

FIT5145 Introduction to Data Science

Module 5

Data Analysis Process

2019 Lecture 9

Monash University

Discussion

In the tutorial you used three different tools for data wrangling:

- ▶ DataWrangler
 - ▶ specialised Data Wrangling tool
 - ▶ intuitive Graphical User Interface (GUI)
 - ▶ no coding required!
- ▶ Python
 - ▶ general purpose open-source programming language
 - ▶ contains packages (Pandas) for manipulating data
- ▶ SAS
 - ▶ general purpose Data Analytics
 - ▶ strange syntax!
 - ▶ very widely used commercial product

Note that there are many other tools we could have used

- ▶ R, Matlab, Java, SPSS.

Unit Schedule: This Week

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

Learning Outcomes (Week 9)

By the end of this week you should be able to:

- ▶ Identify different data analysis problems in a data science project
- ▶ Fit linear regression and polynomial regression models to a given dataset
- ▶ Explain overfitting and underfitting of different models
- ▶ Comprehend bias and variance trade-off
- ▶ Comprehend the importance of “No Free Lunch Theorem”
- ▶ Explain what ensemble models are

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
- ▶ *How we’re teaching computers to understand pictures* by Fei Fei Li, at TED 2015

Implications of Computers that Learn

From [2014 TED talk](#) by Jeremy Howard

Examples: checkers (1956), [IBM Watson at Jeopardy](#) (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

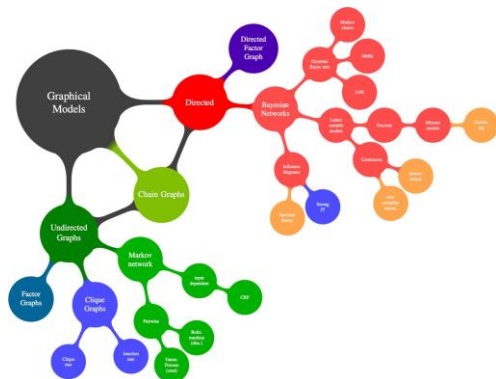
Theory of Data Analysis

Graphical Models

models of the structural aspects 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





Probabilistic Graphical Models



- ▶ represent a huge family of models, see left
- ▶ formally, they are used to represent probability and decision problems
- ▶ we use them to represent the data in a learning task

from David Barber, *Bayesian Reasoning and Machine Learning*, 2012

Node Types

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”);

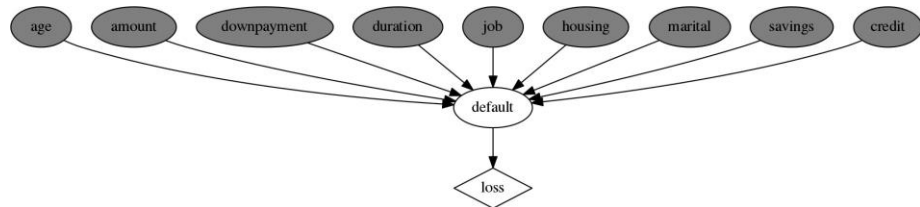
Known variable: no arcs generally, but may show if a related graph has them

Decision: connect to if variable used when making decision;

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

Simple Prediction Task:

Housing Loan Default

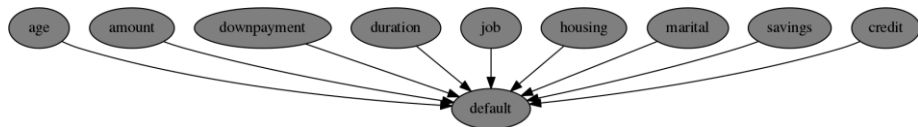


Task is to predict whether an unknown value:

- ▶ 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

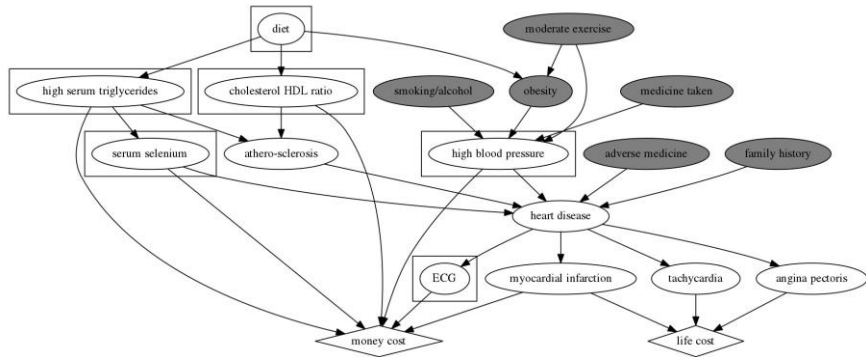


In order to **learn a model**,

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

Complicated Prediction Task:

Heart Disease Diagnosis

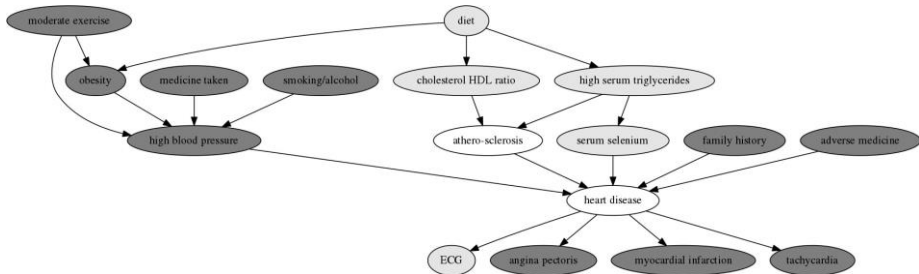


Model contains many variables that link to one another in complicated ways, (called a Bayesian Network)

