Prediction, modelling and classification

UNIVERSITY OF SUSSEX

Dexter R. Shepherd, Paula Seidler, Alejandra Carriero

Recap

We have built circuit python data logging devices

We have gathered datasets

We have looked at how to load in our datasets and visualise them

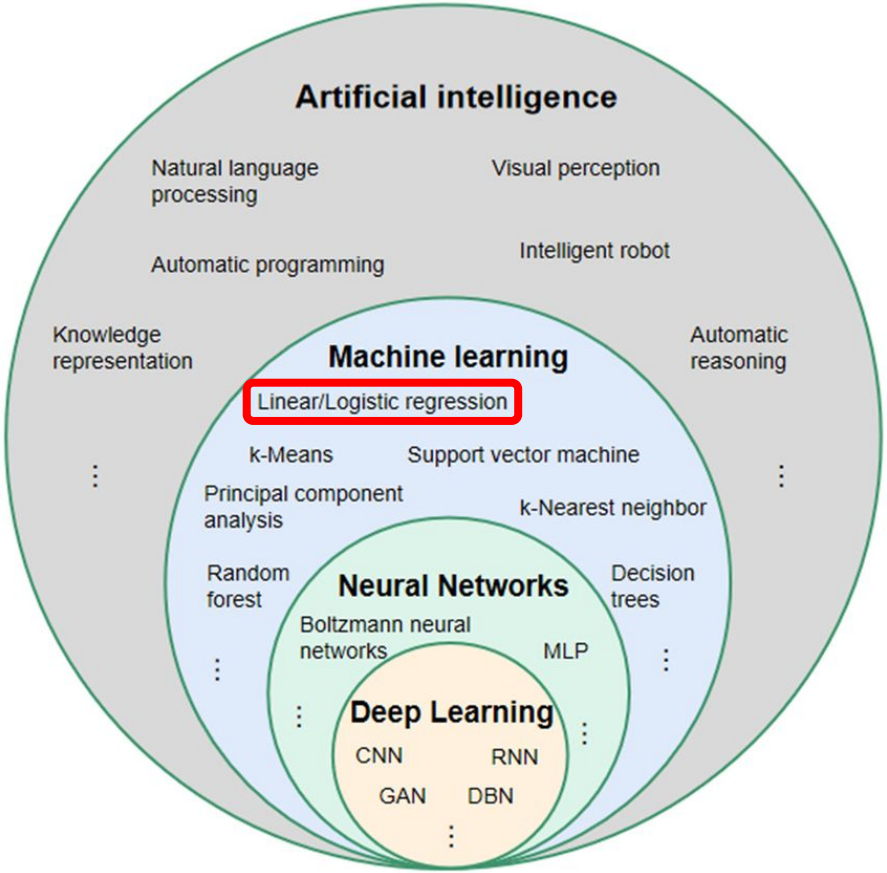
Contents

Introduction

Supervised and unsupervised learning

Data sets and what to look out for

Regression models

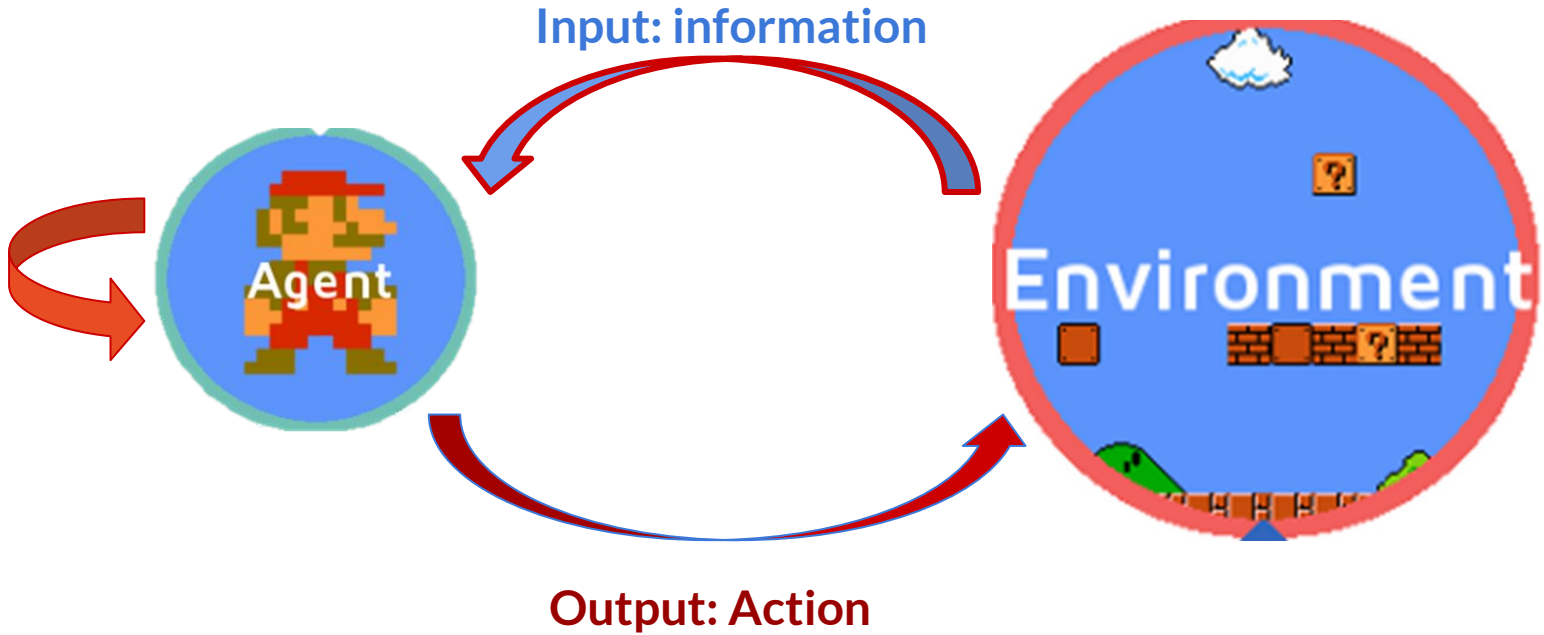

Artificial intelligence
Natural language processing
Visual perception
Automatic programming
Intelligent robot
Knowledge representation
Automatic reasoning
Machine learning
Linear/Logistic regression
k-Means
Support vector machine
Principal component analysis
k-Nearest neighbor
Random forest
Decision trees
Neural Networks
Boltzmann neural networks
MLP
Deep Learning
CNN
RNN
GAN
DBN

