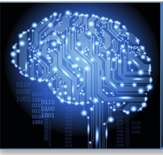


FIT1043 Lecture 5
Introduction to Data Science

Mahsa Salehi

Faculty of Information Technology, Monash University

Semester 2, 2022

Student Feedback Survey

- Hope you enjoyed the unit so far!
- Spend a few mins now to fill in these two anonymous surveys
  - Lecture survey
  - Tutorial survey

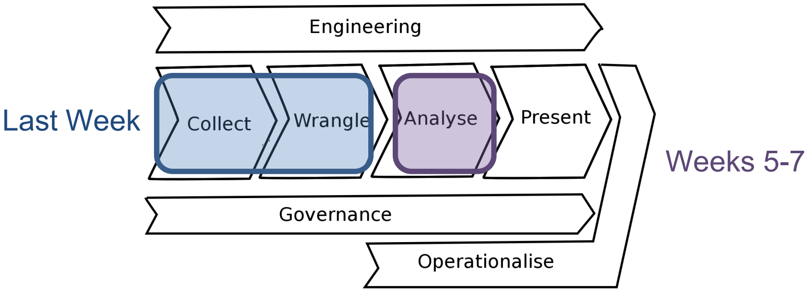
Assignment 1

- Due 29 August 11:55pm
- Dataset: Monthly smartcard replacements dataset in Queensland
- Any questions:
    - Post to forum (Ed discussion)
    - Email: fit1043.clayton-x@monash.edu
    - Email your tutors
    - Attend consultations: click here for the times and locations. We have additional zoom consultations Weeks 4 and 5 for assignment 1.

Unit Schedule

Week	Activities	Assignments
1	Overview of data science	Weekly Lecture/tutorial active participation assessment
2	Introduction to Python for data science	
3	Data visualisation and descriptive statistics	
4	Data sources and data wrangling	
5	Data analysis theory	Assignment 1
6	Regression analysis	
7	Classification and clustering	
8	Introduction to R for data science	
9	Characterising data and "big" data	Assignment 2
10	Big data processing	
11	Issues in data management	
12	Industry guest lecture	Assignment 3

Our Standard Value Chain



Outline

- What is model?
- What are predictive models?
- How to evaluate predictive models?
- Machine learning styles
- What is learning theory
- Linear Regression
- Polynomial regression

Learning Outcomes (Week 5)

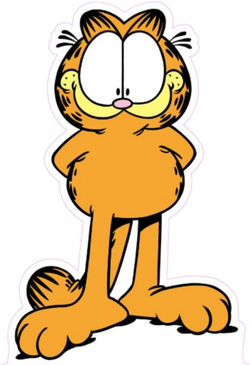
By the end of this week you should be able to:

- Explain what are models and predictive models
- Analyse predictive models in different examples
- Understand how to evaluate predictive models
- Analyse how to estimate linear regression model
- Apply linear regression and polynomial regression on different data sets using Python

What is Model?

What is Model?

Can you draw a CAT..



Better model: closer to reality

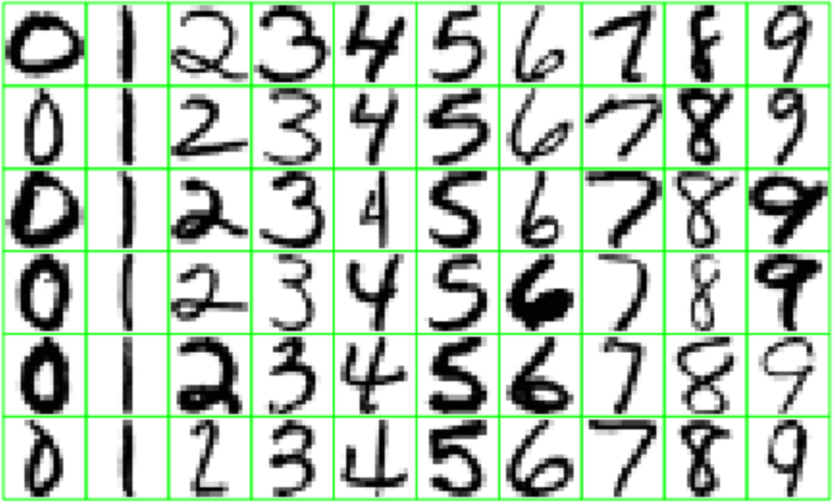
FLUX Question

Do you think you drew a perfect model?

A. Yes
B. No
C. Not sure

What is Model?

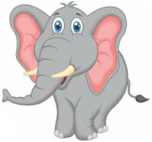
What is Model?



FLUX Question

Which group does this horse belong to?