- ▶ 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
- ▶ in healthcare, segments are called **cohorts**
 - ▶ used for patient management and staff organisation
- ▶ **but how do you do the grouping?**

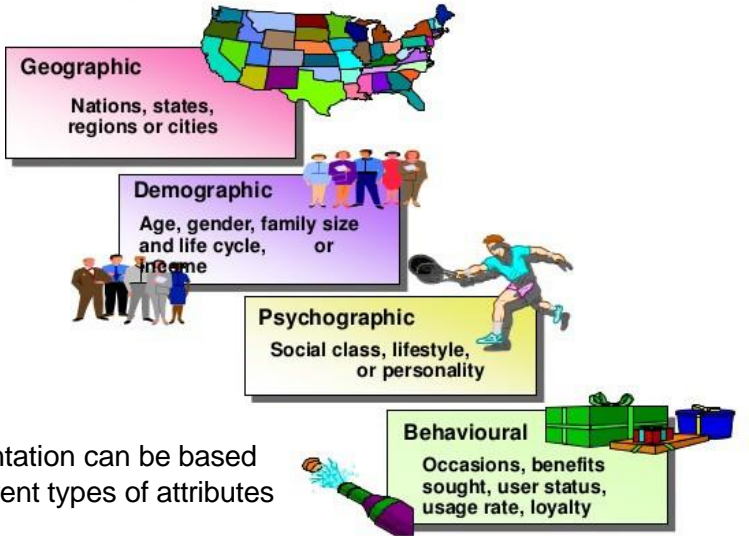


Example segmentation:

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

Market Segmentation

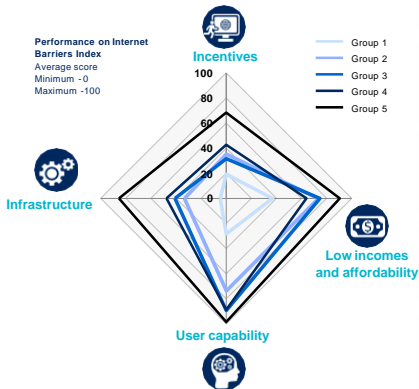
Bases for Segmenting Consumer Markets



Segmentation can be based on different types of attributes

With Already High Mobile Penetration in More Developed / Affluent Countries... New Users in Less Developed / Affluent Countries Harder to Garner, per McKinsey

Countries fall into one of 5 groups based on barriers they face to Internet adoption



Group 1: High barriers across the board; offline populations that are young, rural, and have low literacy

Countries: Bangladesh, Ethiopia, Nigeria, Pakistan, Tanzania
Offline population, 2014: 548 million
Internet penetration, 2014: 18%

Group 2: Medium to high barriers with larger challenges in incentives and infrastructure; mixed demographics

Countries: Egypt, India, Indonesia, Philippines, Thailand
Offline population, 2014: 1,438 million
Internet penetration, 2014: 20%

Group 3: Medium barriers with greatest challenge in incentives; rural and literate offline populations

Countries: China, Sri Lanka, Vietnam
Offline population, 2014: 753 million
Internet penetration, 2014: 49%

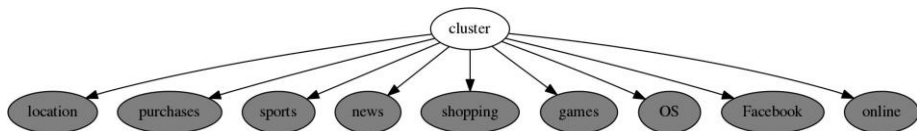
Group 4: Medium barriers with greatest challenge in low incomes and affordability; offline populations predominantly urban / literate / low income

Countries: Brazil, Colombia, Mexico, South Africa, Turkey
Offline population, 2014: 244 million
Internet penetration, 2014: 52%

Group 5: Low barriers across the board; offline populations that are highly literate and disproportionately low income and female

Countries: Germany, Italy, Japan, Korea, Russia, USA
Offline population, 2014: 147 million
Internet penetration, 2014: 82%

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 Analytics Is”](#) starting at 1:30

FLUX Question

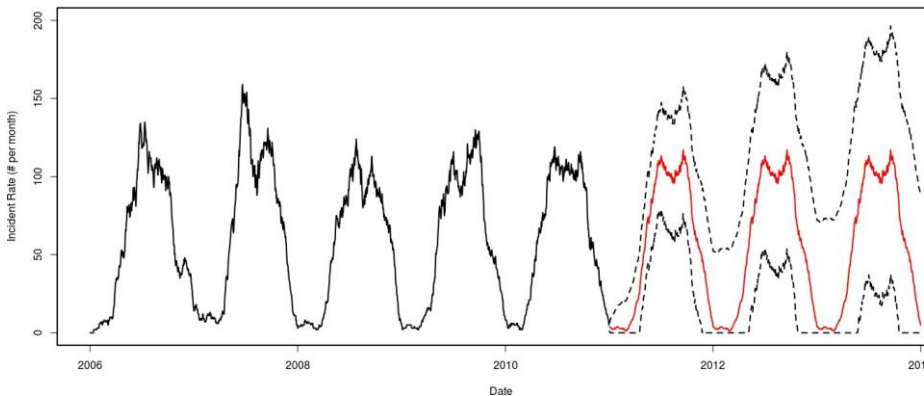
Which one of the following tasks is not a segmentation task?