Why prediction is important for us

Why prediction is important for us

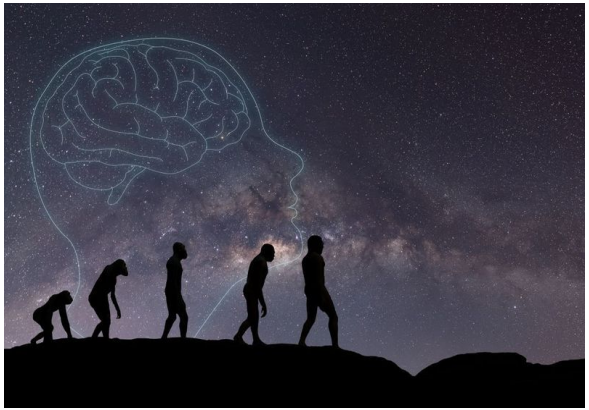
Why prediction is important for us

Predict what is coming next



UNIVERSITY OF SUSSEX

The brain is designed for pattern recognition



Pattern recognition helped our ancestors survive by allowing them to **predict** and respond to environmental changes, such as *weather patterns* or *animal behaviors*:

- Dark clouds + intense wind ⇒ very likely rain (seek repair)
- Tiger growling and assuming an aggressive stance ⇒ likely going to attack (run away)

UNIVERSITY OF SUSSEX

Patterns are how we make sense of the world

Patterns are all around us and we recognise them, whether explicitly or implicitly.

We recognise facial patterns in objects

We recognise patterns in music (even if we cannot really recognise what they are exactly)

Pattern recognition is crucial for prediction

Computers can identify patterns that we cannot identify

Enters the digital age

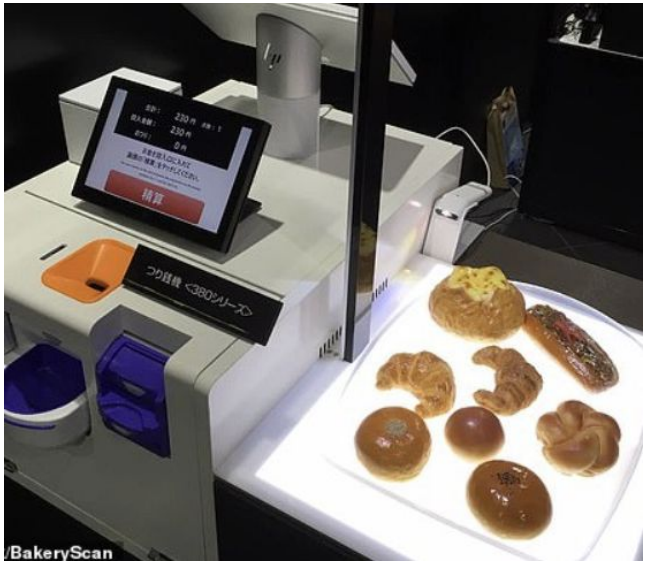
Mail Online

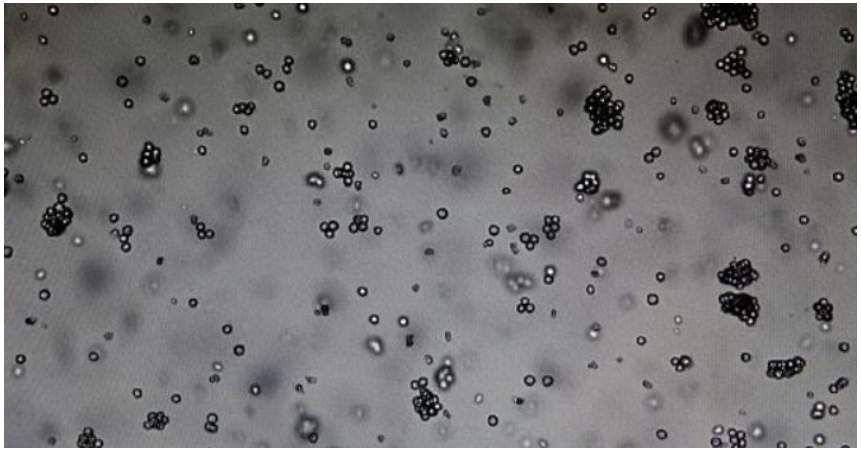
Home | News | Royals | U.S. | Sport | TV | Showbiz | Femail | Health | Science | Money | Trave

Latest Headlines | Blue Origin | SpaceX | NASA | Apple | Google | Twitter | Microsoft

AI designed to distinguish croissants from crullers and other pastries proves capable of identifying cancerous cells on microscope slides with 99% accuracy

- BakeryScan is designed to recognize different pastries to assist at bakeries
- It is equipped with deep learning for object recognition to complete the task
- However, the system was tested in hospitals to see if it can spot cancerous cells
- The system scans a microscope slide, identifies cells and measures the nucleus





The algorithm used to recognise the types of bakery items on a tray was repurposed to recognise types of malignant cancerous cells from the benignant

UNIVERSITY OF SUSSEX

How can the computers identify patterns?

