FIT1043

Week 1-7

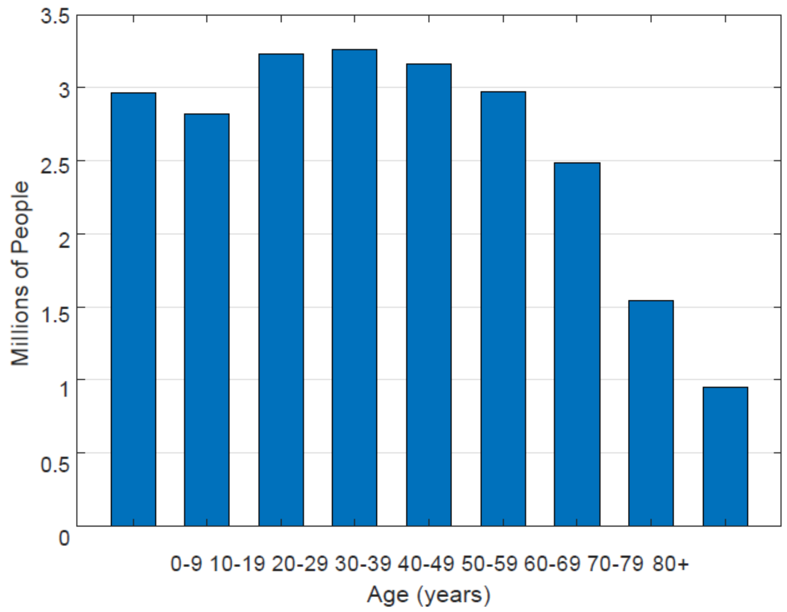
授课老师：Joe

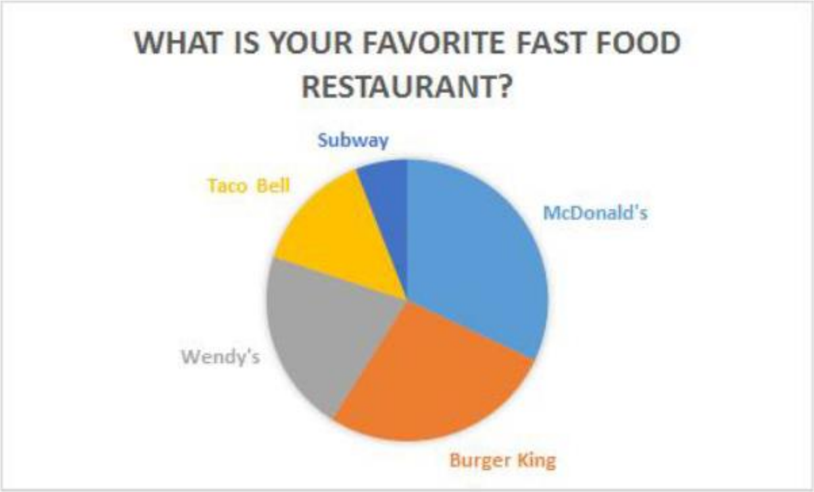
DATA VISUALISATION

- For categorical data, standard visualisations include:
  - Frequency tables
  - Bar graphs
  - Pie charts
- For numeric data (continuous and discrete), we can use:
  - Histograms
  - Box plots

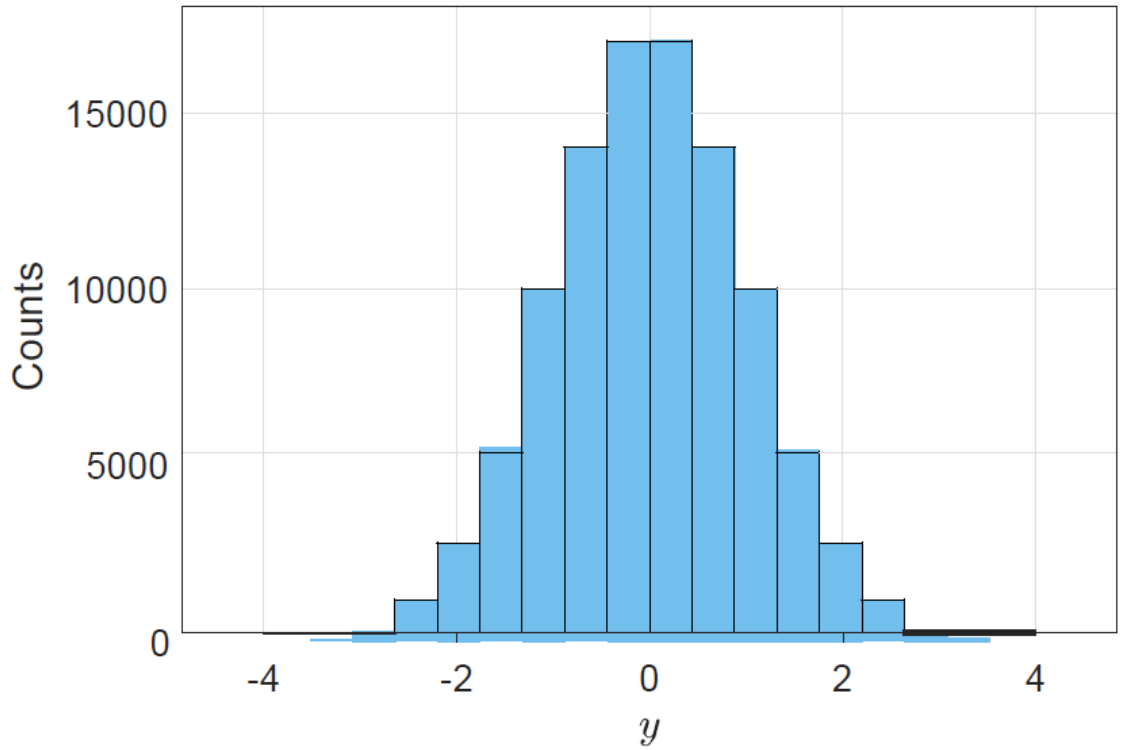
FIT1043

Age (years)	Number of People
0-9	2,967,425
10-19	2,818,778
20-29	3,231,395
30-39	3,265,526
40-49	3,164,712
50-59	2,977,883
60-69	2,488,396
70-79	1,540,373
80+	947,411





FIT1043



Descriptive statistics

Centrality

- Mean
- Mode
- Median

Which option is the Mean, Median and Mode of the following set of values respectively?
1,2,2,3,4,7,9

A. 4,2,3
B. 5,3,2
C. 4,3,3
D. 4,3,2

Type	Example	Result
Mean	(1+2+2+3+4+7+9) / 7	**4**
Median	1, 2, 2, **3**, 4, 7, 9	**3**
Mode	1, **2**, **2**, 3, 4, 7, 9	**2**

- The mean uses *all* the values of the sample
  - Any change to any sample changes the mean
  - The mean can be changed as much as desired by changing just one sample by a large enough amount
- The median uses at most two of the values of the sample

  Is very resistant to changes to the samples not in the middle

Percentiles

- More generally, we can define the percentiles
  - The p-th percentile is the value, $Q(\mathbf{y}, p)$ such that $p\%$ of the values of the sample are lower than $Q(\mathbf{y}, p)$
- The median is simply the 50th percentile, $Q(\mathbf{y}, 50)$
- Other important percentiles are the 1st and 3rd quartiles
  - i.e., the 25th and 75th percentiles