- A. Group all the shopping items available on web.
- B. Identification of areas of similar land use in an earth observation database
- C. Weather prediction based on last month's temperature



Time Series Forecasting

Projected bicycle collision rates in Montreal



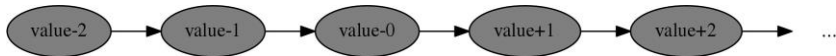
from [bayesianbiologist](#)

Time Series: 1st Order

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



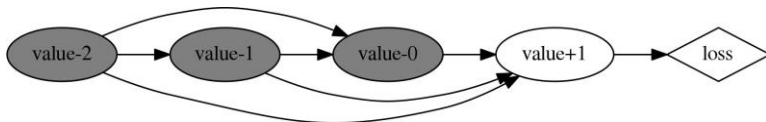
Training data consists of one or more series of values:



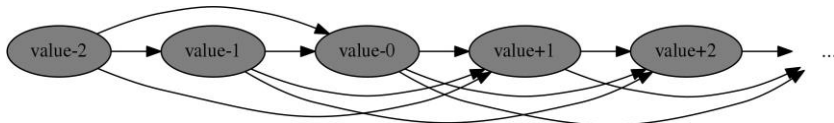
Time Series: 3rd Order

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

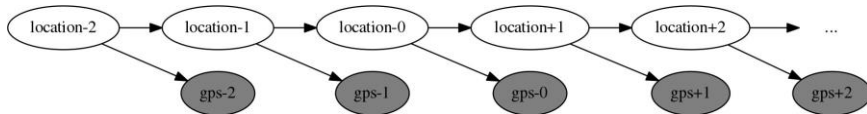


Training data is again just sequences of data:



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

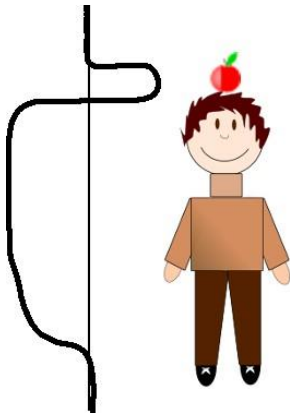
- ▶ to evaluate the quality of results derived from learning, we need notions of value
- ▶ so we will review quality and value

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

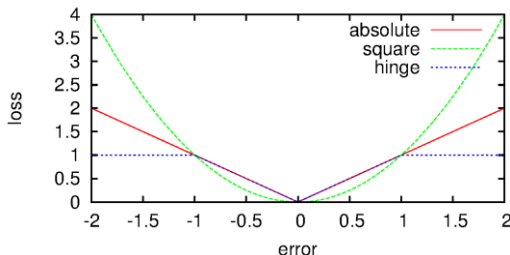
- ▶ may be the quality of your prediction
- ▶ may be the consequence of your actions
(making a prediction is a kind of action)
- ▶ can be measured on a positive or negative scale

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.:

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

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

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

Data Analysis Algorithms Regression

From [The Elements of Statistical Learning](#)

by T. Hastie, R. Tibshirani and J. Friedman

Regression

What is Regression?

- Look for relationships amongst variables

Real World Example:

- Identify the relation between salary and experience, education, and role

Terminology

Variables can be:

- ▲ *Independent Variables/Inputs/Predictors*

E.g., experience, education, role

- ▲ *Dependent Variables/Outputs/Responses*

E.g., salary of employee

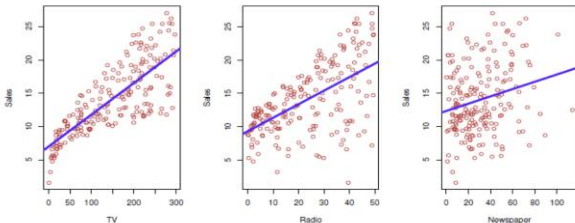
Observation is a data point, row, or sample in a dataset

- ▲ **E.g.**, an employee's salary, experience, education, role.

When Use Regression

▲ To determine how multiple variables are related
E.g., determine *if* and *to what extent* the experience or education impact salaries

▲ To predict a value
E.g., predict electricity consumption given the outdoor temperature, time of day, and number of residents in that household



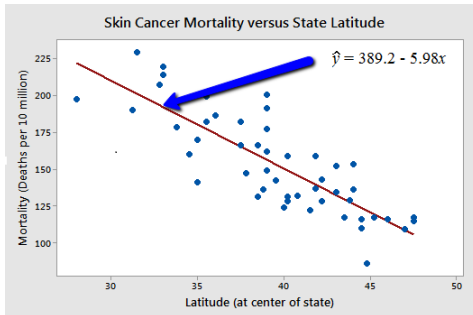
Example: Sales \sim TV, Radio, newspaper

Simple Linear Regression (two-dimensional space):

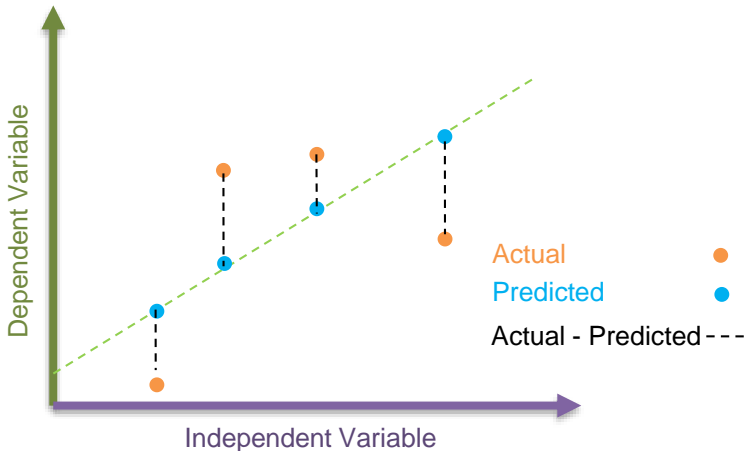
Regression fits a very simple equation to the data:

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

Here $\hat{y}(x; \vec{a})$ is prediction for y at the point x using the model parameters $\vec{a} = (a_0, a_1)$, i.e. the intercept and slope terms.