Group A



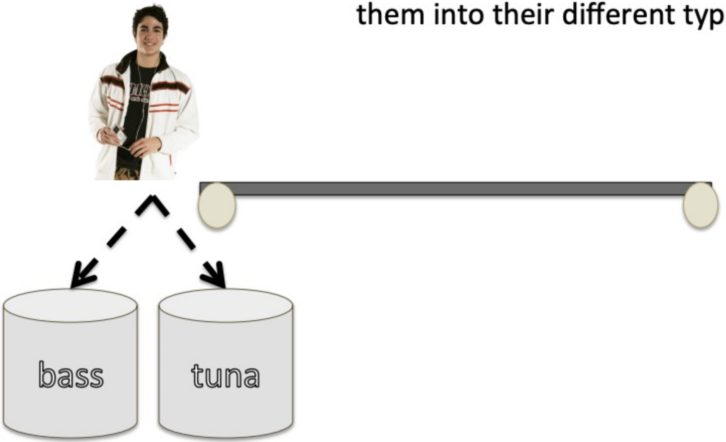
Group B

A brief Introduction to Predictive Models For Data Science

(Example from Duda & Hart, PaCern Classification & Scene Analysis, 1973)

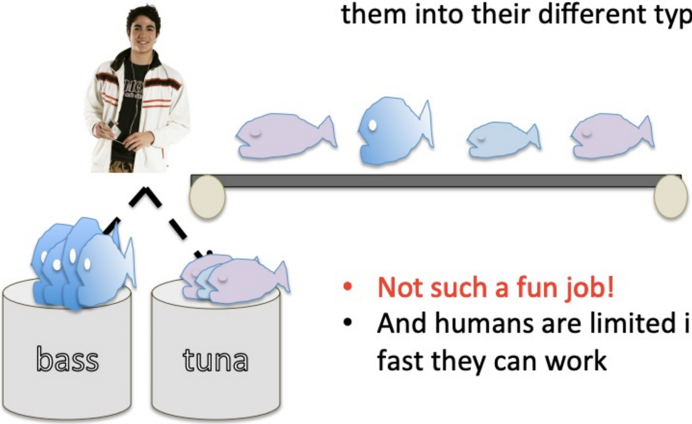
Motivating Example

On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types



Motivating Example

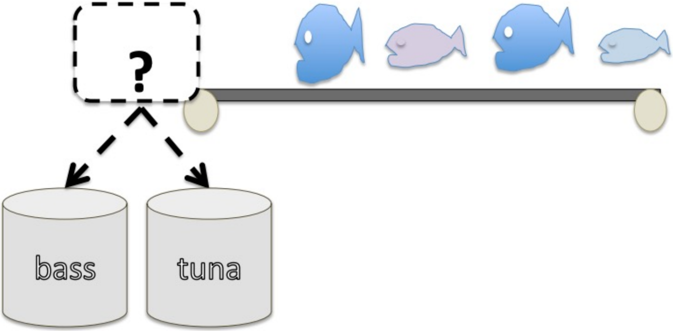
On a fishing boat, a conveyor belt loads fish and a worker separates them into their different types



- Not such a fun job!
- And humans are limited in how fast they can work

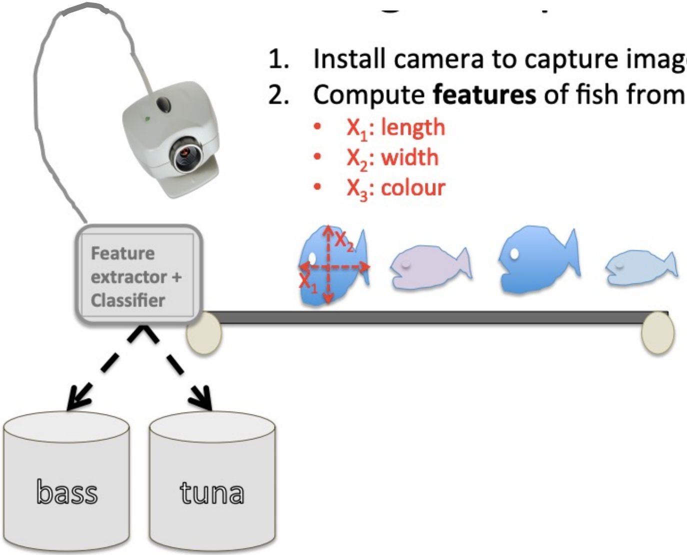
Motivating Example

Question: Can we build a system to do the task automatically?

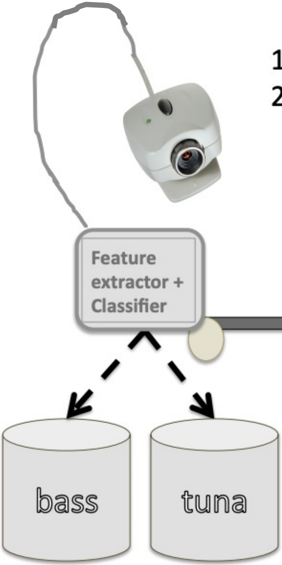


Motivating Example

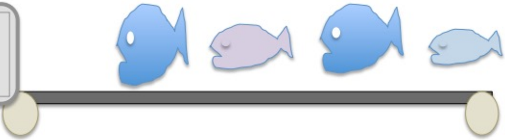
1. Install camera to capture image of fish
2. Compute **features** of fish from image:
   - X_1: length
   - X_2: width
   - X_3: colour