Types of learning

Supervised

You have data that is labelled

Typically used in most ML problems

Requires gathering of large datasets

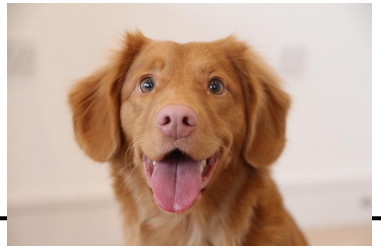
Un supervised

Algorithms learn patterns exclusively from unlabelled

Used for specific algorithms such as clustering, dimensionality reduction, and association

Supervised example

We have our data (known as X)



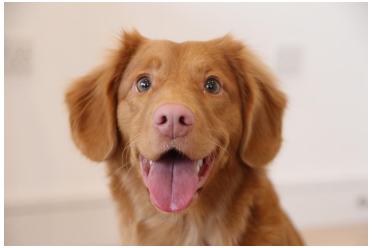
Supervised example

We have our labels (known as y)



y= Class=cat


y= Class=dog



Supervised example

Our model is trained to take in our X an predict y

Predictions are typically referred to as y hat or $\hat{y}$

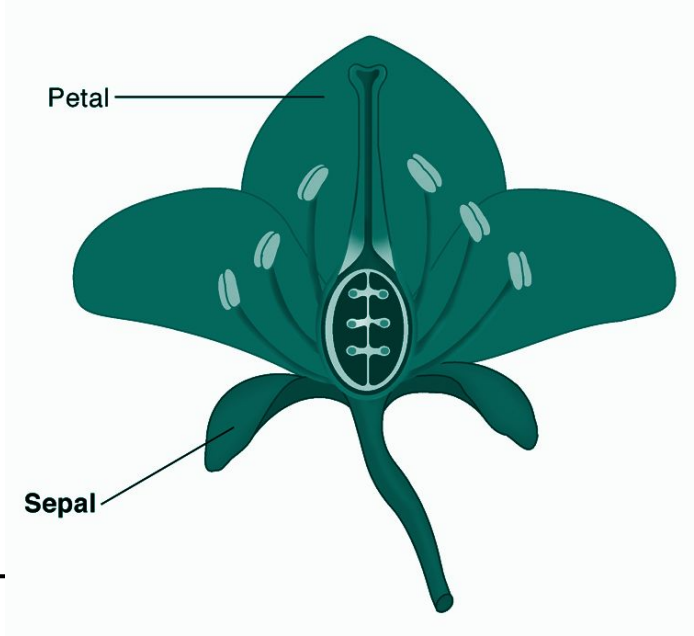
Model

Class: cat
Class: cat
Class: cat

Un-supervised example

We could have data about a flowers, and although not know what species of flower it is, we can classify it into groups which may help us.

Item	Petal length	Sepal width
1	3	2
2	2	5
3	3	5



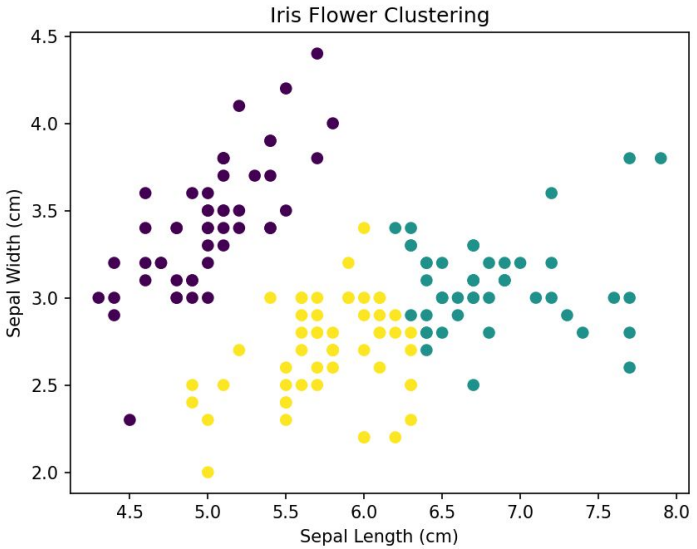
Un-supervised example

Using a clustering method known as k-means we can clearly segment the data into classes

This method did not need labelling

The colour represents the type of flower

Purple Setosa, Yellow Versicolour, and Blue Virginica

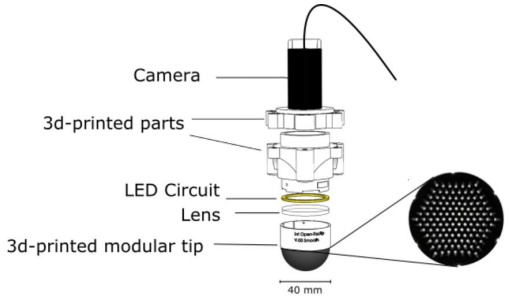


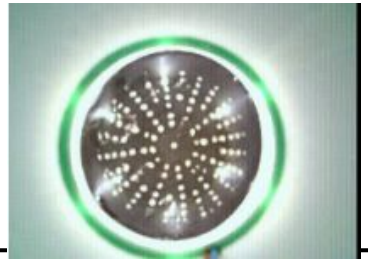
Supervised vs unsupervised

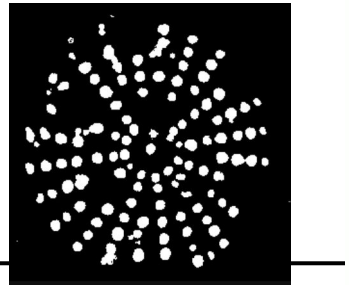
Labelling data costs time and money

Data needs to be high quality and relevant for unsupervised

Unsupervised is helpful when data gathering is challenging eg. labelling something you don't always know. Eg. Netflix working out what other shows to recommend







UNIVERSITY OF SUSSEX

Splitting this data

Data splitting allows us to build robust models that perform well on training data and generalise effectively to new, unseen data.

