Details of ML part 1

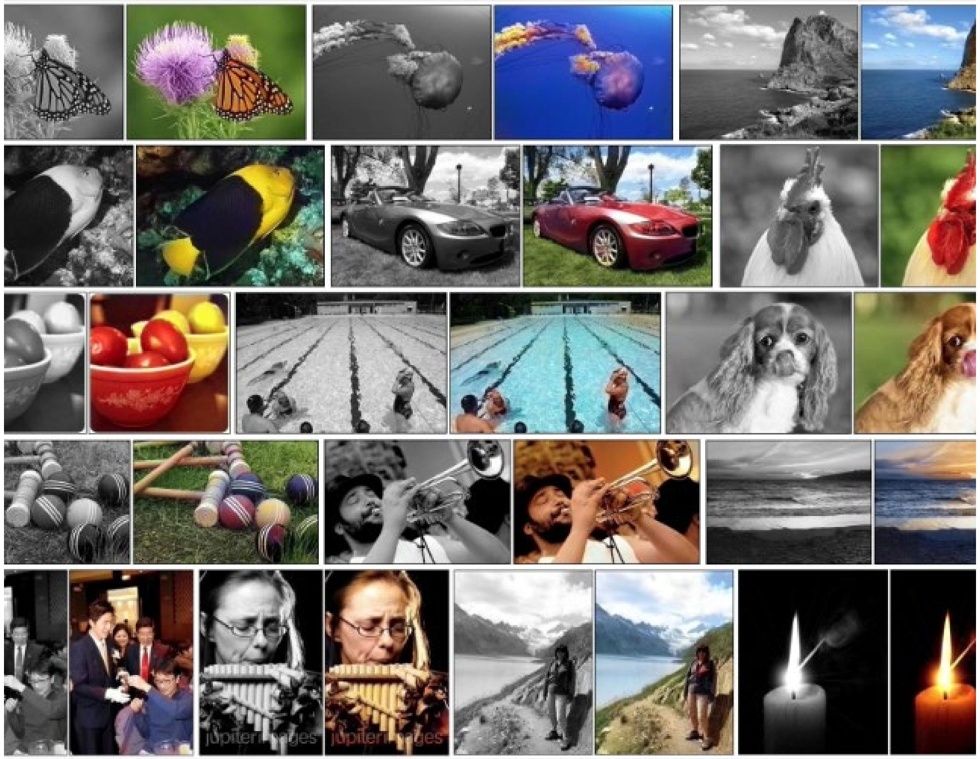
Machine learning tasks

- Supervised learning
  - regression: predict numerical values
  - classification: predict categorical values, i.e., labels
- Unsupervised learning
  - clustering: group data according to "distance"
  - association: find frequent co-occurrences
  - link prediction: discover relationships in data
  - data reduction: project features to fewer features
- Reinforcement learning