Motivating Example



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

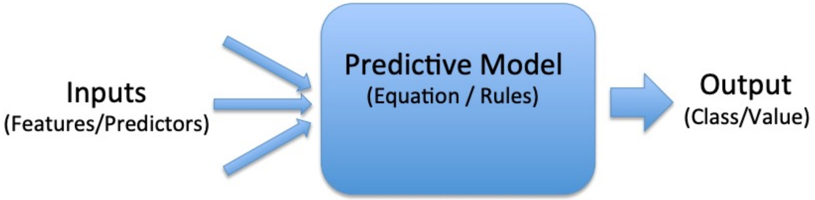


3. Learn rule(s) for classifying fish
   - if $X_1/X_2<1.4$ and X_3=blue => **bass**
   - if $X_1/X_2>2.1$ and X_3=(lightblue,pink) => **tuna**
   - else => **unknown**

Predictive Models

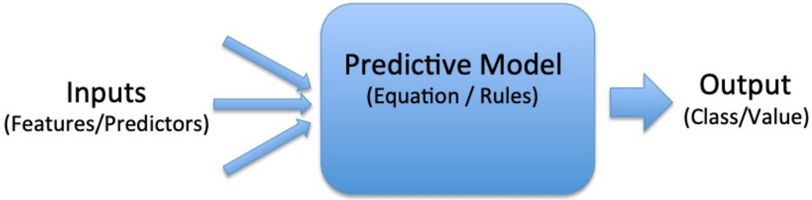
A predictive model is any model that makes a prediction

- 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.



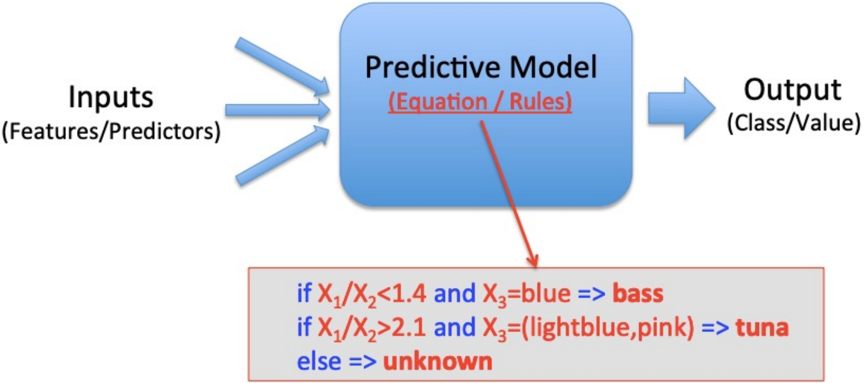
Predictive Models

- 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
- Although there are many other types of models (e.g. ranking, translation, etc.)

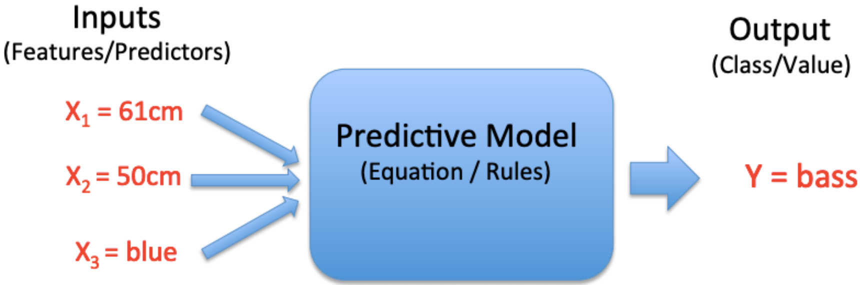


Predictive Models

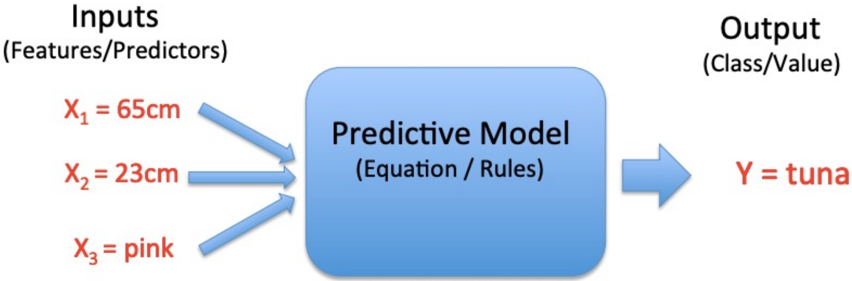
The predictive model uses equations/rules to map the input features to output values



Predictive Models



Predictive Models

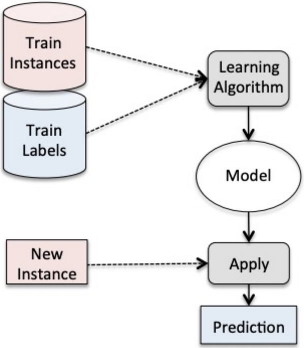


Models are learnt from Examples

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

Training a Model

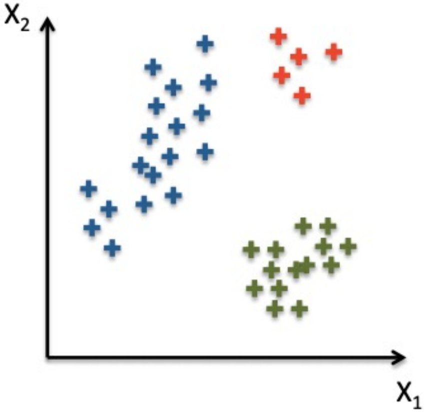
Predictive models are learnt from training data and then applied to make predictions on new instances



How are models learnt?