Best Fitting Line



Aim is that the predicted response, be as close as possible to the actual response.

Calculating Parameters- Intuition

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

- 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:

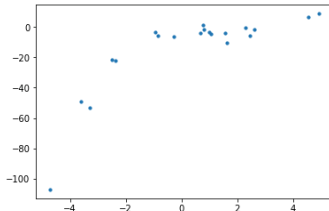
$$\text{mean square error} = MSE_{train} = \frac{1}{N} \sum_{i=1}^N (\hat{y}(x_i; \vec{a}) - y_i)^2$$

- Compare the derivatives to zero

Real-world Example (Python)

```
#import required packages
import numpy as np
import matplotlib.pyplot as plt

#provide data
np.random.seed(0)
x = 2 - 3 * np.random.normal(0, 1, 20)
y = x - 2 * (x ** 2) + 0.5 * (x ** 3) + np.random.normal(-3, 3, 20)
plt.scatter(x,y, s=10)
plt.show()
```



Real-world Example (Python)

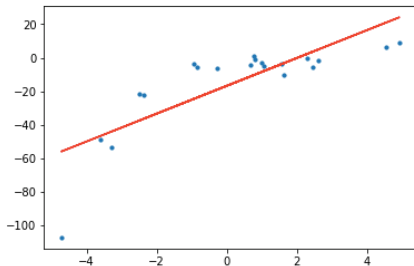
```
#import required packages
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression

#provide data
np.random.seed(0)
x = 2 - 3 * np.random.normal(0, 1, 20)
y = x - 2 * (x ** 2) + 0.5 * (x ** 3) + np.random.normal(-3, 3, 20)

# transforming the data to include another axis
x = x[:, np.newaxis]
y = y[:, np.newaxis]

#create a linear regression model
model = LinearRegression()
model.fit(x, y)
y_pred = model.predict(x)

#display the best fit line
plt.scatter(x, y, s=10)
plt.plot(x, y_pred, color='r')
plt.show()
```



▲ Linear regression is unable to capture the patterns in the data. This is an example of under-fitting.

▲ To overcome under-fitting, we need to increase the complexity of the model

Polynomial Regression

▲ Assume the polynomial relationship between the inputs and output.

▲ E.g., 10th order (aka degree) polynomial

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

FLUX Question

What is a polynomial regression?

- A. Fitting many lines to data.
- B. Fitting a curve defined by a polynomial function to data.
- C. Fitting a curve to a line.



Real-world Example (cont.)

```
#import required packages
import operator
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
from sklearn.metrics import mean_squared_error, r2_score
from sklearn.preprocessing import PolynomialFeatures

#provide data
np.random.seed(0)
x = 2 - 3 * np.random.normal(0, 1, 20)
y = x - 2 * (x ** 2) + 0.5 * (x ** 3) + np.random.normal(-3, 3, 20)

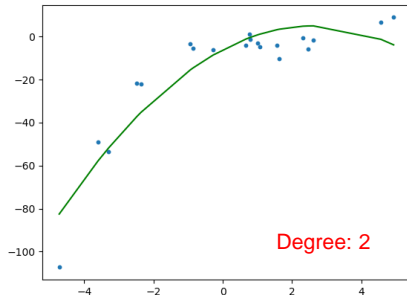
# transforming the data to include another axis
x = x[:, np.newaxis]
y = y[:, np.newaxis]

#create polynomial regression
polynomial_features = PolynomialFeatures(degree= 2)
x_poly = polynomial_features.fit_transform(x)

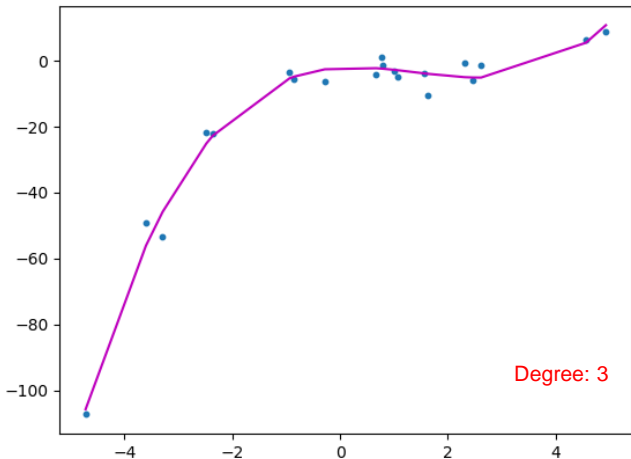
model = LinearRegression()
model.fit(x_poly, y)
y_poly_pred = model.predict(x_poly)

rmse = np.sqrt(mean_squared_error(y,y_poly_pred))
r2 = r2_score(y,y_poly_pred)
print(rmse)
print(r2)

plt.scatter(x, y, s=10)
# sort the values of x before line plot
sort_axis = operator.itemgetter(0)
sorted_zip = sorted(zip(x,y_poly_pred), key=sort_axis)
x, y_poly_pred = zip(*sorted_zip)
plt.plot(x, y_poly_pred, color='m')
plt.show()
```