Spread

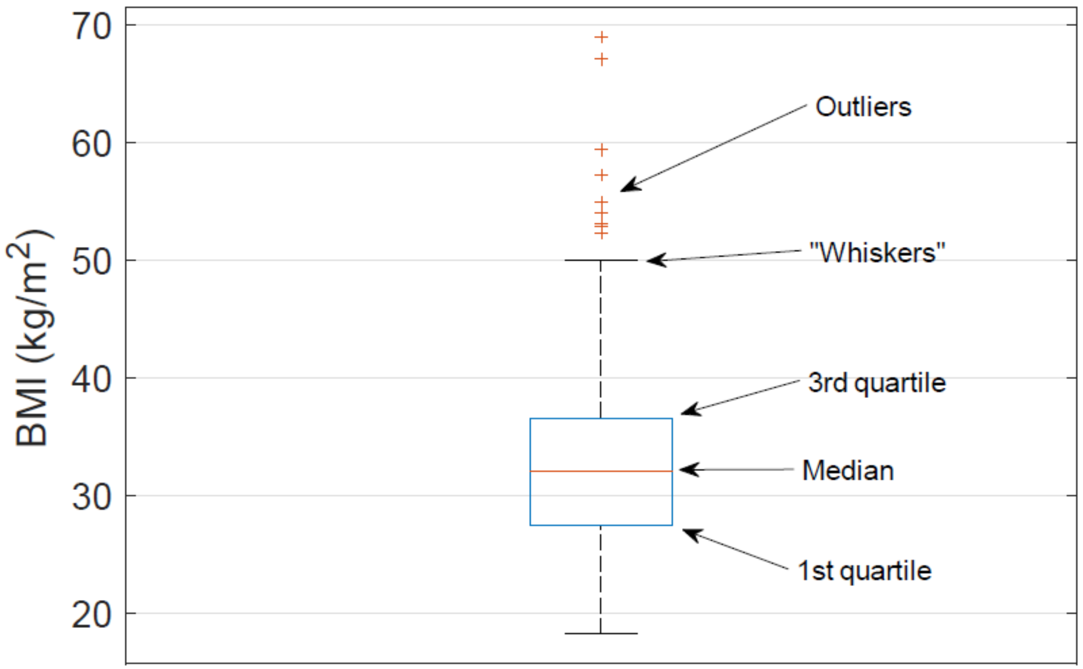
The most straightforward is the range

$$\mathrm{rng}(\mathbf{y}) = \max\{\mathbf{y}\} - \min\{\mathbf{y}\}$$

The most common measure of spread used is the sample standard deviation

$$s(\mathbf{y}) = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(y_j - \bar{y})^2}$$

FIT1043



70
60
50
40
30
20
BMI (kg/m^2)
Outliers
"Whiskers"
3rd quartile
Median
1st quartile

Pearson correlation measures linear association

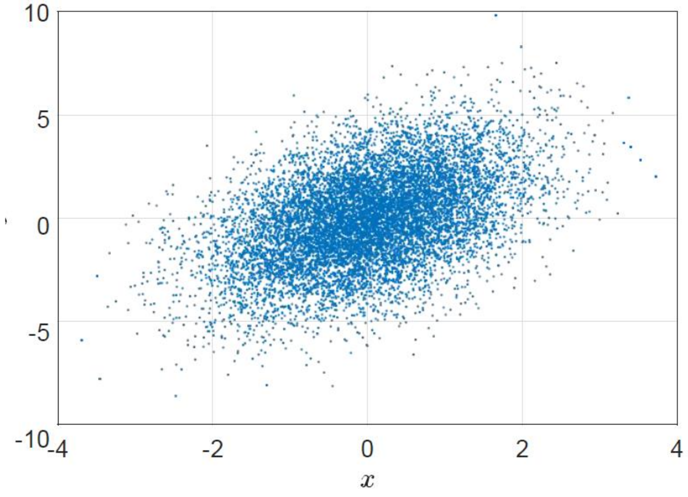
$$R(\mathbf{x},\mathbf{y}) = \frac{\sum_{j=1}^{n}(x_j - \bar{x})(y_j - \bar{y})}{n\, s(\mathbf{x}) s(\mathbf{y})}$$

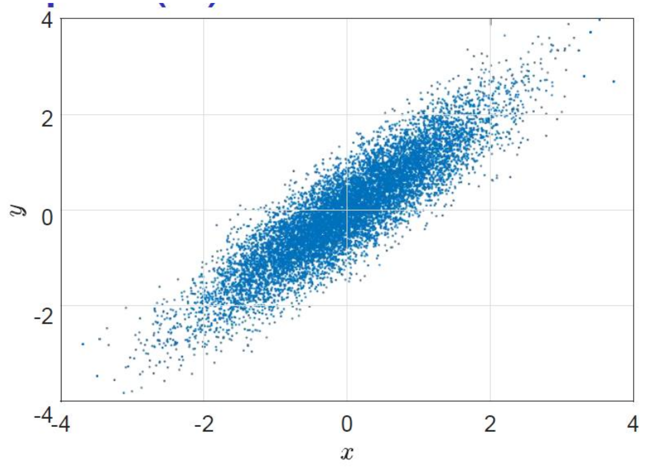
- Correlation is always between -1 (completely negatively correlated) and 1 (completely positively correlated)
- A correlation of zero implies there is no linear association

$\Rightarrow$ does not imply no non-linear association

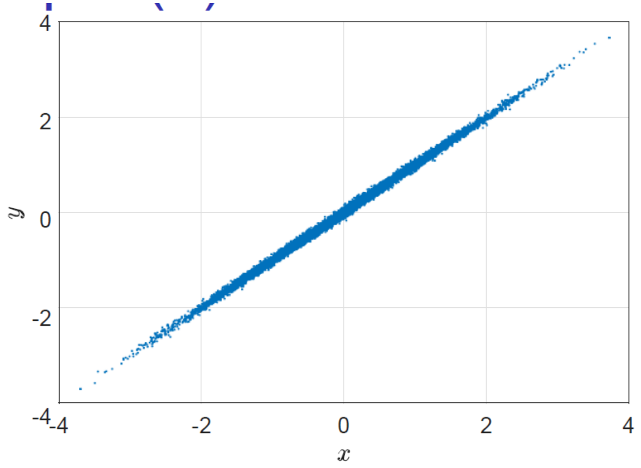
Remember: correlation not equal causation!

FIT1043





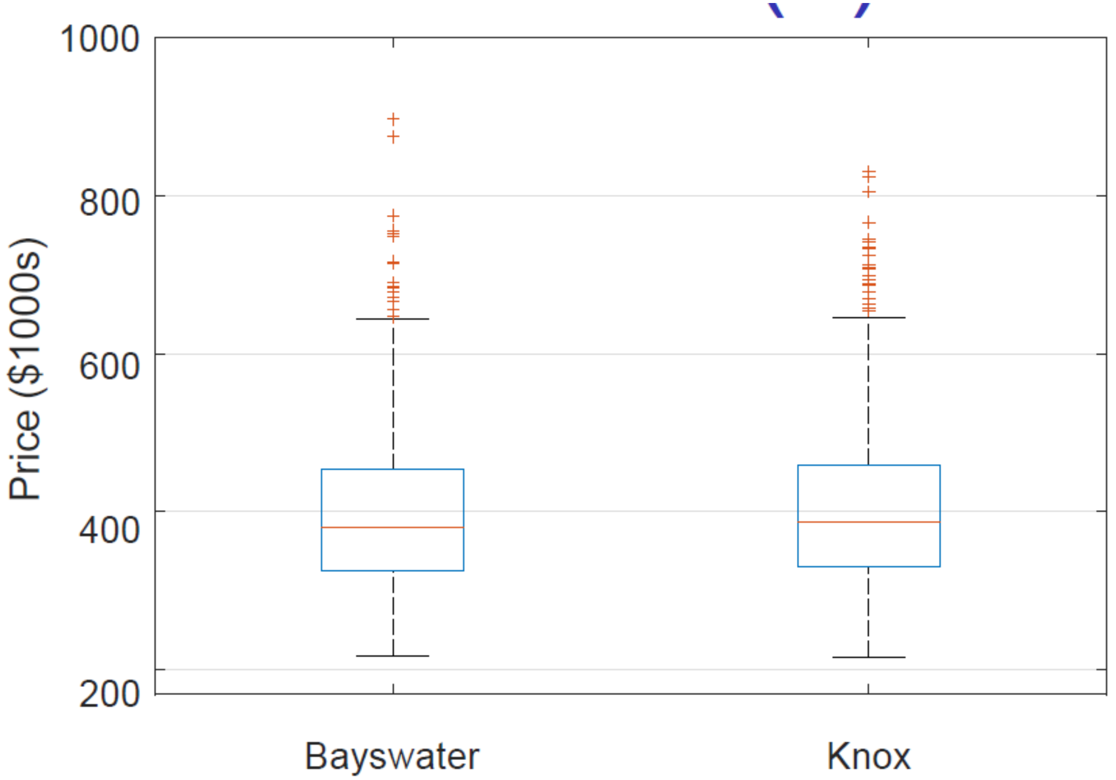
R = 0.9



R ≈ 0.999

- If **x** is categorical, and **y** is numeric, how to visualise?
- A standard approach is the side-by-side boxplot
  - Divide the data between categories, then plot boxplots for each group
  - Do the boxplots look different?
- If **x** and **y** are both categorical, we can use a side-by-side bargraph instead
  - Are the distributions/bargraphs different between categories? If so, there is a possible association