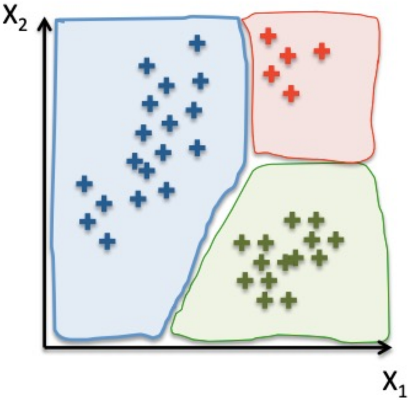
- Each training instance (fish in our case) is just a point in some feature space

- Here the colour denotes the class
  - (blue = bass, green = tuna, red = unknown)



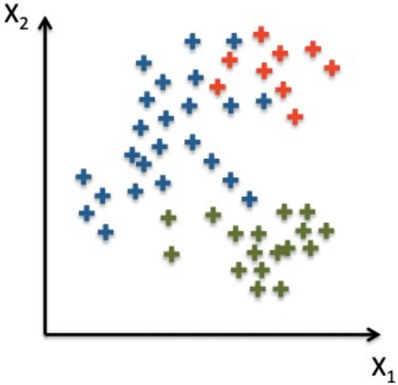
How are models learnt?

- Many (classification) learning algorithms work by dividing the feature space into regions of the same type



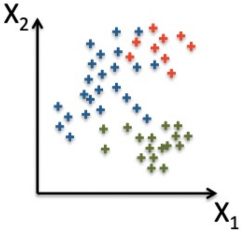
In Practice

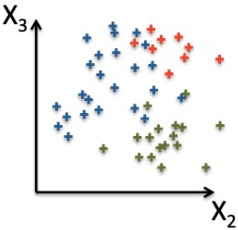
- In practice, the data is usually overlapping
- Making it hard to separate the classes

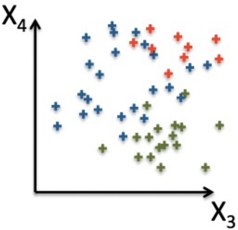


In Practice

- And we have many feature dimensions
- With some features more useful than others

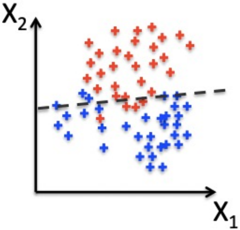






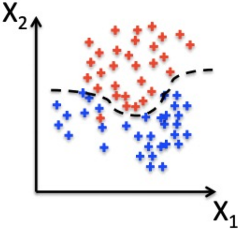
Different Models

- There are many different types of models that we can train to classify objects



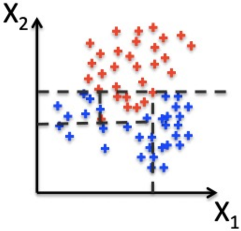
Linear classifiers

e.g. Logistic Regression, Linear SVMs



Non-linear Classifiers

e.g. Neural Nets, SVM with RBF kernel



Decision Tree Learners

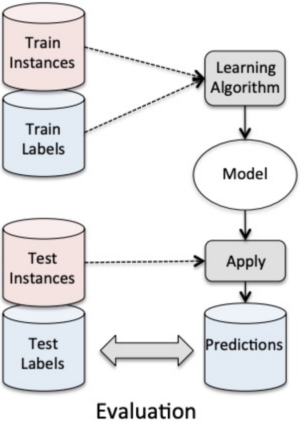
e.g. Random forests

FLUX Question

How can we decide which model is better?

Testing models

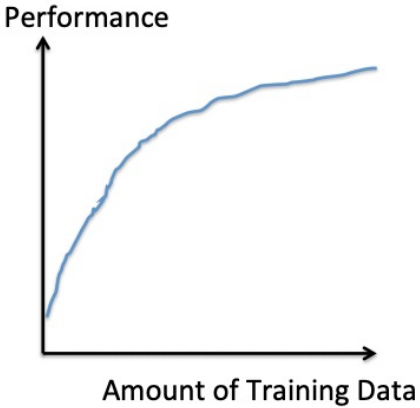
- We evaluate predictive models based on how well they predict the labels for test instances (not used in training)



Performance of predictive models

Generally:

- The more training data the better the test performance
- And (providing there is sufficient training data) the more features the better performance



Introduction to Machine Learning

Introduction to Machine Learning

- **What is Machin Learning?**

  - From Wikipedia: “…is the scientific study of algorithms and statistical models that computer systems use to perform a specific task without using explicit instructions, relying on patterns and inference instead.”

