Types of Machine Learning

Types of Machine Learning

- Supervised Learning
- Unsupervised Learning

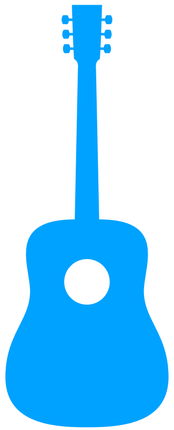
Supervised Learning

- Like learning under a teacher's supervision
- ML algorithm given data along with responses
- Learns from looking at questions and answers

Supervised Learning

- Classification
  - Spam or Ham
  - Mammal or Fish or Reptile or Bird or Amphibian
- Regression (predict numeric value)
  - Income of shopper

Supervised Learning



Cause

Effect

X causes Y

x Variables

- Attributes which an ML algorithm focuses on are called features
- Each data point is a list (or vector) of such features
- Input to an ML algorithm is a feature vector
- Feature vectors are called x variables
- Also called independent variables or predictors

y Variables

- Attributes which an ML algorithm tries to predict are called labels
- Labels can be:
  - Categorical (classification)
  - Continuous (regression)
- Labels are referred to as y variables
- Also called dependent variable

Data Set

Sender	Subject	IP	Body	Result
abc@bank.com	Statement for February 2018	73.45.167.1	Dear Customer, Your account statement…	Ham
familymember@gmail.com	Sunday lunch	103.209.2.92	Are we OK with tapas or should we try…	Ham
xyz@suspicious.net	Click for FREE gift!	204.1.36.82	Valued Customer, You have been selected…	Spam
spammer@donottrustme.com	Miracle cure for baldnesss!!	54.172.109.5	Steve, Donnottrustme has launched…	Spam
kat@ecommercesite.com	Invoice for your purchase of headphones	119.3.5.49	The invoice for the transaction on 24-Feb…	Ham

ATTRIBUTES
FEATURE
TARGET VARIABLE
LABELS

Supervised Learning

$$y = f(x)$$

Most machine learning algorithms aim to "learn" the function f which links the features x to the labels y