FIT1043



DATA WRANGLING

Sources of Data Quality Issues

- Interpretability issue
- Data format issue
- Inconsistent and faulty data
- Missing and incomplete data
- Outliers
- Duplicates

Dirty data

Mark Johnson, 31, 21/Aug/1985, 180, M, 0433010010, Melbourne VIC

Mr. Christian, Peter, 34, 21-09-1982, , M, 0433010118, Sydney NSW

Ethan Steedman, 32, 01/01/1982, 170, M, 0433210019, Sydney NSW

FIT1043

Inconsistency

- common cases:
  - upper vs. lower case
  - inconsistency in domain value representation, e.g., 0 vs. No, 1 vs. Yes
- detecting and fixing
  - investigate unique domain values (`unique()`)
  - make the representation consistent, e.g., replace

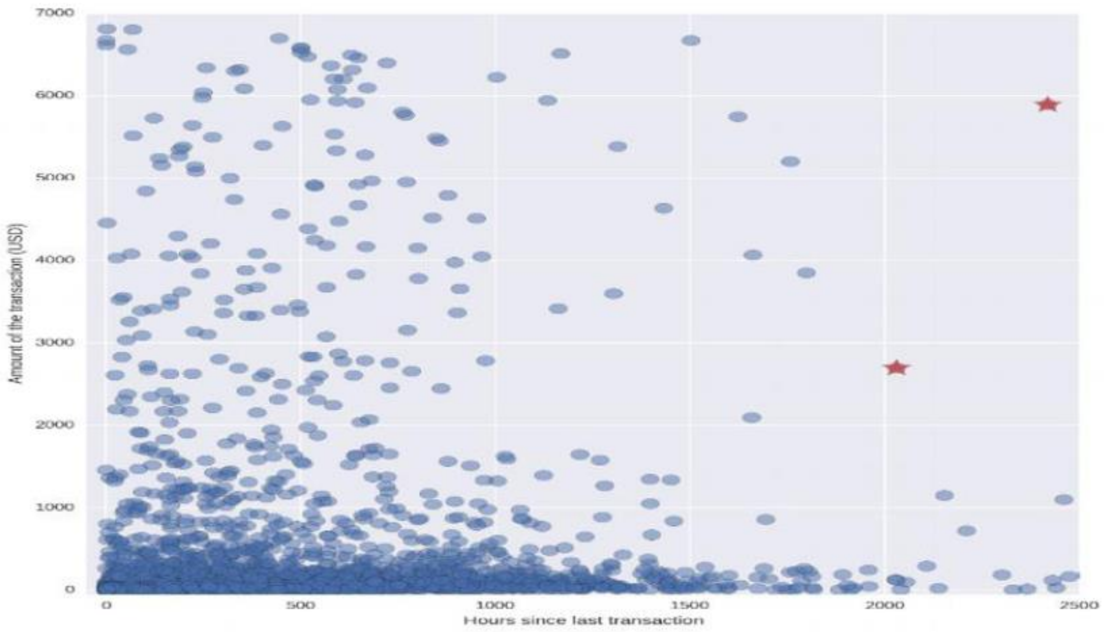
Misspelling

- investigate unique domain values (`unique()`)
- string matching
  - calculate domain value frequencies (`value_counts()`)
  - for all values, find matches for the infrequent values
  - replace infrequent values with the best match (if it exists) from the more frequent values.

Missing values

```
32,1,1,95,0,?,0,127,0,.7,1,?,?,1
34,1,4,115,0,?,?,154,0,.2,1,?,?,1
35,1,4,?,0,?,0,130,1,?,?,?,7,3
36,1,4,110,0,?,0,125,1,1,2,?,6,1
38,0,4,105,0,?,0,166,0,2.8,1,?,?,2
38,0,4,110,0,0,0,156,0,0,2,?,3,1
38,1,3,100,0,?,0,179,0,-1.1,1,?,?,0
38,1,3,115,0,0,0,128,1,0,2,?,7,1
38,1,4,135,0,?,0,150,0,0,?,?,3,2
38,1,4,150,0,?,0,120,1,?,?,?,3,1
40,1,4,95,0,?,1,144,0,0,1,?,?,2
```

Outliers

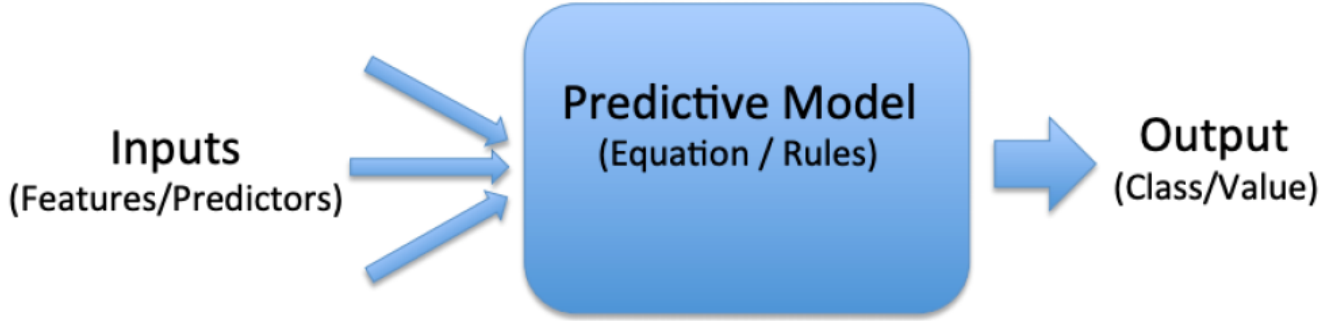

7000
6000
5000
4000
3000
2000
1000
0
Amount of the transaction (USD)
0
500
1000
1500
2000
2500
Hours since last transaction

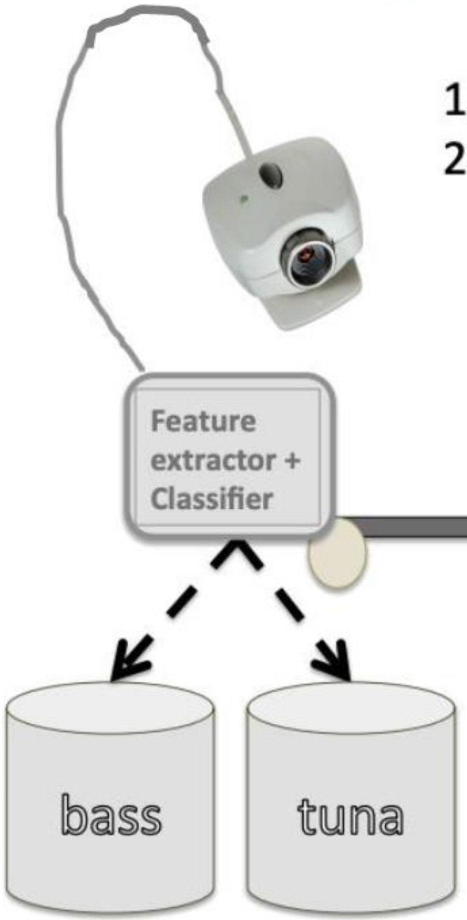
Data Analysis Theory

Predictive Models

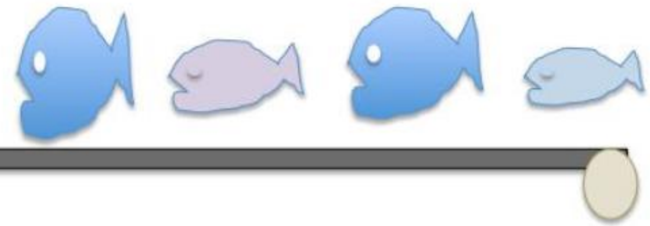
A predictive model is any model that makes a prediction

- Usually based on a set of features describing an object.
- The prediction could be:
    - A binary outcome (spam, not-spam)
    - Categorical (bass, tuna, other)
    - A real value (the age of the fish)
    - A vector of real values (probability of bass, tuna)
    - Etc.