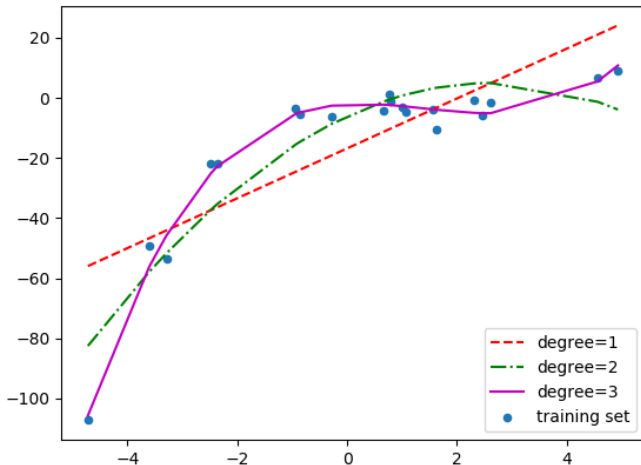
Polynomial Regression



FLUX Question



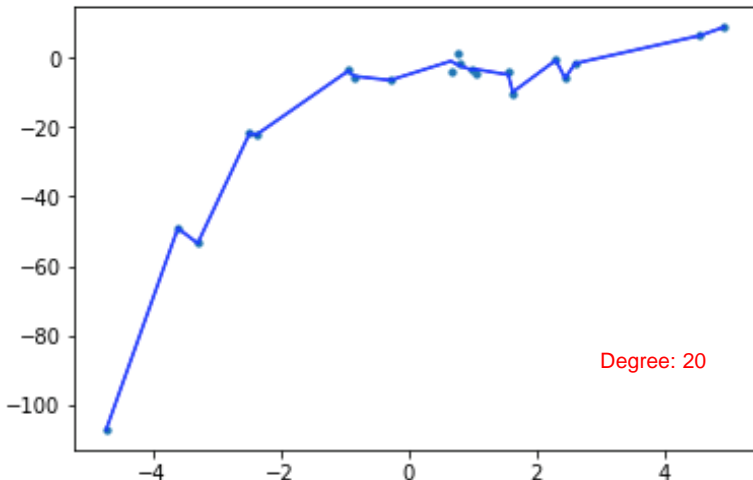
What is the best degree? 1, 2 or 3?



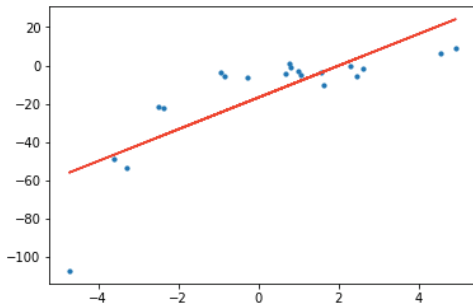
FLUX Question



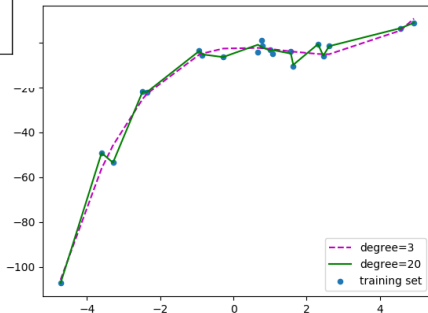
Is this fit better than previous fits?



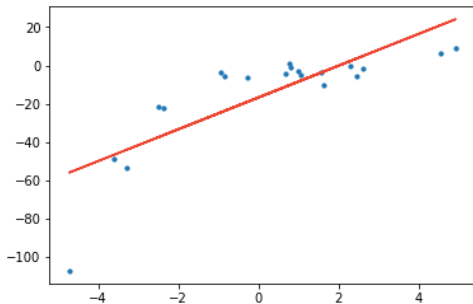
Underfitting and Overfitting



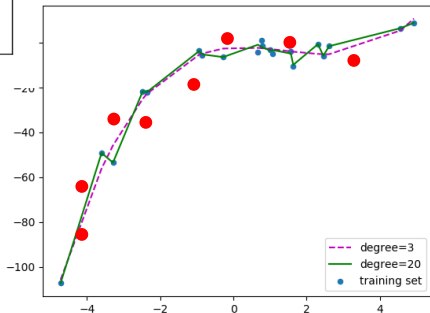
Under-Fitting



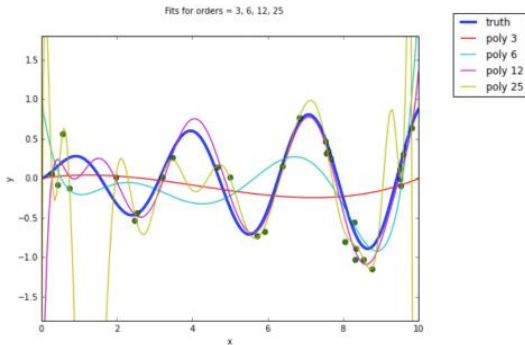
Underfitting and Overfitting



Over-Fitting



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.