  - From Emerj: “Machine Learning is the science of getting computers to learn and act like humans do, and improve their learning over time in autonomous fashion, by feeding them data and information in the form of observations and real-world interactions.

FLUX Question

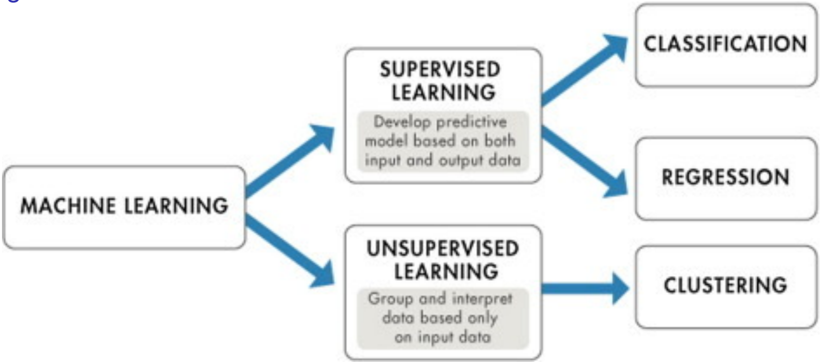
Why is Machine Learning important?

Introduction to Machin Learning

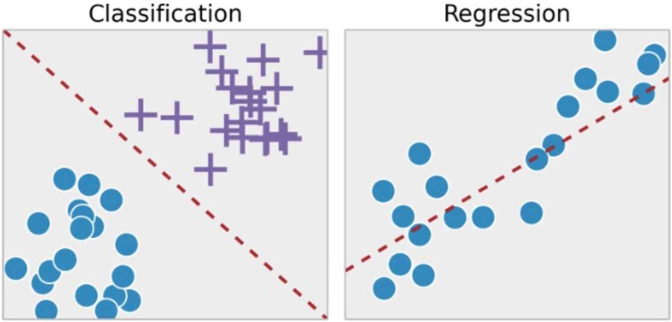
- **How to develop a Machin Learning model?**
  - Choose a measure of success
  - Setting an evaluation protocol
  - Developing a Benchmark Model
  - Developing a Better Model and tunning its Hyperparameters

Learning Styles in Machin Learning Algorithms

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



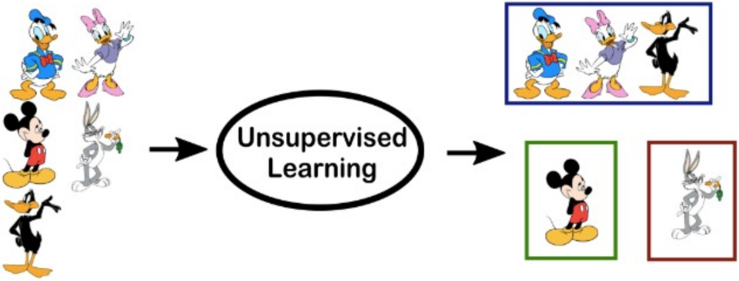
Learning Styles: Supervised Machin Learning (Brownlee, 2019)



Learning Styles: Supervised Machin Learning (Brownlee, 2019)

- All data is labelled and the algorithms learn to predict the output from the input data.
- The goal is to approximate the mapping function so well that when you have new input data (x), you can predict the output variable (Y) for that data.
- **Example Problems**: **Fish example**
    - Classification:
        - The output variable is a category (e.g. “red” or “Blue”)
    - Regression:
        - The output variable is a real value (e.g. “dollars” or “weight”)
- **Example Algorithms:**
    - Linear regression for regression problems.
    - Random forest for classification and regression problems.
    - Support vector machines for classification problems.

Learning Styles: Unsupervised Machine Learning (Brownlee, 2019)



Learning Styles: Unsupervised Machine Learning (Brownlee, 2019)

- All data is unlabelled and the algorithms learn to inherent structure from the input data.
- The goal is to model the underlying structure or distribution in the data in order to learn more about the data.
- **Example Problems**: **face similarity detection**
  - Clustering:
    - Discover the inherent groupings in the data (e.g. grouping customers by purchasing behaviour)
  - Association:
    - Discover rules that describe large portions of your data (e.g. people that buy X also tend to buy Y)
- **Example Algorithms:**
  - k-means for clustering problems.
  - Apriori algorithm for association rule learning problems.

Theory of Data Analysis

Introduction to Learning Theory

What is Learning Theory

From Wikipedia: (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

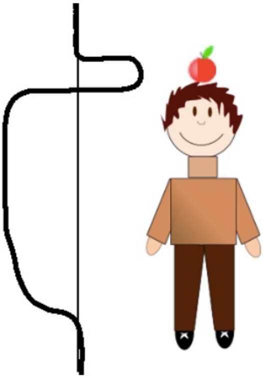
► to evaluate the quality of results derived from learning, we need notions of value

► 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



- 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?

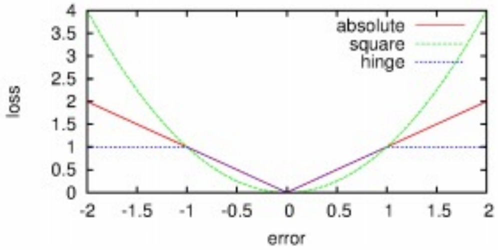
The boy is smiling! its hard to find a cartoon with an apple on a boy's head

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.:

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

Regression

$ $$ $$$ $$$$

?