Training data

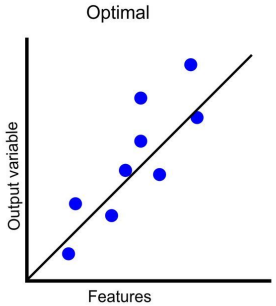
This is the section of the data set we use to train the model, our model might get really good at predicting this, but that does not mean the model works well

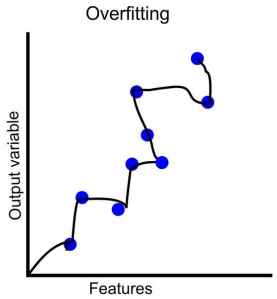
Testing data

The data we use to test how the model performs on unseen data. This gives a good idea of model robustness.

Overfitting

Overfitting occurs when a machine learning model learns the training data too well, capturing noise or random fluctuations in the data rather than the underlying patterns.



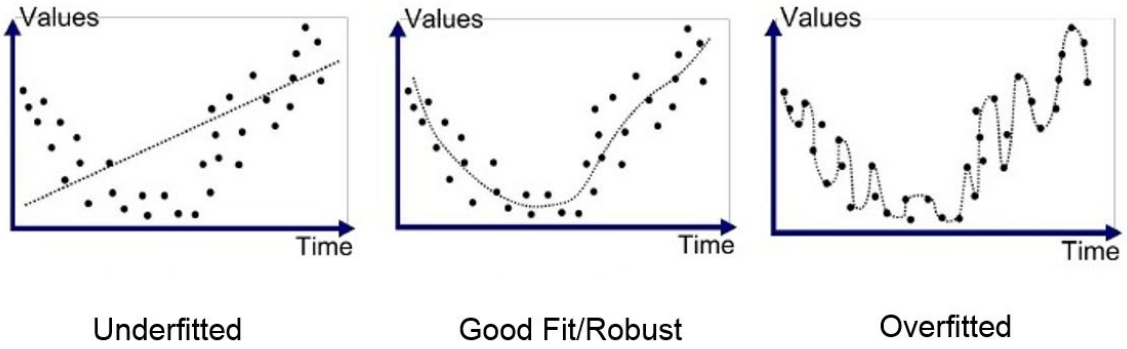


UNIVERSITY OF SUSSEX

Under fitting

Model has not been trained enough

Data is not meaningful enough

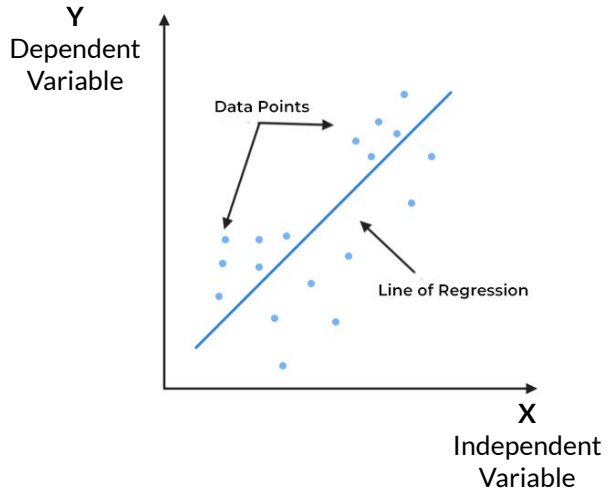


Underfitted | Good Fit/Robust | Overfitted

Regression

Overview of regression analysis

Regression in simple terms is a statistical method used to understand and quantify the relationship between two or more variables. It helps us predict or estimate the value of one variable based on the values of other variables.



UNIVERSITY OF SUSSEX

Key terms in regression

Y

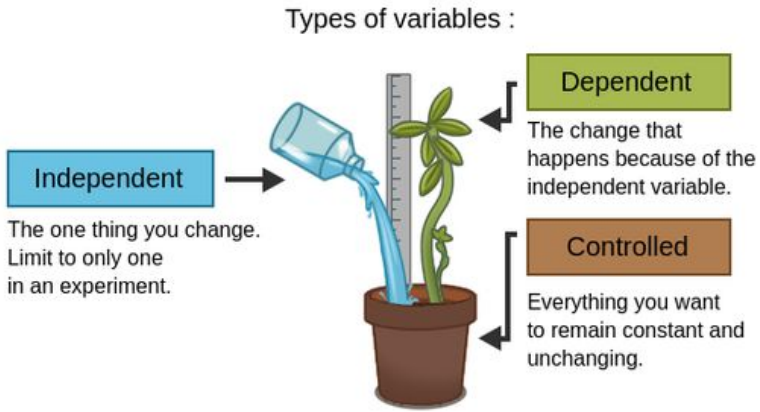
Dependent variable – The variable being predicted

X

Independent variable – The variable used to predict the dependent

Types we will cover:

- Linear regression
- Multiple linear regression



UNIVERSITY OF SUSSEX

But first!