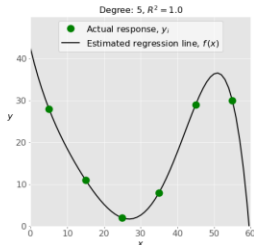
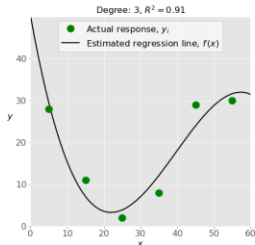
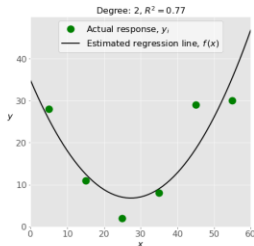
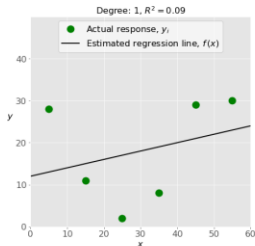
▲ This phenomenon is referred to as overfitting

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

FLUX Question

Which model could be well-fitted?



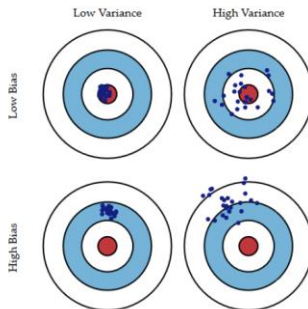
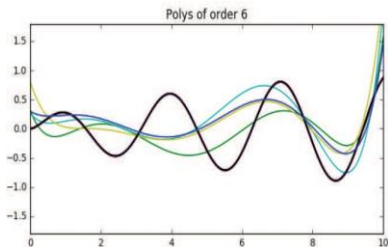
Theory of Data Analysis, cont.

Bias and Variance

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?

Bias and Variance

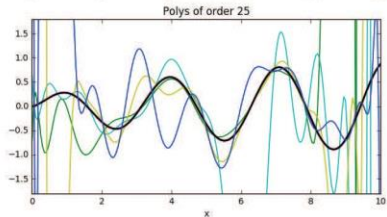
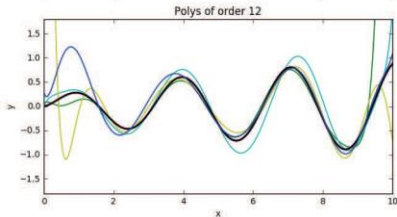
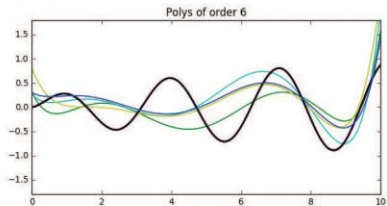
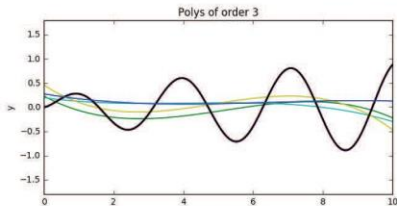


Bias: measures how much the prediction differs from the desired regression function.

Variance: measures how much the predictions for individual data sets vary around their average.

Bias-Variance Examples

Simple polynomials on different data of size 30



FLUX Question

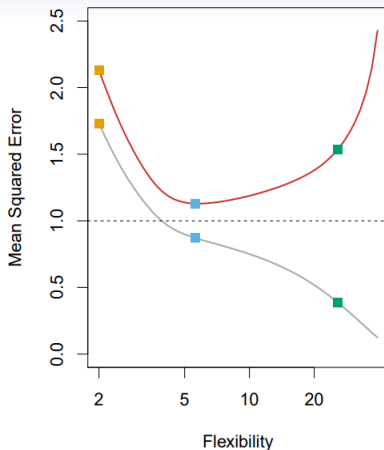
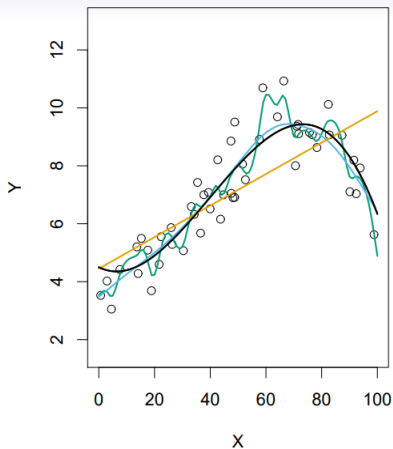
Which of the polynomials in the previous slide is a better model?

- A. Order 3
- B. Order 6
- C. Order 12
- D. Order 25



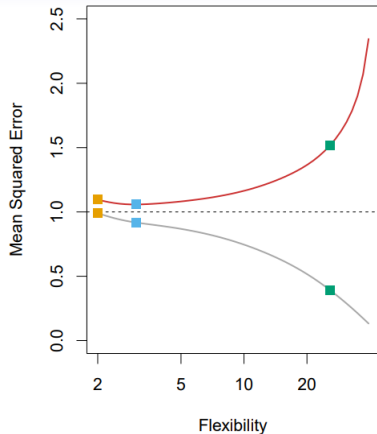
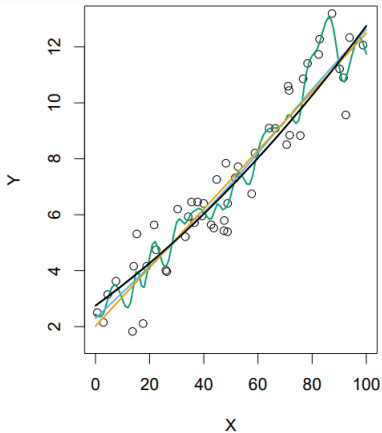
Bias vs Variance Trade-off

