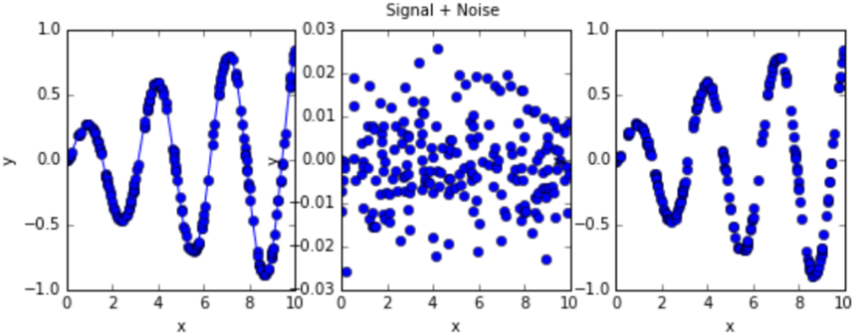
- notion used in communications (pictures, video, etc.)
- problem is we usually don't know what the noise is!

Signal versus Noise, cont.



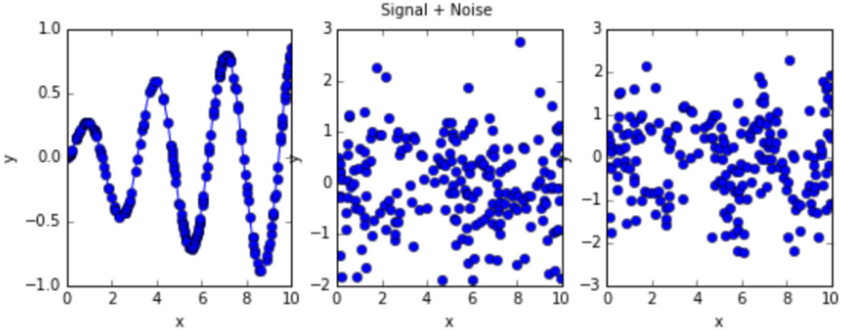
- same idea, applied to regression

Signal versus (Low) Noise



- noise level is “low” compared to signal strength
- “low” is relative and hard to quantify in practice

Signal versus (High) Noise



- noise level is “high” compared to signal strength
- “high” is relative and hard to quantify in practice

No Free Lunch Theorem

Wolpert and McCready proved:

> *if a [learning] algorithm performs well on a certain class of problems then it necessarily pays for that with degraded performance on the set of all remaining problems*

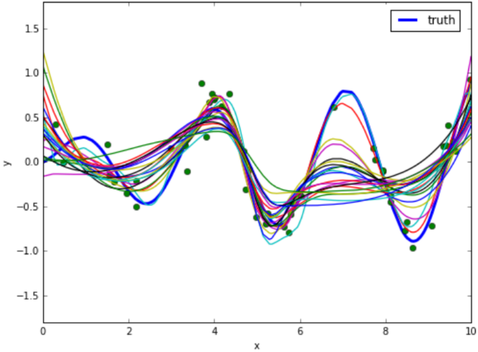
- there is no universally good machine learning algorithm (when one has finite data)

e.g. Naive Bayesian classification performs well for text classification with smaller data sets

e.g. linear Support Vector Machines perform well for text classification

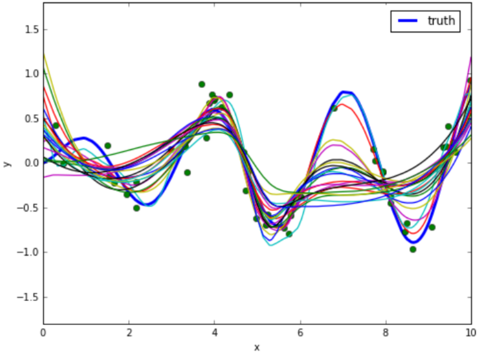
Ensembles

- given only data, we do not know the truth and can only estimate what may be the "truth"
- an ensemble is a collection of possible/reasonable models
- from this we can understand the variability and range of predictions that is realistic