Supervised Learning

- Linear Regression
- Logistic Regression

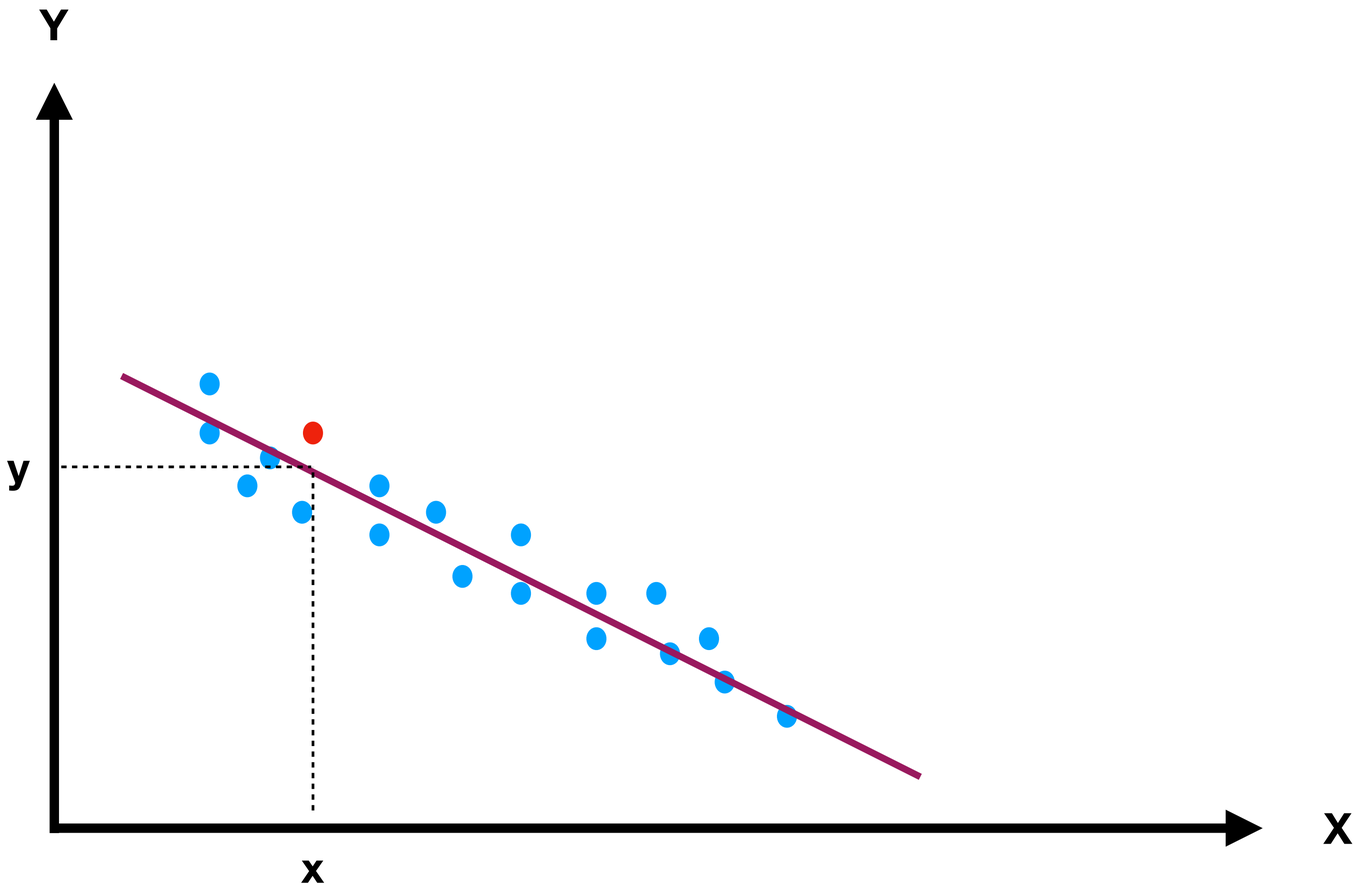
Linear Regression

$$y = Wx + b$$

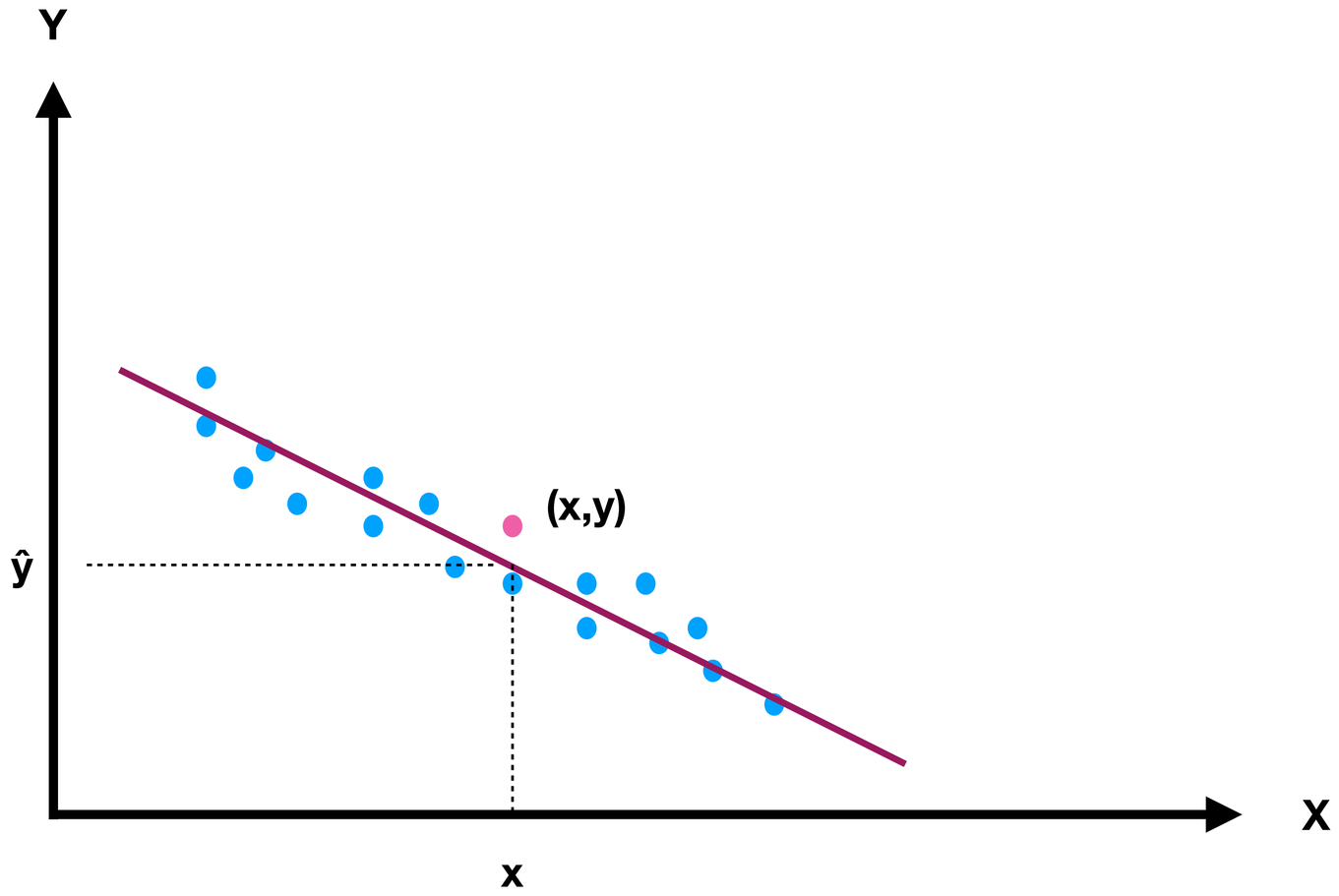
$$f(x) = Wx + b$$

Linear regression specifies, up-front,
that the function f is linear

The "best" regression line



The "best" regression line



The "best" regression line

Residual: $e = y - \hat{y}$

Data: $(x_1, y_1), (x_2, y_2), (x_3, y_3), (x_4, y_4)\ldots$

Residuals: $(y_1-\hat{y}_1), (y_2-\hat{y}_2), (y_3-\hat{y}_3), (y_4-\hat{y}_4)\ldots$

The "best" regression line

Goal: Minimize error for the training data

e.g. Least Square method

Minimize: $(y_1-\hat{y}_1)^2+(y_2-\hat{y}_2)^2+(y_3-\hat{y}_3)^2+(y_4-\hat{y}_4)^2\ldots$