Scenario 1



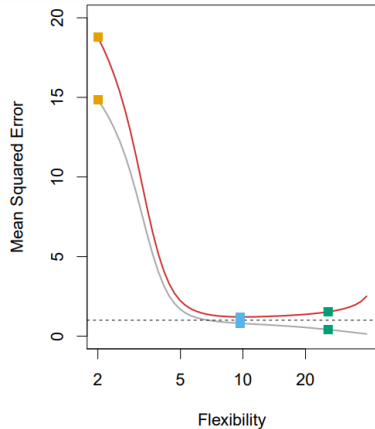
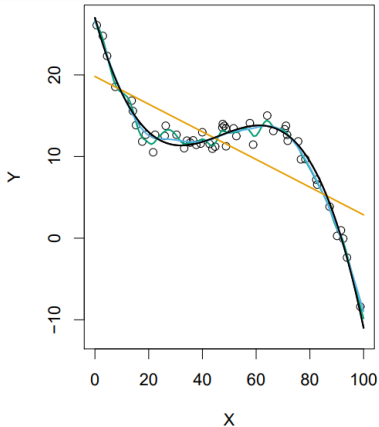
Bias vs Variance Trade-off

Scenario 2



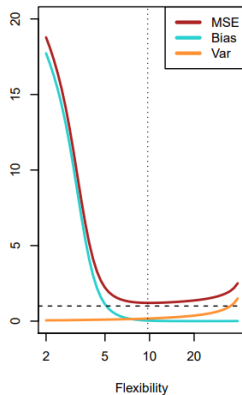
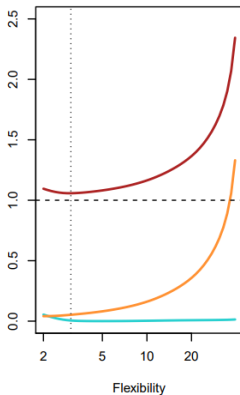
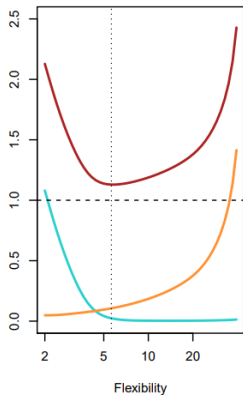
Bias vs Variance Trade-off

Scenario 3

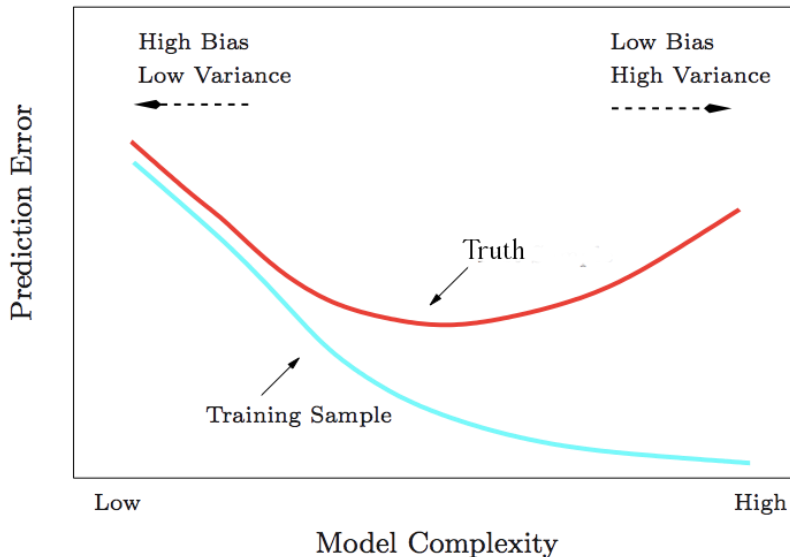


Bias vs Variance Trade-off

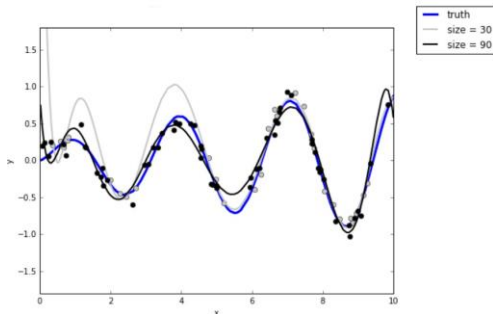
Optimum Degree



Bias-Variance Tradeoff



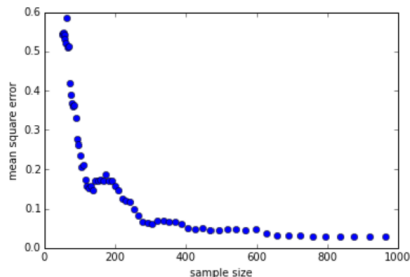
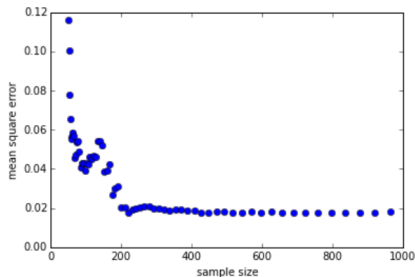
More Data Improves the Fit



- ▶ 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)

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

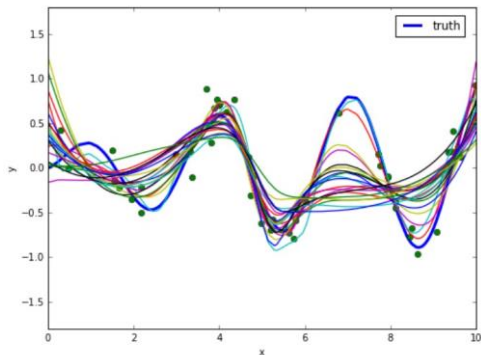
- ▶ there is no universally good machine learning algorithm (when one has finite data)