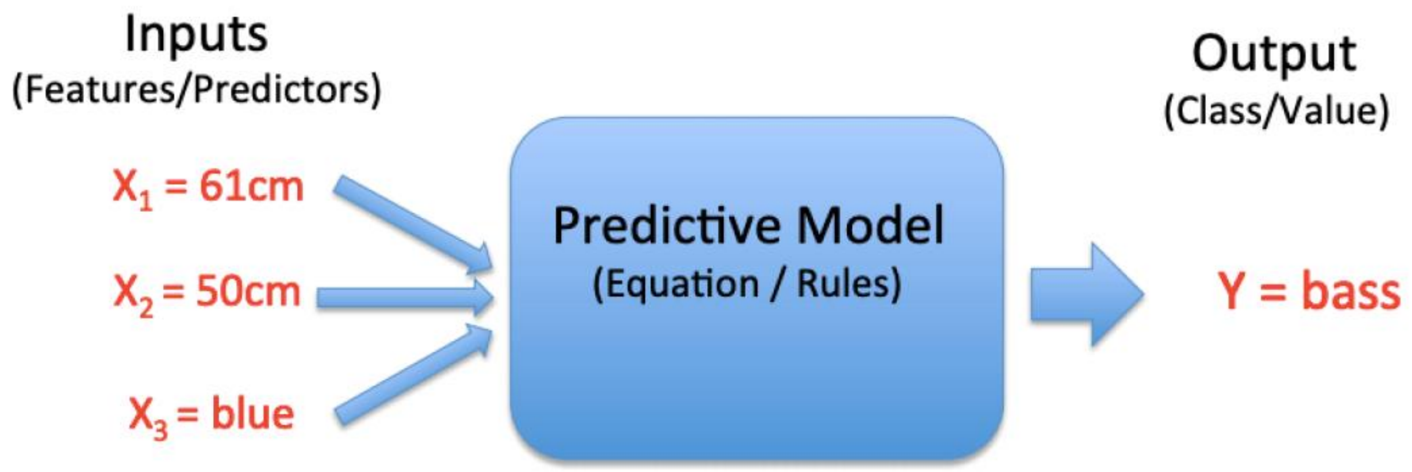


1. Install camera to capture image of fish
2. Compute **features** of fish from image:
   - X_1: length
   - X_2: width
   - X_3: colour



3. Learn rule(s) for classifying fish
   - if $X_1/X_2<1.4$ and X_3=blue => **bass**
   - if $X_1/X_2>2.1$ and X_3=(lightblue,pink) => **tuna**
   - else => **unknown**

FIT1043

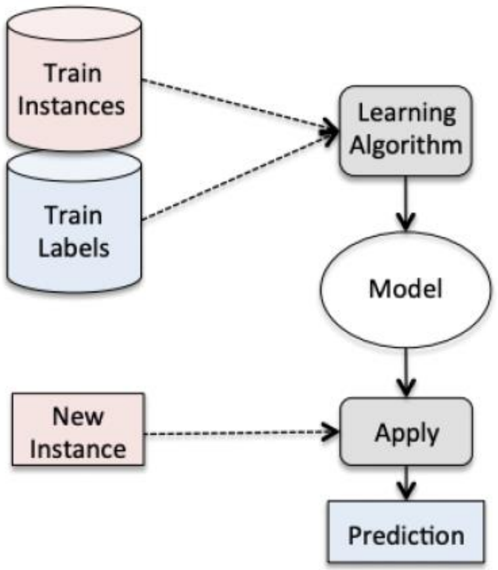

Inputs
(Features/Predictors)
X1 = 61cm
X2 = 50cm
X3 = blue
Predictive Model
(Equation / Rules)
Output
(Class/Value)
Y = bass

FIT1043

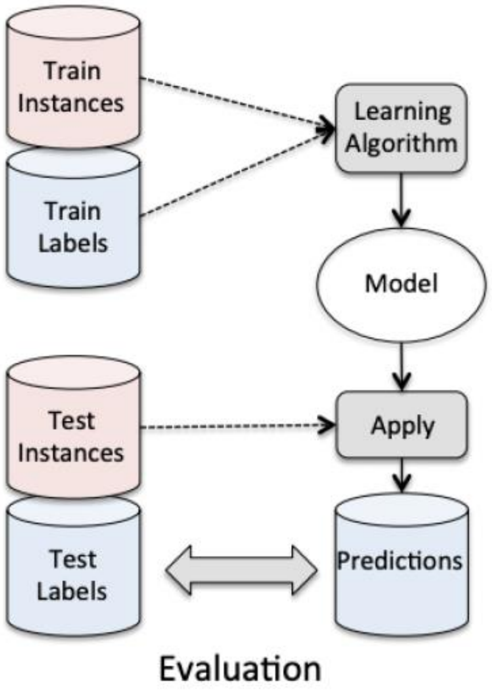
- If the predicted value is binary/categorical we usually refer to the model as a classifier
- If it predicts real values we refer to it as regression

FIT1043

Instance	X1 = length	X2 = width	X3 = colour	Y = class
	55	51	blue	bass
	65	23	pink	tuna
	67	54	blue	bass
	54	20	light-blue	tuna
	62	26	pink	tuna
	44	62	blue	bass
	47	55	light-blue	bass
	73	31	pink	tuna
	54	48	light-blue	bass
	57	23	light-blue	tuna

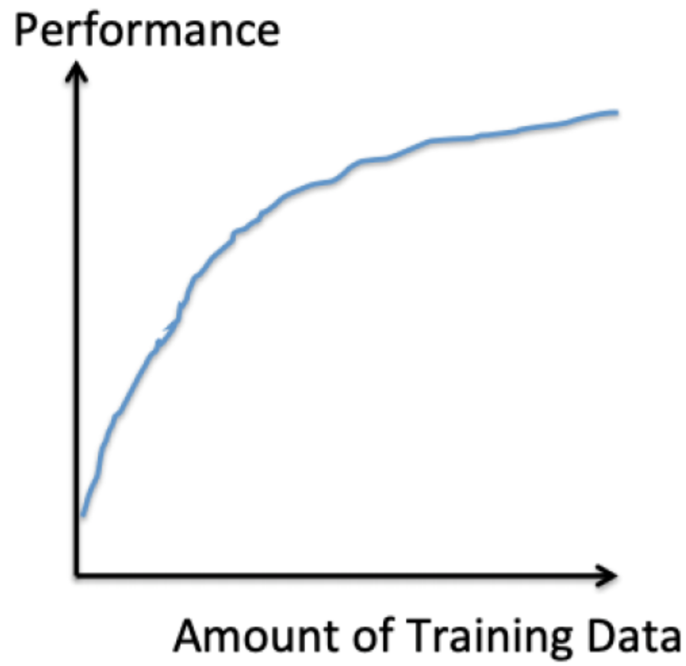

Train Instances
Train Labels
Learning Algorithm
Model
New Instance
Apply
Prediction

How can we decide which model is better?