FLUX Question

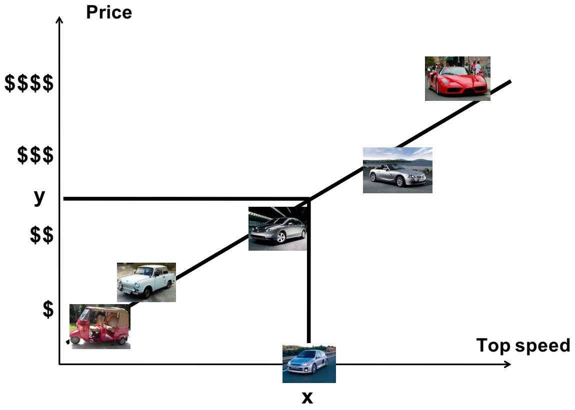
A B C D

$ $$ $$$ $$$$

?

How much is this car worth?

Regression

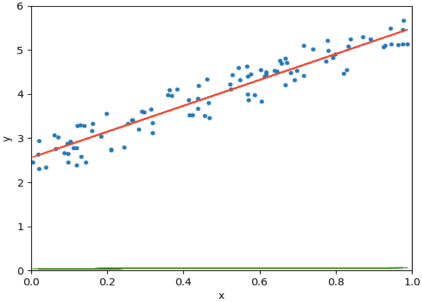


Linear Regression

Regression fits a very simple equation to the data:

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

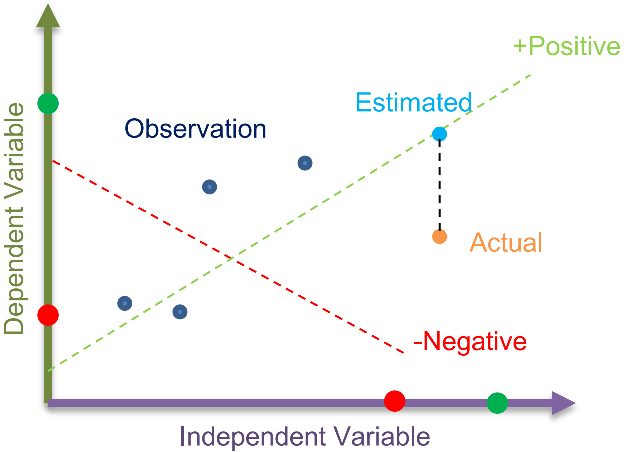
- Data is shown with blue dots, red line is the "linear fitted model"



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

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

How to Calculate Linear Regression



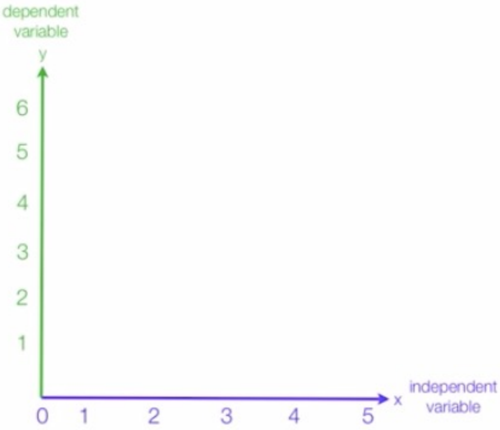
How to Calculate Linear Regression

Example from [this link]

independent variable

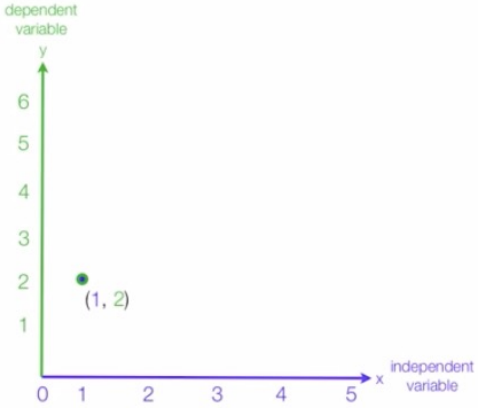
x
1
2
3
4
5

How to Calculate Linear Regression



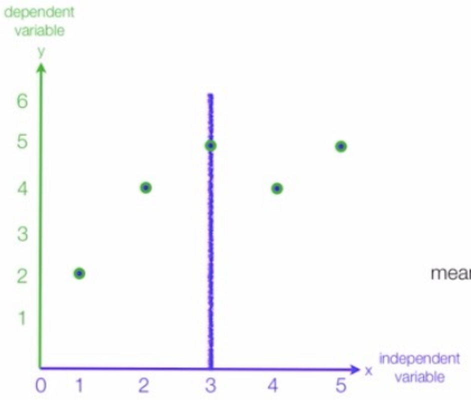
independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