Linear Regression

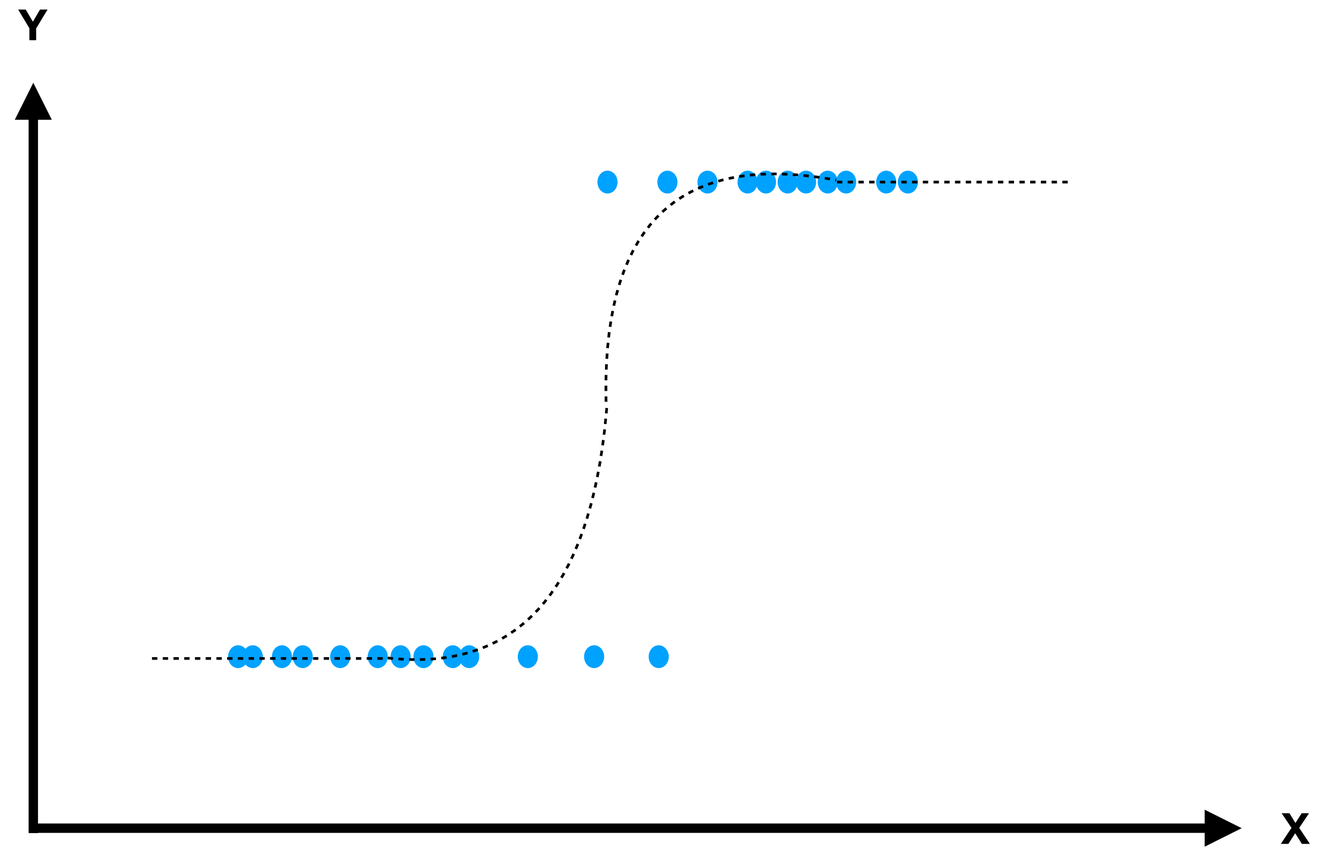
- Use only when errors are normally distributed
- Can also use with multiple independent variables