Regression

Colorize B&W images automatically

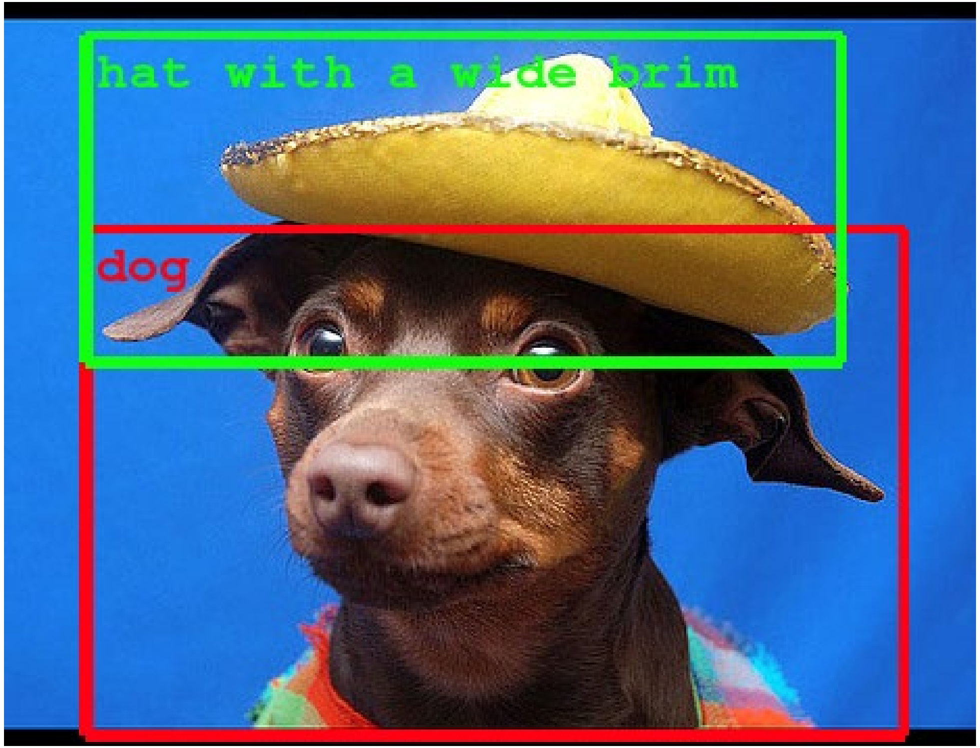
https://tinyclouds.org/colorize/



Classification

Object recognition

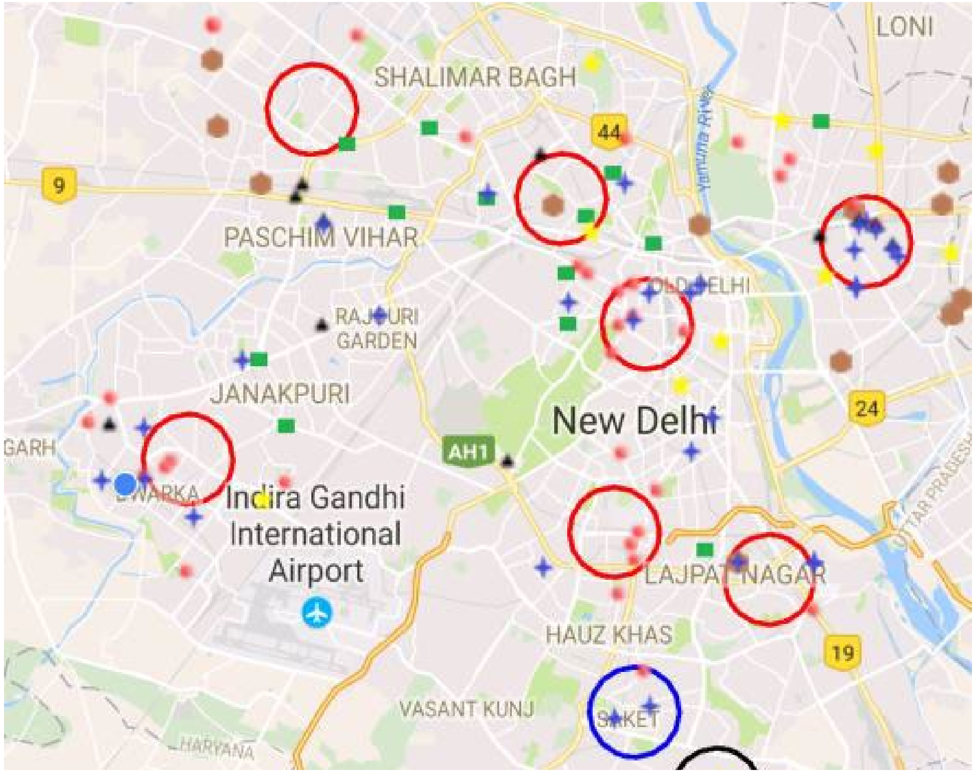
https://ai.googleblog.com/2014/09/building-deeper-understanding-of-images.html



Clustering

Crime prediction using k-means clustering

http://www.grdjournals.com/uploads/article/GRDJE/V02/I05/0176/GRDJEV02I050176.pdf



Reinforcement learning

Learning to play Break Out

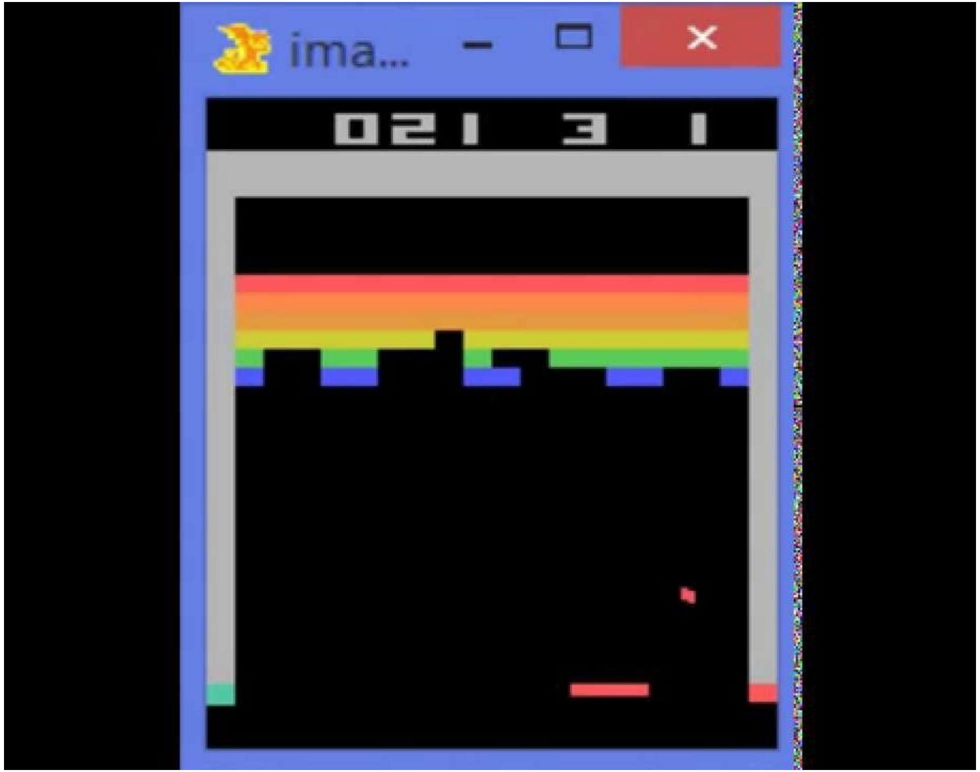
https://www.youtube.com/watch?v=V1eYniJ0Rnk