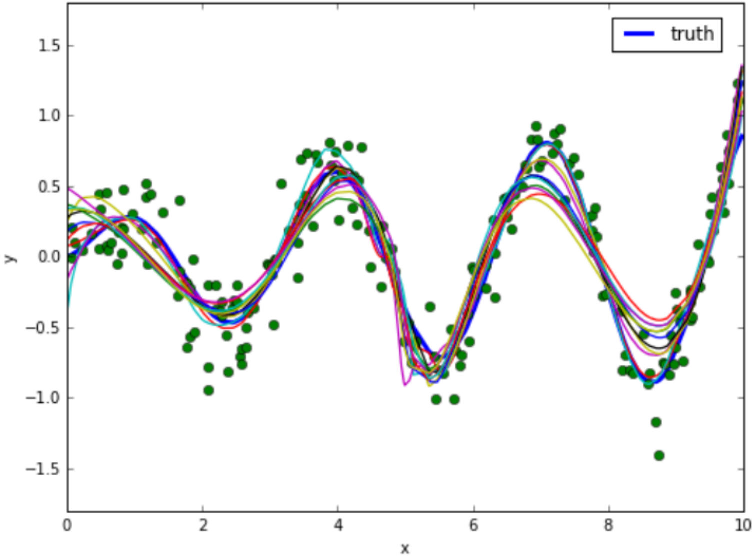


Ensembles (cont.)

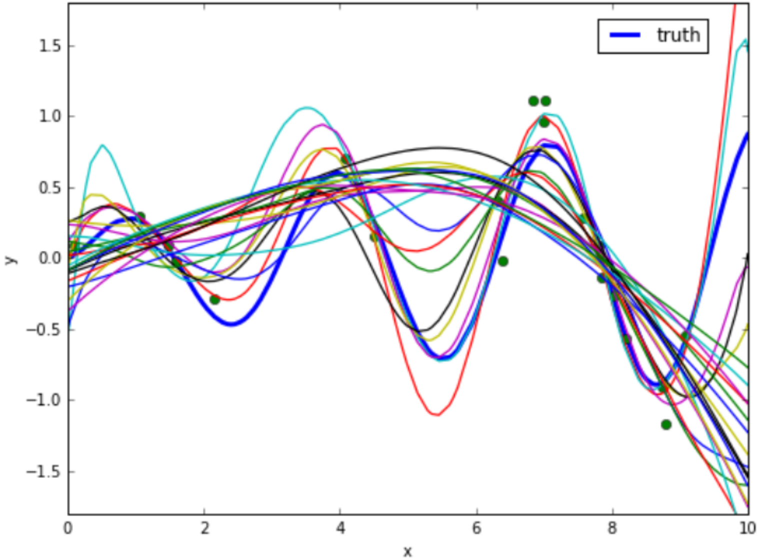
- generating an ensemble is a whole statistical subject in itself
- often we average the predictions over the models in an ensemble to improve performance $\hat{y}(x) = \frac{1}{M}\sum_{i=1}^{M}\hat{y}^{(i)}(x)$



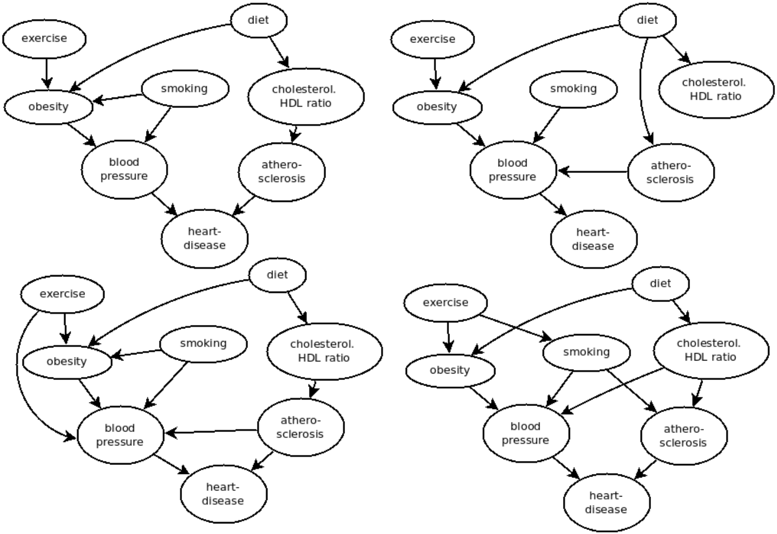
Ensembles: Large Data



Ensembles: Small Data



Ensemble of BayesNet Models



This Week's iPython Activity

- you will be given some iPython notebooks to test out concepts in learning
- the lab computers have Anaconda (with iPython/Jupyter loaded)
- you should also be able to log onto a Jupyter Notebook server (using your authcate and password) at: *https://jupyterhub.erc.monash.edu/hub/*
- the tutors can work you through it

Unit Schedule: Next Week

Module	Week	Content
1.	1	overview and look at projects
	2	(job) roles, and the impact
2.	3	data business models
	4	application areas and case studies
3.	5	characterising data and "big" data
	6	data sources and case studies
4.	7	resources and standards
	8	resources case studies
5	**9**	**data analysis theory**
	10	**data analysis process**
6.	11	issues in data management
	12	data management frameworks