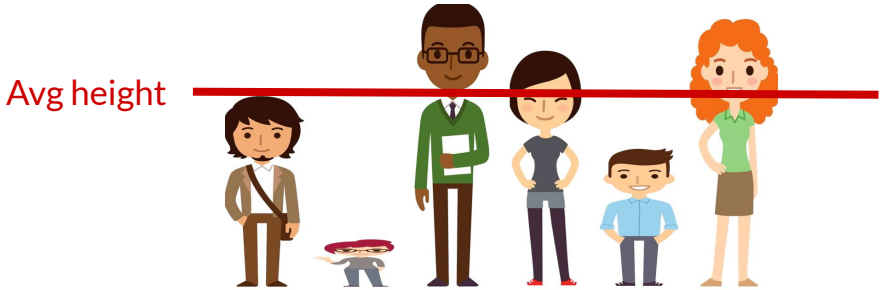
Let's be aware of a few important aspects!

Variance

Variance is a measure of how spread out a set of numbers is.

It tells us how much the numbers in a group differ from the average (mean) of the group.

The variance here is LOW. Their heights don't differ much from the average height



The variance here is HIGH. Their heights differ significantly from the average

Variance

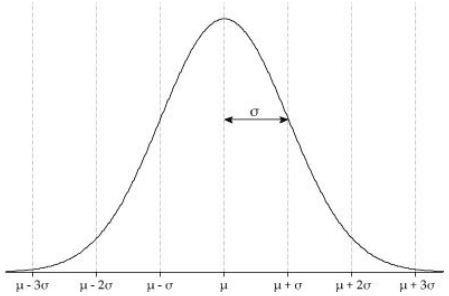
Variance is a measure that indicates the extent of spread of data points from the mean.

$$\sigma^2 = \frac{\Sigma (x - \mu)^2}{N}$$ Population Variance

Sample Variance is an estimate of variance based on a sample.

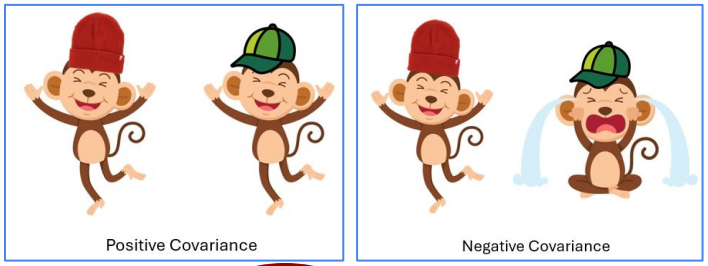
$$s^2 = \frac{\Sigma (x - \bar{x})^2}{n - 1}$$ Sample Variance

They give a scaling factor for the relationship between the independent and dependent variables in regression.



UNIVERSITY OF SUSSEX

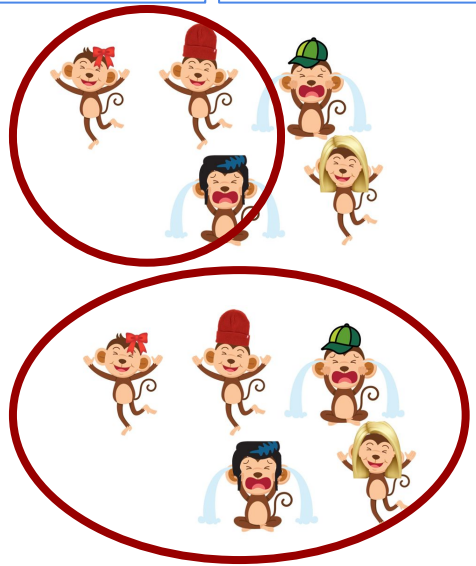
Covariance



Covariance is a measure of how two things change together.

Sample covariance is checking this with a small group, a subset of the population

Population covariance is checking this with everyone



UNIVERSITY OF SUSSEX

Covariance

Covariance is a measure that indicates the extent to which two random variables change together.

Sample covariance is used when we only have access to a sample from the population and want to estimate the population covariance.

Population covariance provides a measure of how two variables change together in the entire population.

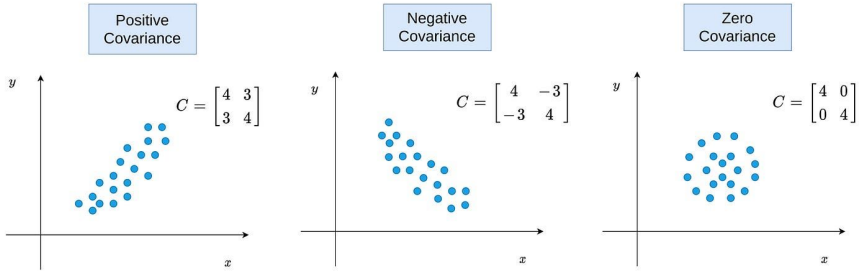
Point Xi | Mean X | Point Yi | Mean Y

For Population

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

For Sample

$$Cov(x,y) = \frac{\sum (x_i - \bar{x}) * (y_i - \bar{y})}{(N-1)}$$



If Cov(X, Y) > 0, the variables move together

If Cov(X, Y) < 0, they move oppositely

If Cov(X, Y) = 0, they are not correlated

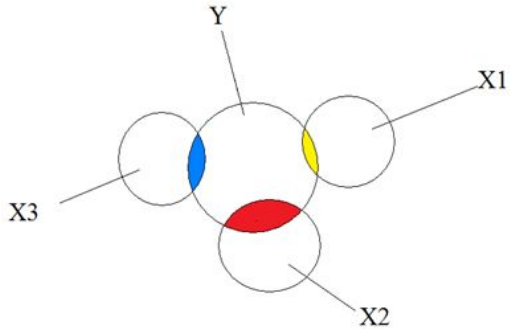
UNIVERSITY OF SUSSEX

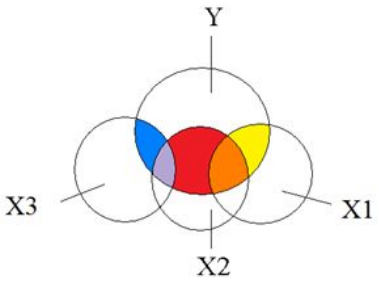
Correlated variables

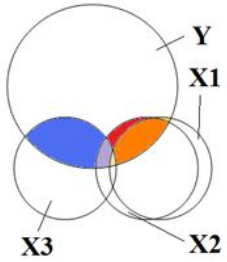
When two or more independent variables are correlated, it can create issues

We call this Multicollinearity

The model is trying to estimate the unique effect of each predictor on Y, but there just isn't enough unique information about X1 and X2 to calculate it.