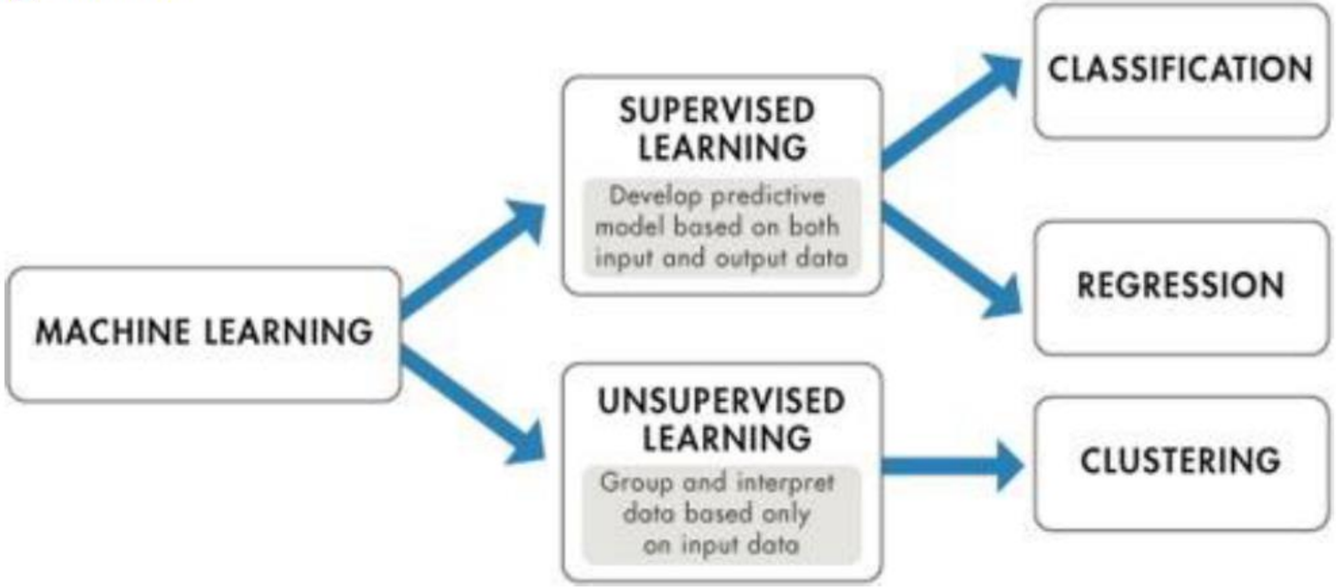
Generally:

- The more training data the better the test performance

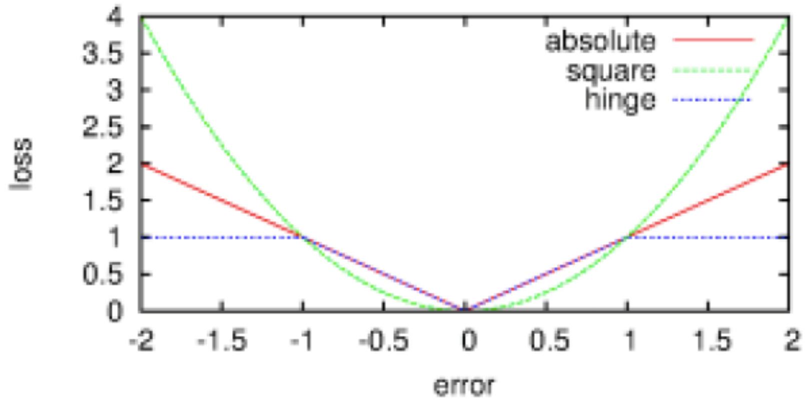


FIT1043

Brownlee, J. (2019). Supervised and Unsupervised Machine Learning Algorithms



Quality is a Function of Error



Error measures the distance between the prediction and the actual value

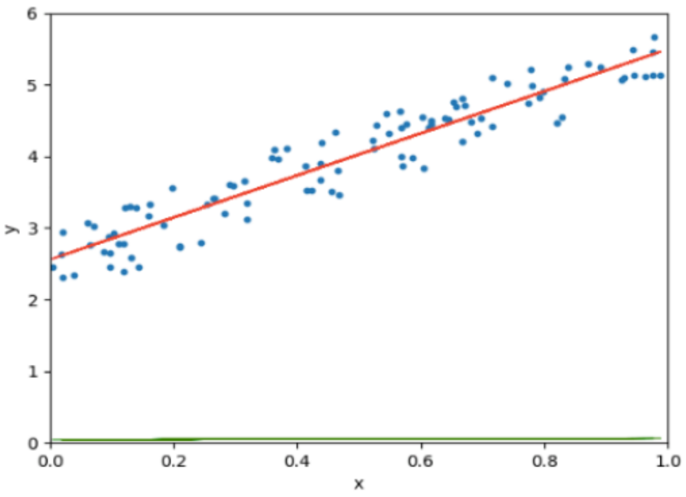
- “0” means no error, prediction was exactly right
- We can convert error to a measure of quality using a loss function, e.g.:

absolute-error(x) = |x|
square-error(x) = x ∗ x
hinge-error(x) = |x| if |x| ≤ 1
1 otherwise

Regression fits a very simple equation to the data:

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

- Data is shown with blue dots, red line is the "linear fitted model"



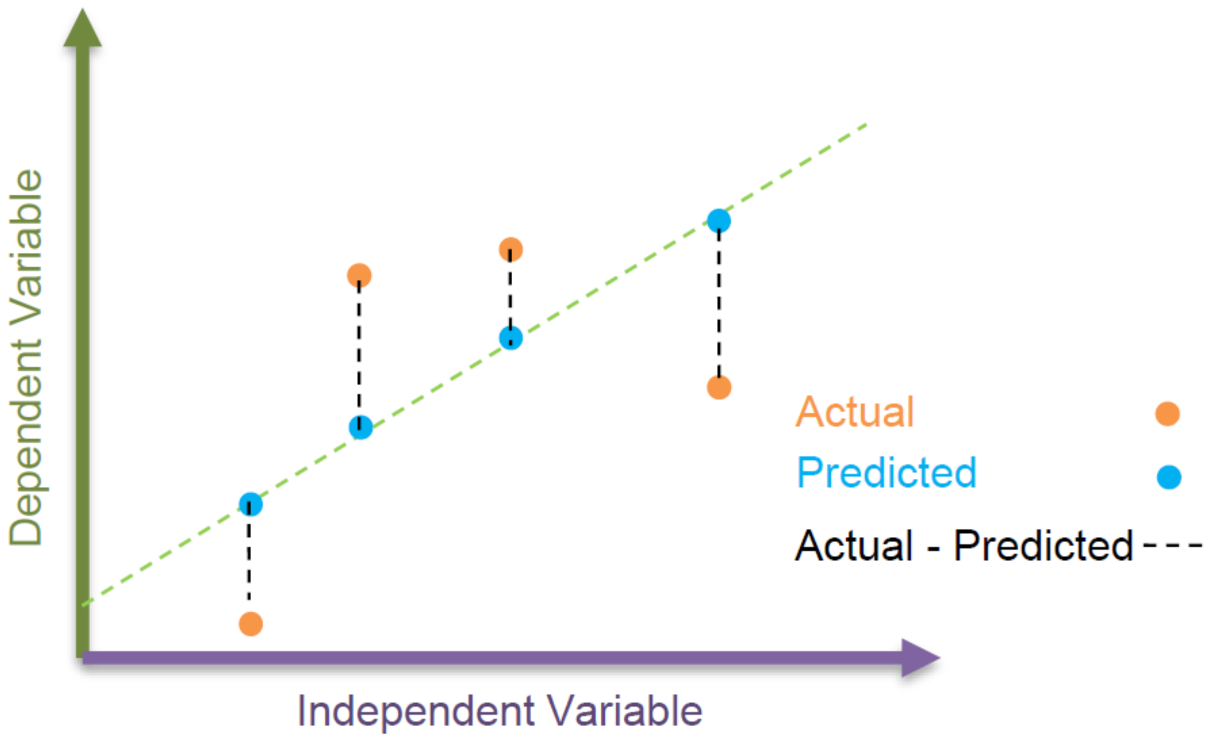
- Here $\hat{y}(x; \vec{a})$ is the for prediction for y at the point x using the model parameters $\vec{a}$ = (a_0, a_1), i.e. the intercept and slope terms.
- Given some data pairs (x_1, y_1), ...,(x_N, y_N), we fit a model by finding the vector $\vec{a}$ that minimises the loss function:

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

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

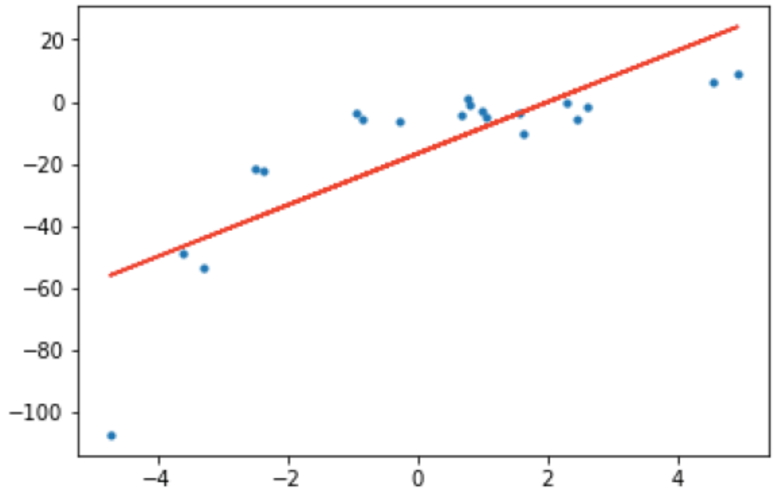
$$\hat{y}(x;\vec{a}) = a_0 + a_1 x + a_2 x^2 + ... a_9 x^9 + a_{10} x^{10} = \sum_{i=0}^{10} a_i x^i$$

Best Fitting Line

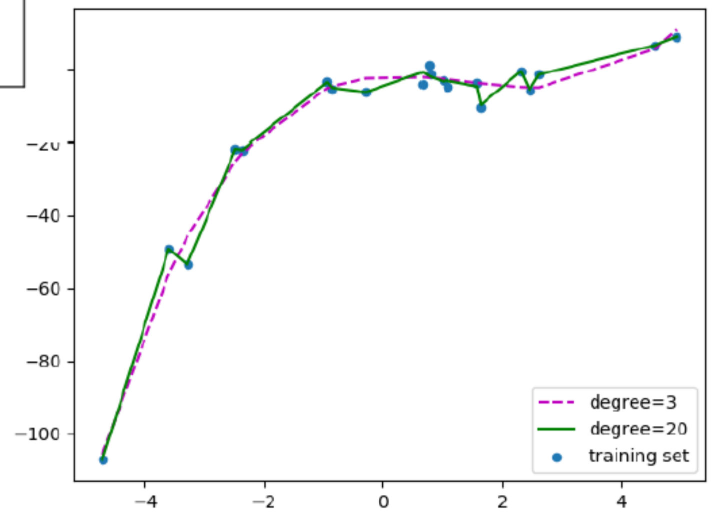


Aim is that the predicted response, be as close as possible to the actual response.

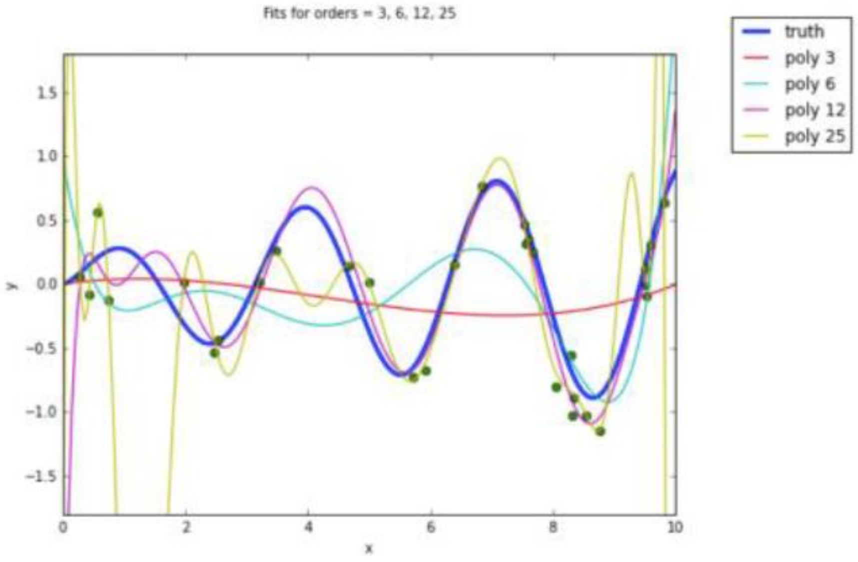
Underfitting and Overfitting



Under-Fitting



Overfitting



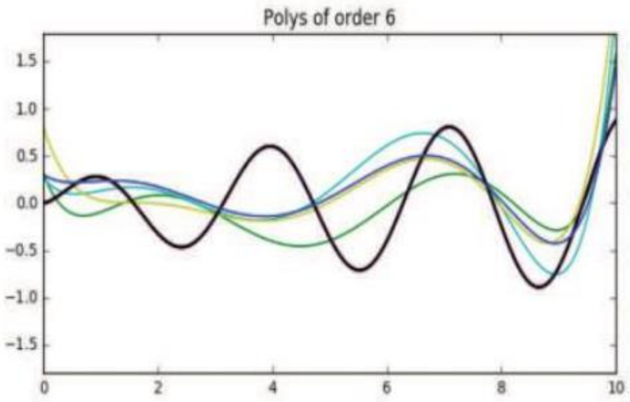
The more parameters a model has, the more complicated a curve it can fit.