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**

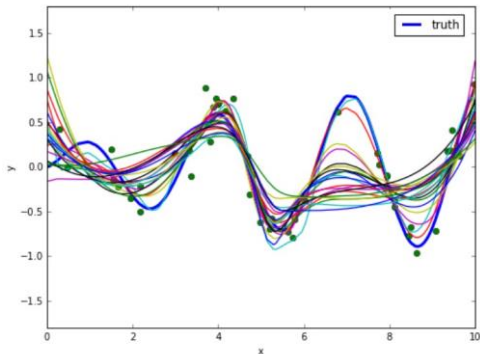
Ensembles

- ▶ 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

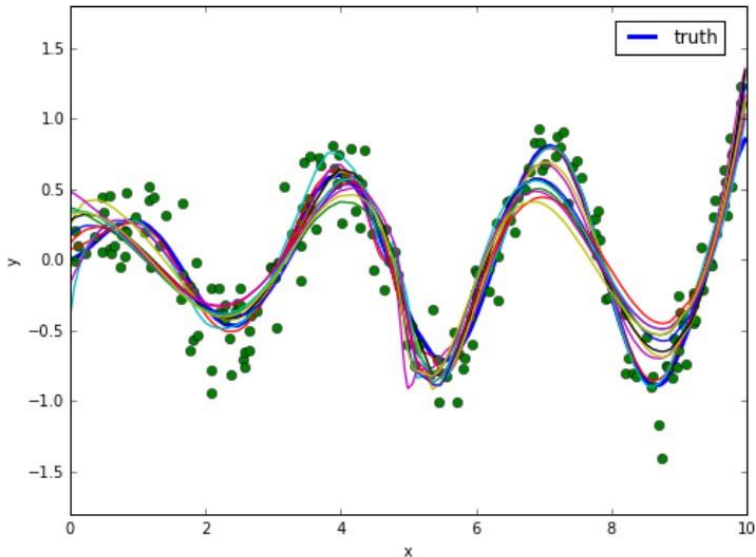


Ensembles (cont.)

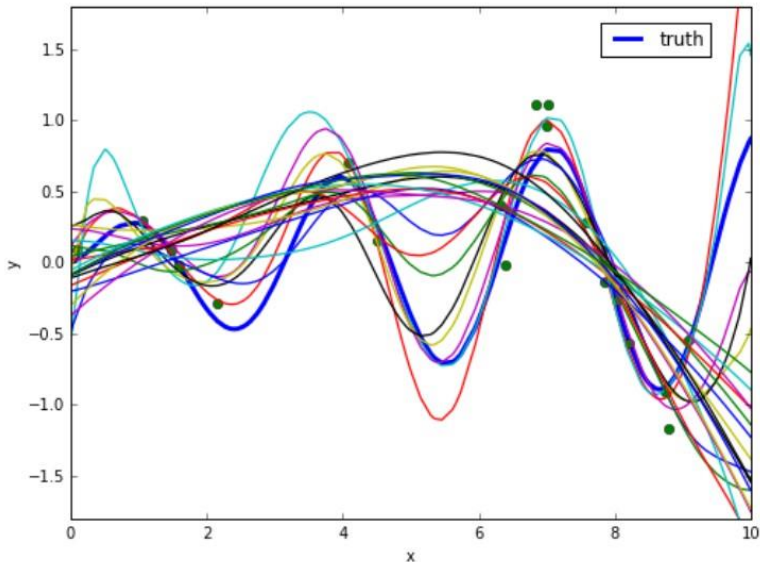
- ▶ **generating an ensemble** is a whole statistical subject in itself
- ▶ often we average the predictions over the models in an ensemble to improve performance $\hat{y}(x) = \frac{1}{M} \sum_{i=1}^M \hat{y}^{(i)}(x)$



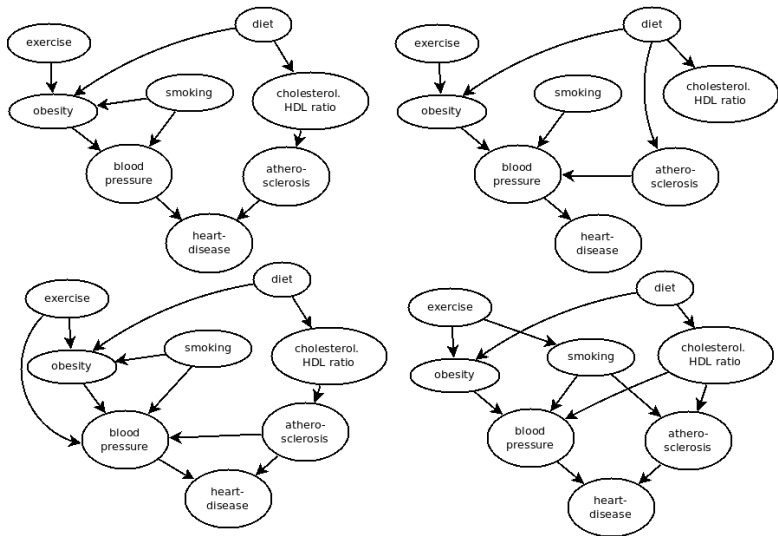
Ensembles: Large Data



Ensembles: Small Data



Ensemble of BayesNet Models



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 and case studies
	4	
3.	5	characterising data and "big" data data sources and case studies
	6	
4.	7	resources and standards resources case studies
	8	
5	9	data analysis theory data analysis process
	10	
6.	11	issues in data management GUEST SPEAKER & EXAM INFO
	12	