Example

Using the number of hours spent studying and the number of hours sleep the night before the exam to predict an exam score

If the correlation between the multiple variables is not considered… then there can be multicollinearity problems

Perhaps studying more can lead to a better result

Perhaps the number of hours spent sleeping is also correlated.

Linear regression

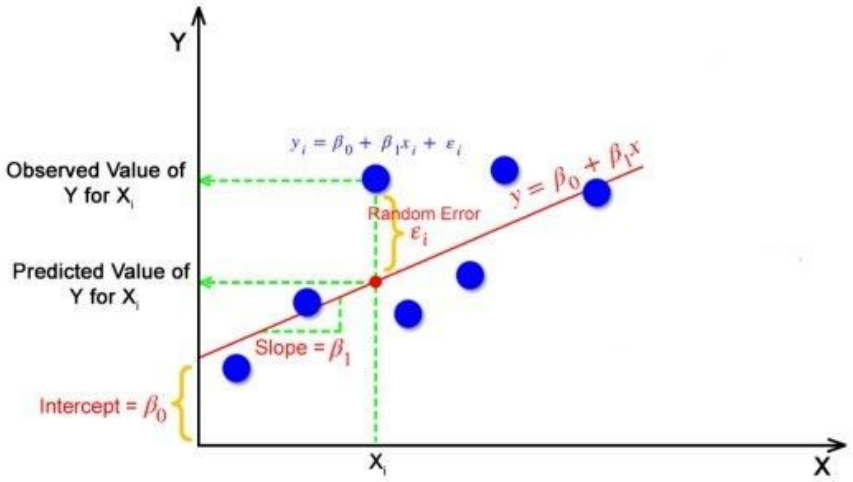
Simplest form of regression

Purpose: Predict and understand the relationship between variables

-> fit red line!

Types of Data: Continuous Data

Examples: Temperature vs. Plant Growth, Rainfall vs. Soil Moisture, …



Y = b + m × X

Fitted Line: Y = β0 + β1 × X + ϵ

Intercept Slope Error

Steps in Linear Regression

Steps:

Why?

1. Calculate Mean, Variance and Covariance

   → Mean → Central tendency - we want the line to pass through the mean
   Covariance → Is there a positive or negative relationship between X and Y?
   Variance → Measures how spread out the values of X are around the mean

2. Calculate Slope(β1) =Cov(X,Y)/Var(X)

   → Slope → How much does Y change per unit increase in X?

3. Calculate Intercept β0=mean(Y) $-\beta 1 \cdot$ mean(X)

   → Intercept → mean(Y)=β0+β1*mean(X); Solving for β0 ensures the line passes through the mean.

Example

Predicting salary based on experience

$$\text{Salary} = \beta_0 + \beta_1 \times \text{Years of Experience} + \epsilon$$

β0 is the intercept (base salary when experience is zero).

β1 is the slope (change in salary for each additional year of experience).

ϵ represents the error term.

Example

Salary (y)	Years experience (X)
10k	3
11k	4
15k	3
20k	5
30k	10
50k	10
60k	12
61k	11
70k	13

Example

Regression models are trained to fit your data set

Calculate the mean of x and y:

y mean = 36.33…

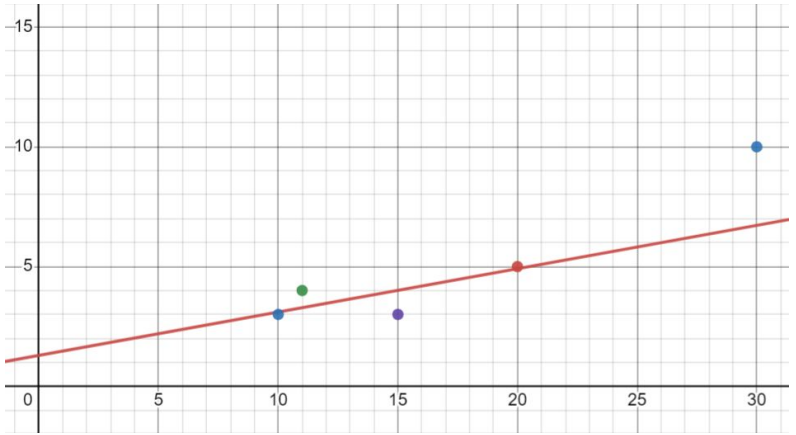
X mean = 7.88…

Then calculate

β1 = covariance(X,Y)/variance(X) = 0.181…

β0 = y mean - β1 * x mean = 1.290…

Y = β1x + β0



Coding this

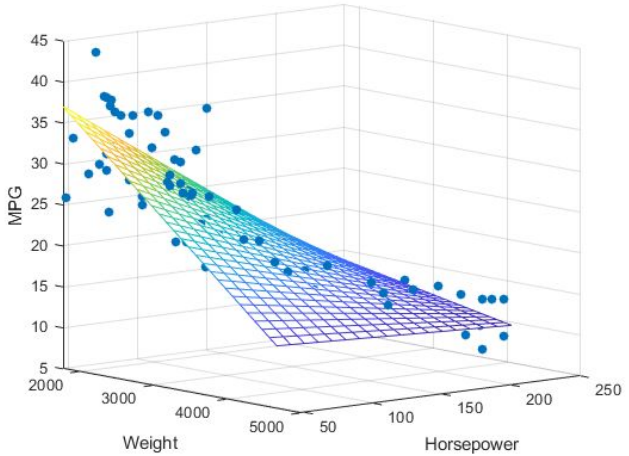
```python
import numpy as np

# Given data
Y = np.array([10, 11, 15, 20, 30, 50, 60, 61, 70])
X = np.array([3, 4, 3, 5, 10, 10, 12, 11, 13])

# Calculate means
mean_X = np.mean(X)
mean_Y = np.mean(Y)

# Calculate covariance and variance
cov_XY = np.cov(X, Y)[0, 1]
var_X = np.var(X)

# Calculate beta_1 (slope)
beta_1 = cov_XY / var_X

# Calculate beta_0 (intercept)
beta_0 = mean_Y - beta_1 * mean_X

print("Beta_0 (Intercept):", beta_0)
print("Beta_1 (Slope):", beta_1)
```

Multi-linear regression

A linear regression model with more than one independent variable

-> **X1, X2, X3,...**

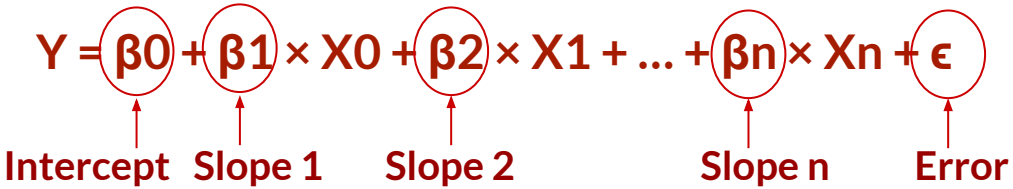
An example of this could be how the moisture of the soil, light intensity and CO_2 levels influence the growth of a plant



UNIVERSITY OF SUSSEX

Adapting the formula

Change in formula



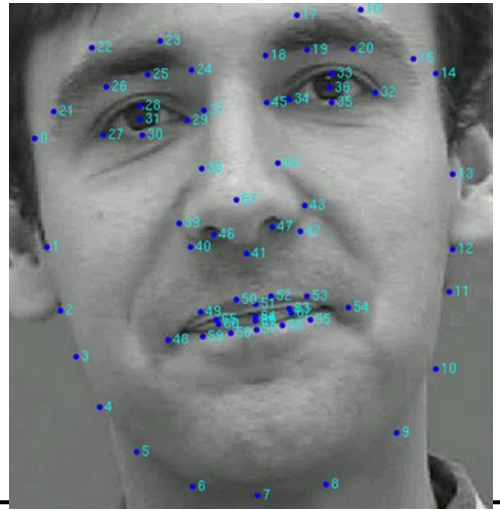
Regression for images

Images are much more complex than the variables we used as examples.

We have a data set of images and label data of coordinates for facial features

Deep learning and NN can be used

Regression is sometimes more efficient



Assumptions

All images are the same size (n,m)

Same number of labels for every image

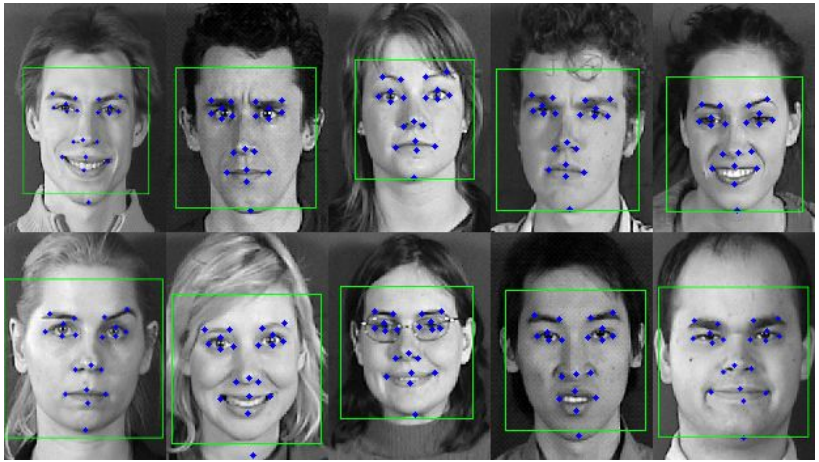
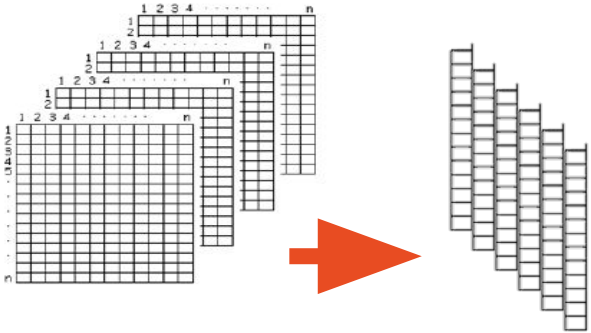


Image to linear data

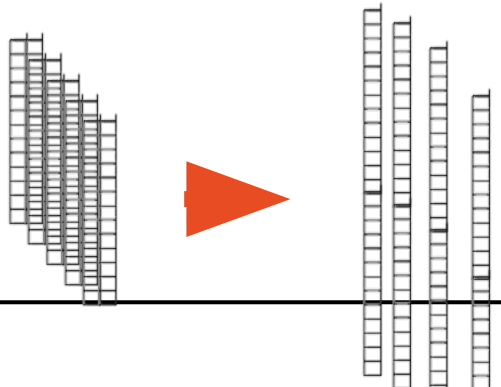
X data:

2D image (n,m) to 1D input layer (n*m)



Y data:

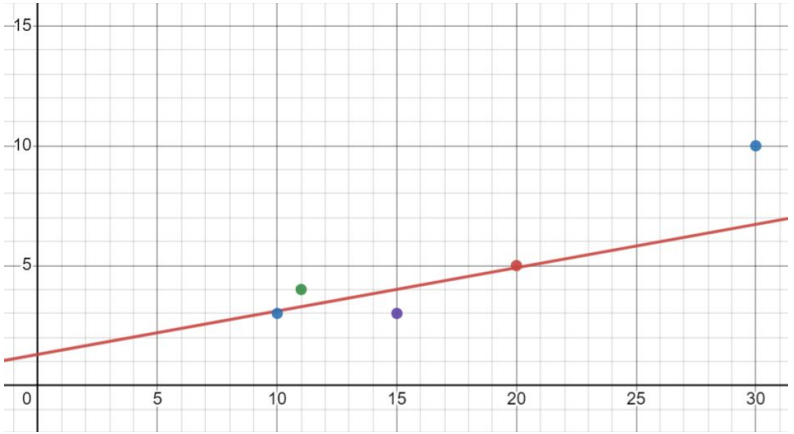
List of coordinates from 2D to 1D



UNIVERSITY OF SUSSEX

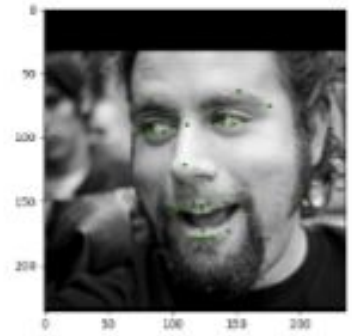
Fit Regression for that linear data

Coordinates
-> flattened

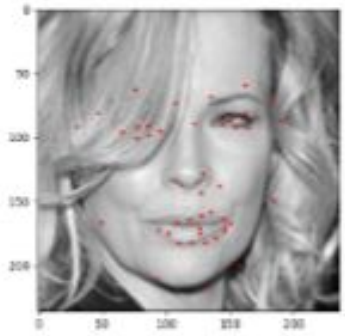


Images -> flattened

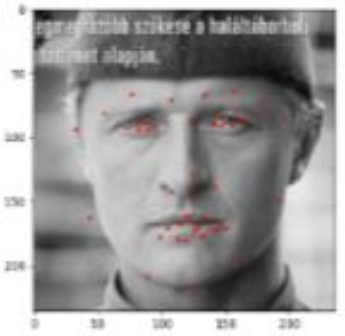
Outcome



(a) Best prediction



(b) Good prediction



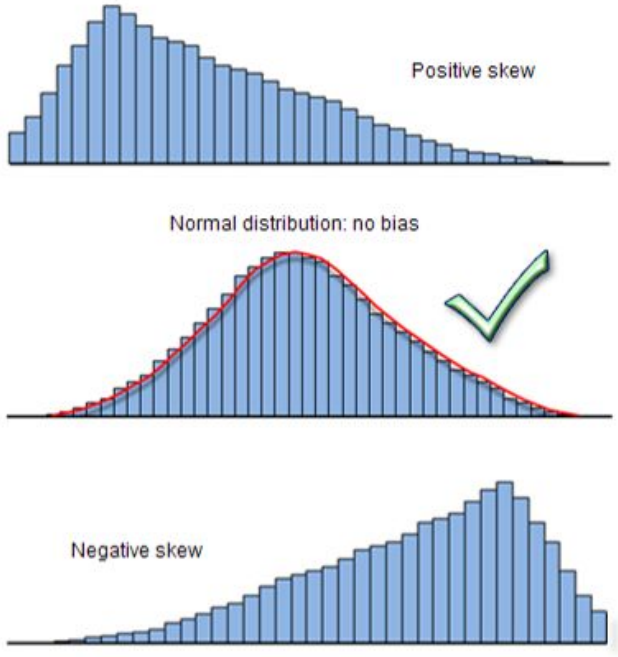
(c) Slightly off

UNIVERSITY OF SUSSEX

Testing validity of models

Using the residuals of the predictions we have the difference between prediction and actual.

In an ideal scenario, residuals should be randomly scattered around zero. If there's a noticeable pattern (e.g., a curve or specific trend) in the residual plot, it indicates that the model might be missing something or that the relationship between variables is not adequately captured.



Conclusions

There are a number of regression models that can be used for simple and complex tasks

When training these models there are a number of factors we must consider to make sure the model works outside the scope of training

We can test the validity of these models by evaluating examples, but also calculating loss

References

“An Introduction to Logistic Regression” - John Whitehead, Department of Economics, East Carolina University

William H. Kruskal and Judith M. Tanur, ed. (1978), "Linear Hypotheses," International Encyclopedia of Statistics. Free Press, v. 1, Evan J. Williams, "I. Regression,“

University Corporation for Atmospheric Research (August 14, 2007). "The WRF Variational Data Assimilation System (WRF-Var)"

David Hochfelder (1998). "Joseph Henry: Inventor of the Telegraph?"

Eckert Jr., John Presper and Mauchly, John W.; Electronic Numerical Integrator and Computer, United States Patent Office, US Patent 3,120,606, filed 1947-06-26, issued 1964-02-04;

Many of the figures are from Towards Data Science: https://towardsdatascience.com/

Labs

First half

Open up the work sheet from the pal python github

https://github.com/SussexPAL/PythonCrashCourse/blob/main/Worksheets/day_4_regression.ipynb

Work through it

We will be using indian agricultural data and regression models

UNIVERSITY OF SUSSEX

Extension

Make a new worksheet that applies regression to your own data.

If your data collection is not a regression task, use someone else's!

You might find that the variables you have gathered have no relationship, this is where robustness comes into play