- 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

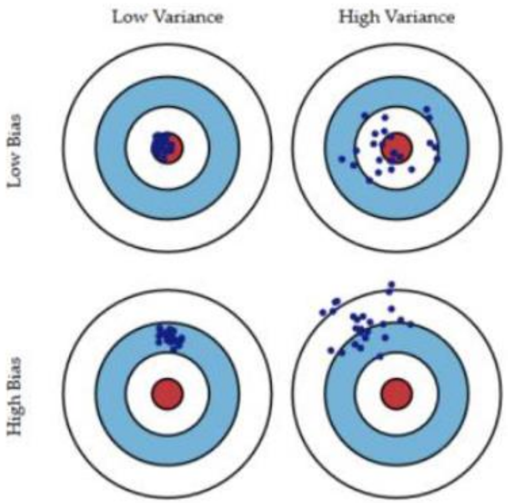
Training Set and Test Set

- 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

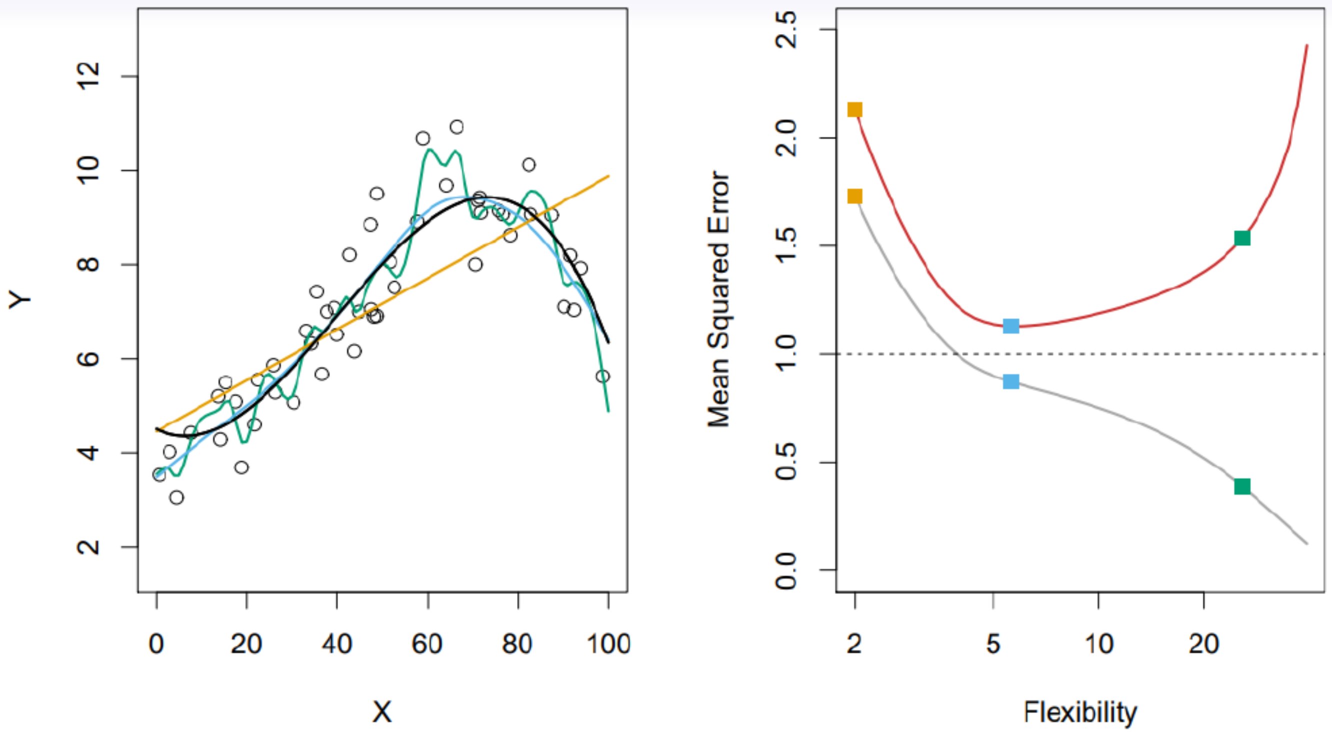


Different data sets of size 30.

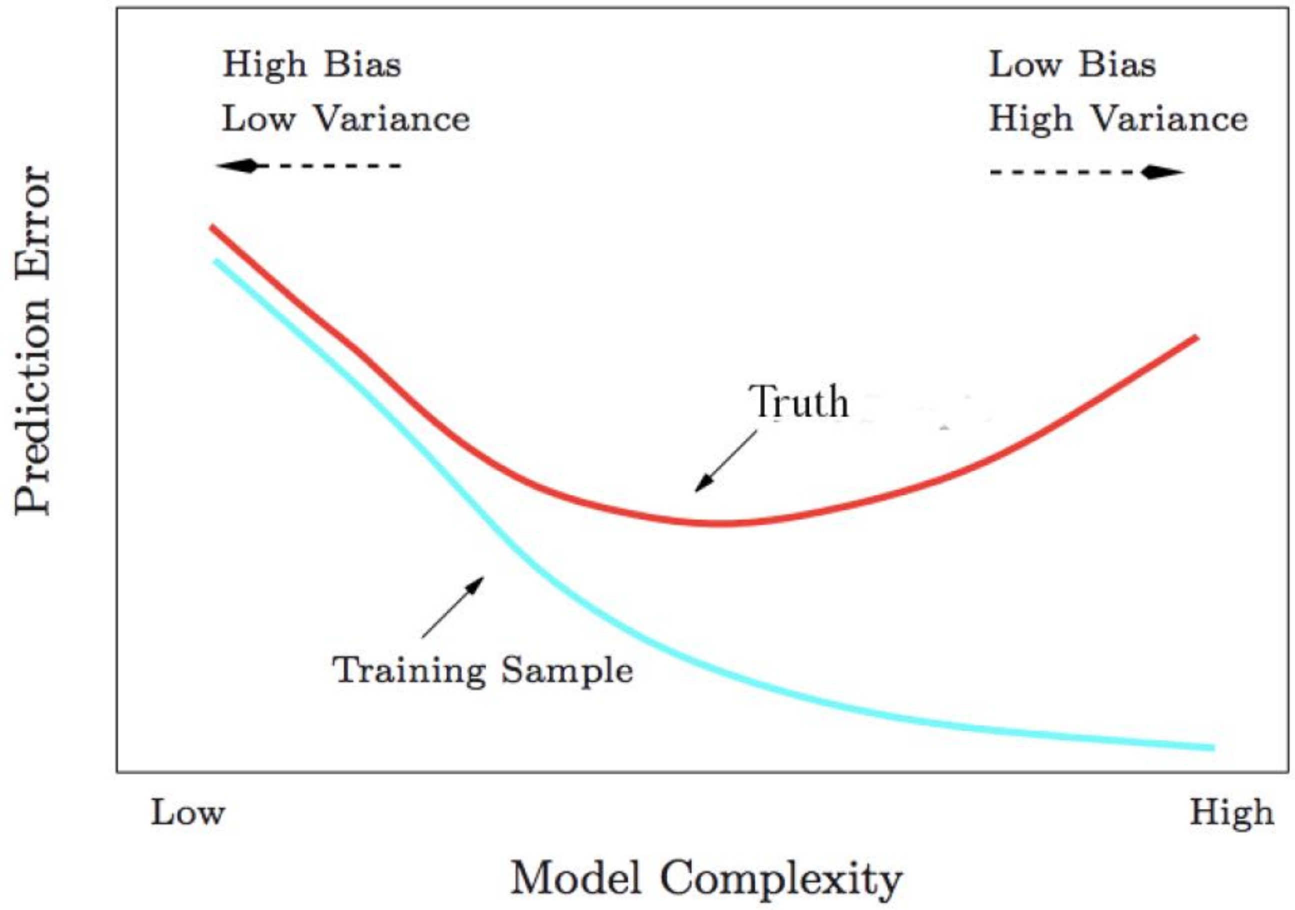


Bias vs Variance Trade-off

Scenario 1



Bias-Variance Tradeoff



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

Classification

Confusion Matrix

- A tool to measure performance for classification

Is accuracy enough?

		Predicted Values	
		Positive(1)	Negative(0)
Actual Values	Positive(1)	True Positive (TP)	False Negative (FN)
	Negative(0)	False Positive (FP)	True Negative (TN)

FIT1043

		Predicted Class		
		Positive	**Negative**	
Actual Class	**Positive**	True Positive (TP)	False Negative (FN) **Type II Error**	**Sensitivity** $\frac{TP}{(TP + FN)}$
	Negative	False Positive (FP) **Type I Error**	True Negative (TN)	**Specificity** $\frac{TN}{(TN + FP)}$
		Precision $\frac{TP}{(TP + FP)}$	**Negative Predictive Value** $\frac{TN}{(TN + FN)}$	**Accuracy** $\frac{TP + TN}{(TP + TN + FP + FN)}$

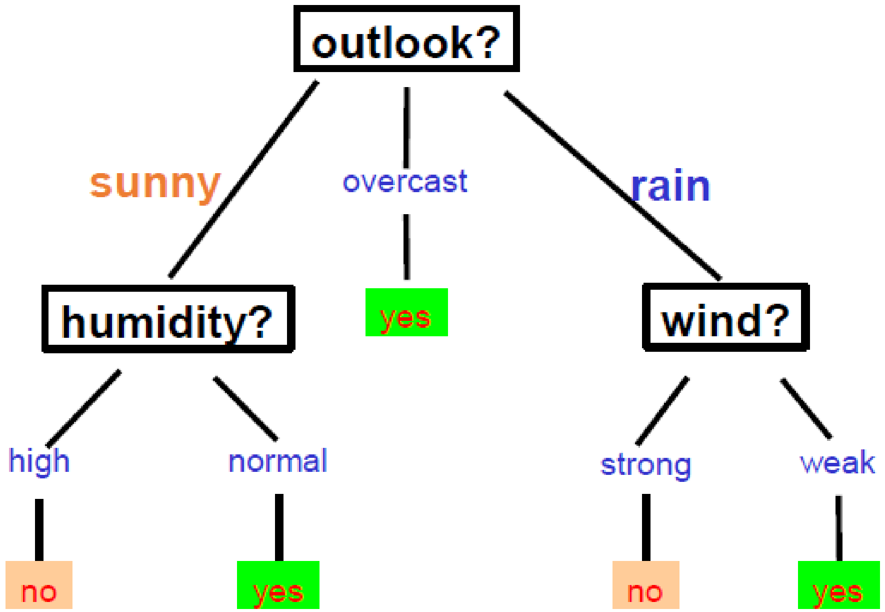
Decision Trees and Regression Trees

What is Decision Trees?