Logistic Regression

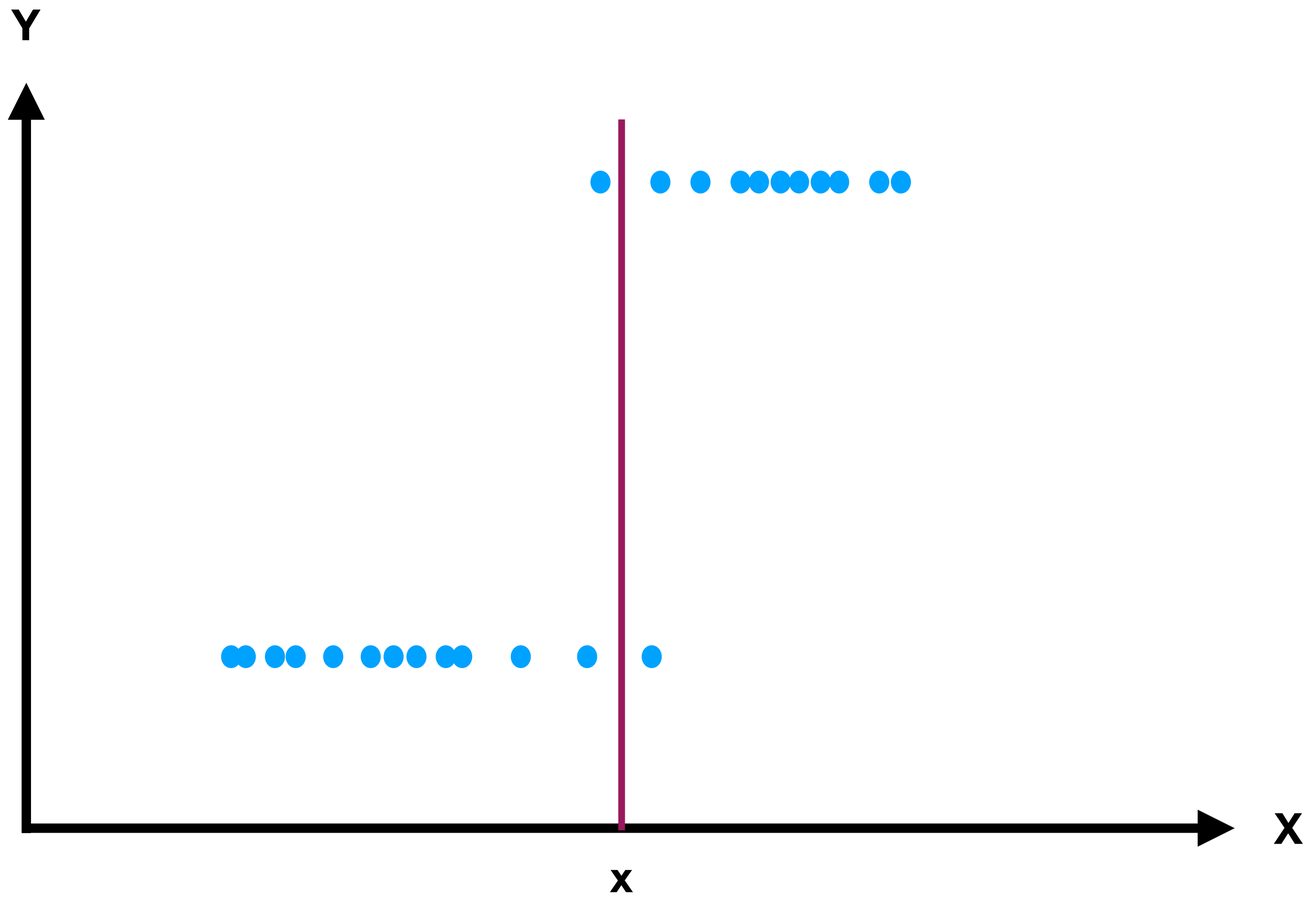
- Used when the dependent variable is categorical
- Independent variables can be continuous or categorical
- Predict probability of each outcome - assign result to category with highest probability

Logistic Regression

Logistic Regression



Logistic Regression



Logistic Regression

Objective function: Minimize cross-entropy

Unsupervised Learning

- There is only input data x - not output data
- Model the underlying structure to learn more about the data
- Algorithms self discover patterns and structure in the data

Unsupervised Learning - algorithms

- Autoencoding - Identify latent factors that drive data (e.g. PCA)
- Clustering - Identify patterns in data items

Looking Within

- Be emotionally self-sufficient
- Learn what matters (to you)
- Identify others who share them…
- …and those who don’t
- Eliminate what does not matter
- Train yourself to navigate the outside world

Why Look Within?

In life

Be emotionally self-sufficient

Learn what matters (to you)

Identify others who share them...

...and those who don't

Eliminate what does not matter

In Machine Learning

Make unlabelled data self-sufficient

Latent factor analysis

Clustering

Anomoly detection

Quantisation

Why Look Within?

ML technique	Use Case
Make unlabelled data self-sufficient	Identify photos of specific individual
Latent factor analysis	Find common drivers of 100 stocks
Clustering	Find relevant document in a corpus
Anomoly detection	Flag fraudulent credit card transactions
Quantisation	Compress 24-bit true colour to 8 bit

Why Look Within?

What	How
Make unlabelled data self-sufficient	Autoencoding
Latent factor analysis	Autoencoding
Clustering	Clustering
Anomoly detection	Autoencoding
Quantisation	Clustering