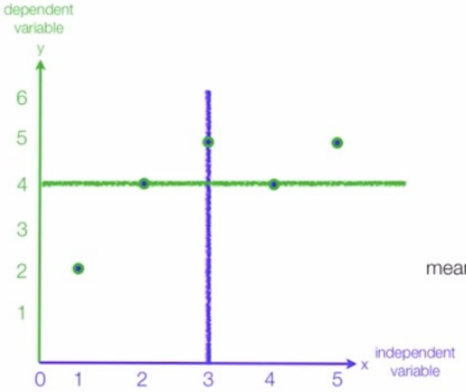
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

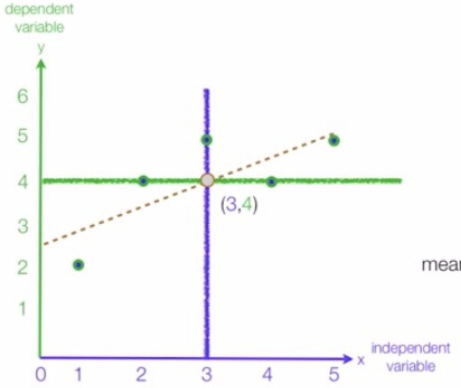
mean 3

How to Calculate Linear Regression



	independent variable	dependent variable
	x	y
	1	2
	2	4
	3	5
	4	4
	5	5
mean	3	4

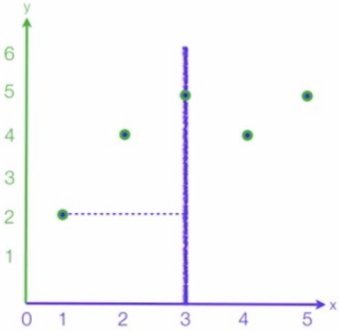
How to Calculate Linear Regression



independent variable	dependent variable
x	y
1	2
2	4
3	5
4	4
5	5

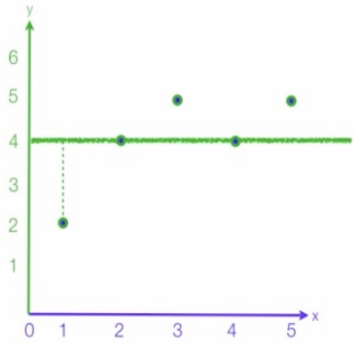
mean

How to Calculate Linear Regression



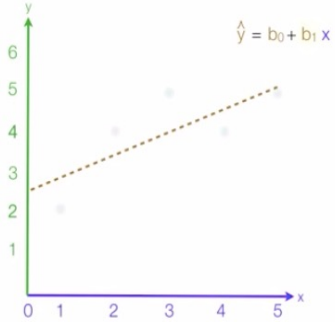
	x	y	$x - \bar{x}$	
	1	2	1 - 3	
	2	4		
	3	5		
	4	4		
	5	5		
mean	3	4		

How to Calculate Linear Regression



	x	y	$x - \bar{x}$	$y - \bar{y}$
	1	2	-2	2 - 4
	2	4	-1	
	3	5	0	
	4	4	1	
	5	5	2	
mean	3	4		

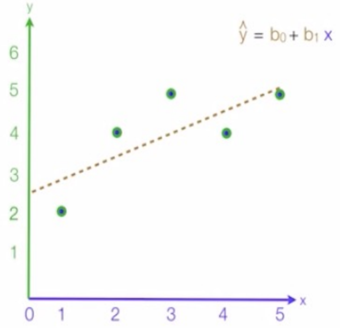
How to Calculate Linear Regression



x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2

mean 3 4

How to Calculate Linear Regression

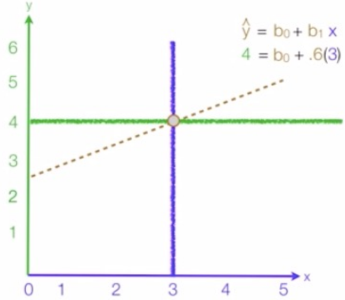


	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
mean	3	4			10	6

sum

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

How to Calculate Linear Regression



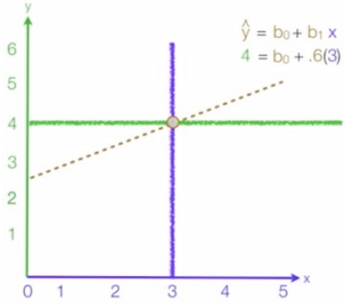
x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
1	2	-2	-2	4	4
2	4	-1	0	1	0
3	5	0	1	0	0
4	4	1	0	1	0
5	5	2	1	4	2
				10	6

mean 3 4

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

$$\begin{aligned} 4 &= b_0 + .6(3) \\ 4 &= b_0 + 1.8 \\ -1.8 & \quad\;\; -1.8 \\ \hline 2.2 &= b_0 \end{aligned}$$