- Predict binary (or categorical)outcomes

What is Regression Trees?

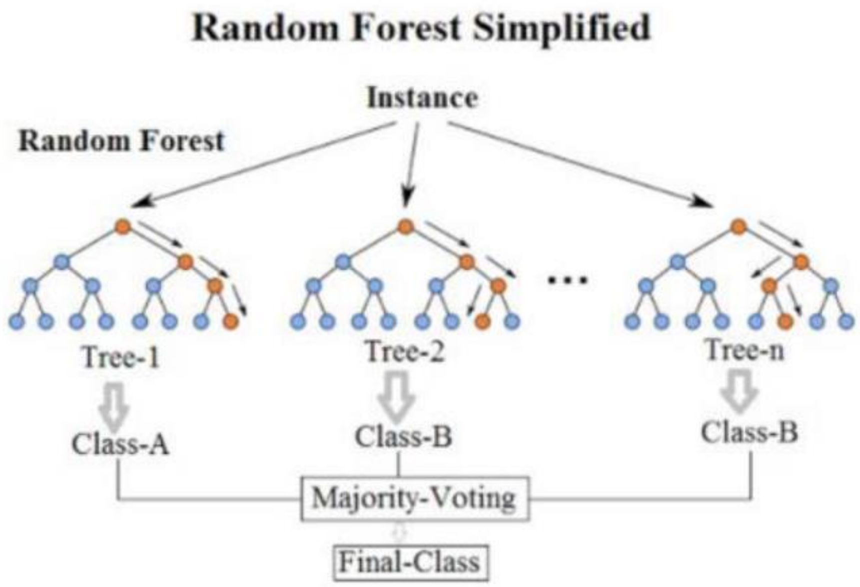
- Predict continuous (i.e. real) values


outlook?
sunny
overcast
rain
humidity?
yes
wind?
high
normal
strong
weak
no
yes
no
yes

- Algorithms for building Decision & Regression trees differ on the criteria (e.g., Entropy) used to:
  - Decide on which feature to split on in each iteration
  - Decide when to stop splitting

What is Random Forest?

- Ensemble learning method that operate by constructing a number of decision trees



What is Clustering?

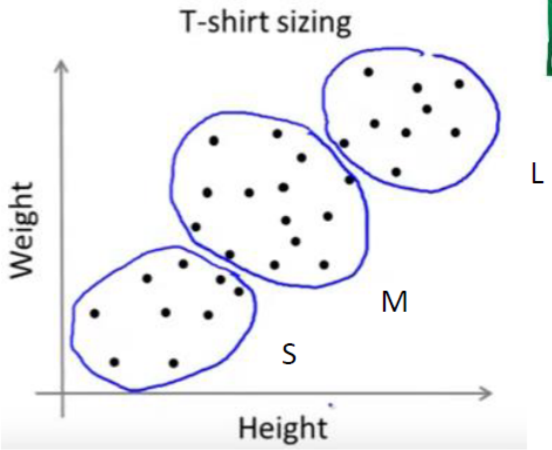
From lecture notes by Andrew Ng

- Grouping a set of data points into different <u>subgroups</u> based on their similarity

called clusters

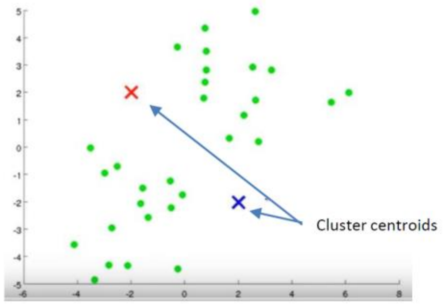
- K-means

- T-shirt manufacturer
- Group into 3 sizes: Small, Medium and Large

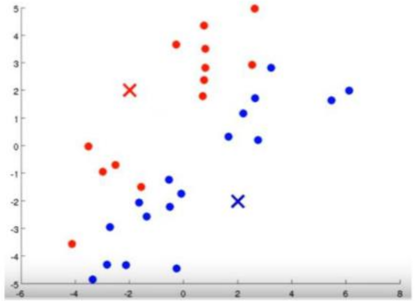


K means

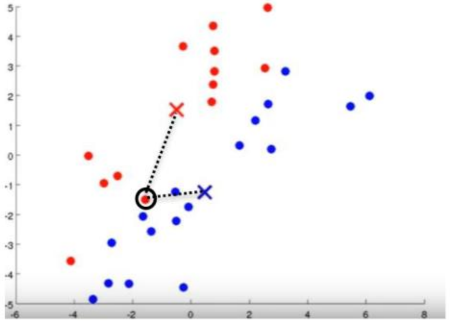
➢ Randomly initialize two points



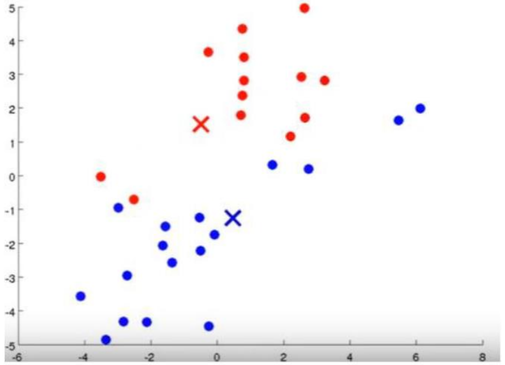
1. Cluster assignment
2. Move centroid



1. Cluster assignment
2. Move centroid

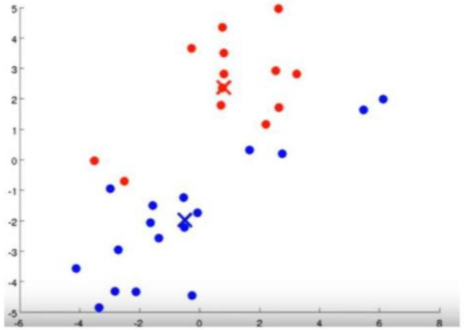


1. Cluster assignment
2. Move centroid

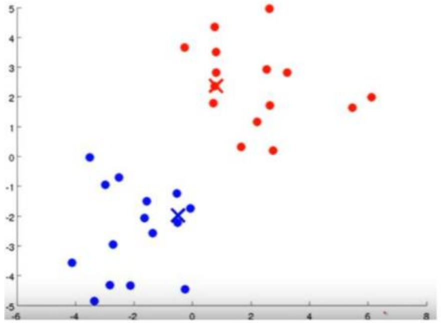


Iterate until there is no changes

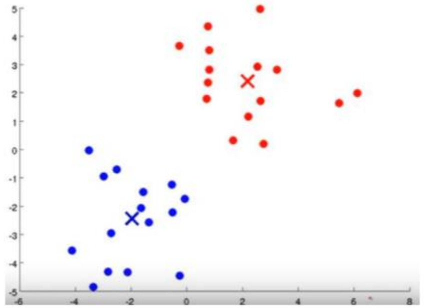
1. Cluster assignment
2. Move centroid



1. Cluster assignment
2. Move centroid



1. Cluster assignment
2. Move centroid



K means is sensitive to initialization !

You have to design the value of K