How to Calculate Linear Regression



$b_0 = 2.2$

$b_1 = .6$

$\hat{y} = 2.2 + .6x$

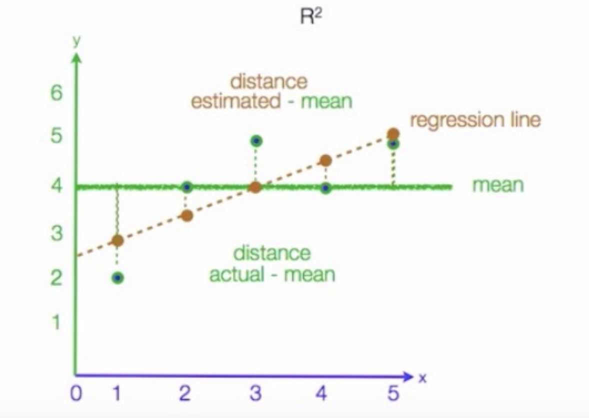
	x	y	$x - \bar{x}$	$y - \bar{y}$	$(x - \bar{x})^2$	$(x - \bar{x})(y - \bar{y})$
	1	2	-2	-2	4	4
	2	4	-1	0	1	0
	3	5	0	1	0	0
	4	4	1	0	1	0
	5	5	2	1	4	2
					10	6
mean	3	4				

$$b_1 = \frac{6}{10} = .6 = \frac{\Sigma(x - \bar{x})(y - \bar{y})}{\Sigma(x - \bar{x})^2}$$

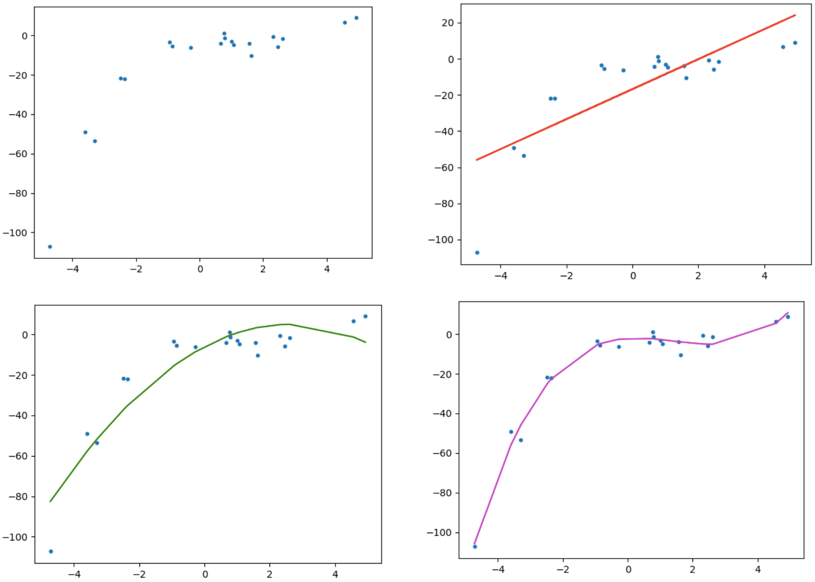
$$4 = b_0 + .6(3)$$

$$\begin{array}{rcl} 4 & = & b_0 + \cancel{1.8} \\ -1.8 & & \cancel{-1.8} \\ \hline 2.2 & = & b_0 \end{array}$$

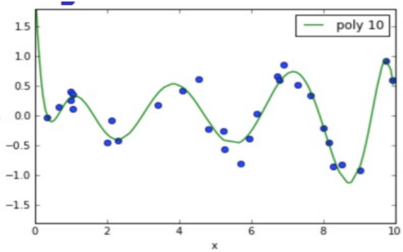
How to Calculate Linear Regression



Polynomial Regression



Polynomial Regression



- Data is shown with blue dots, green line is the "polynomial fitted model"

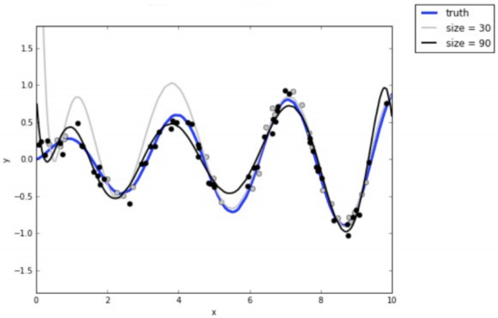
- Polynomial regression uses the same linear regression infrastructure to fit a higher order polynomial.
  In this case we fit a 10-th order polynomial:

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

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

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

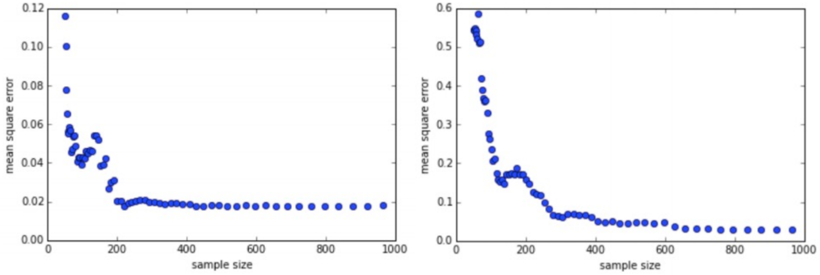
More Data Improves the Fit



- 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

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)

Tutorial/Lab week 5

- Linear regression in Python
  - Using the existing libraries
- Polynomial regression
- Solutions will be released end of the week