Mario AI

https://www.youtube.com/watch?v=5GMDbStRgoc

https://www.youtube.com/watch?v=qv6UVOQ0F44

http://www.youtube.com/watch?v=Xj7-QA-aCus



Reasons for failure

- Asking the wrong question
- Trying to solve the wrong problem
- Not having enough data
- Not having the right data
- Having too much data
- Hiring the wrong people
- Using the wrong tools
- Not having the right model
- Not having the right yardstick

Frameworks

- Programming languages
  - Python
  - R
  - C++
  - …
- Many libraries
  - scikit-learn
  - PyTorch
  - TensorFlow
  - Keras
  - …

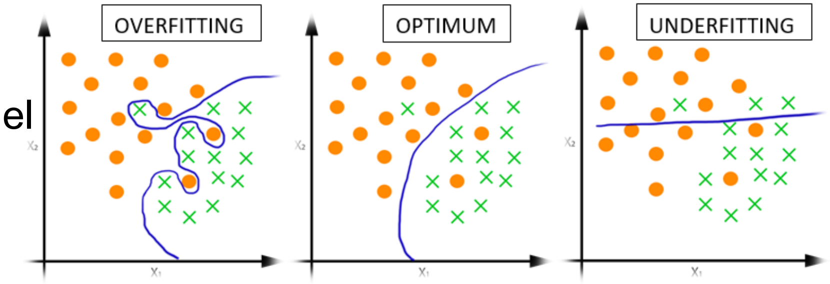
Fast-evolving ecosystem!

classic machine learning

deep learning frameworks

Supervised learning: methodology

- Select model, e.g., decision tree, random forest, support vector machine, …
- Train model, i.e., determine parameters
  - Data: input + output
    - training data $\rightarrow$ determine model parameters
    - testing data $\rightarrow$ yardstick to avoid overfitting
- Prediction model
  - Data: input + output
    - validation data $\rightarrow$ final scoring of the model
- Production
  - Data: input $\rightarrow$ predict output



REGRESSION

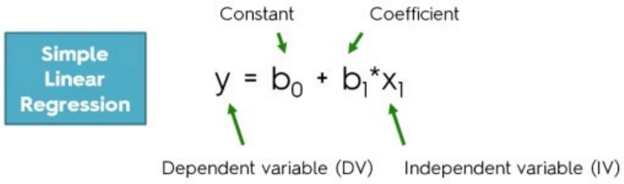
Regression is a statistical measurement used in finance, investing, and other disciplines that attempts to determine the strength of the relationship between one dependent variable and a series of other changing variables or independent variable

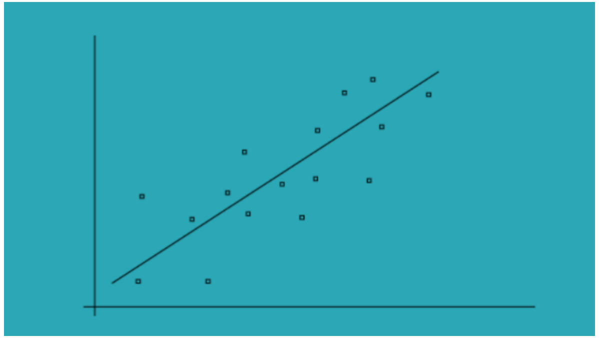
Types of regression

- Linear regression
    - Simple linear regression
    - Multiple linear regression
- Polynomial regression
- Decision tree regression
- Random forest regression

Simple Linear regression

- The simple linear regression models are used to show or predict the relationship between the two variables or factors
- The factor that being predicted is called dependent variable and the factors that is are used to predict the dependent variable are called independent variables

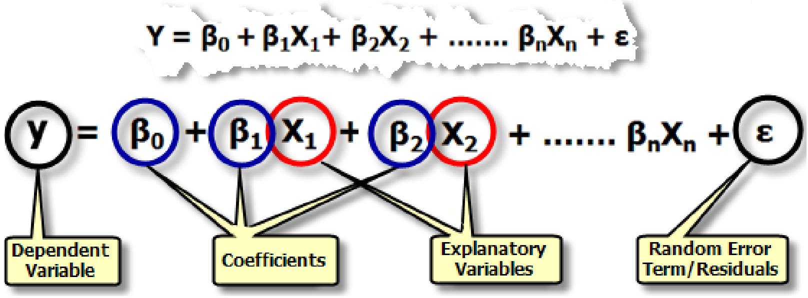




Simple Linear regression

Multiple linear regression

- Multiple regression is an extension of simple linear regression. It is used when we want to predict the value of a variable based on the value of two or more other variables. The variable we want to predict is called the dependent variable (or sometimes, the outcome, target or criterion variable).



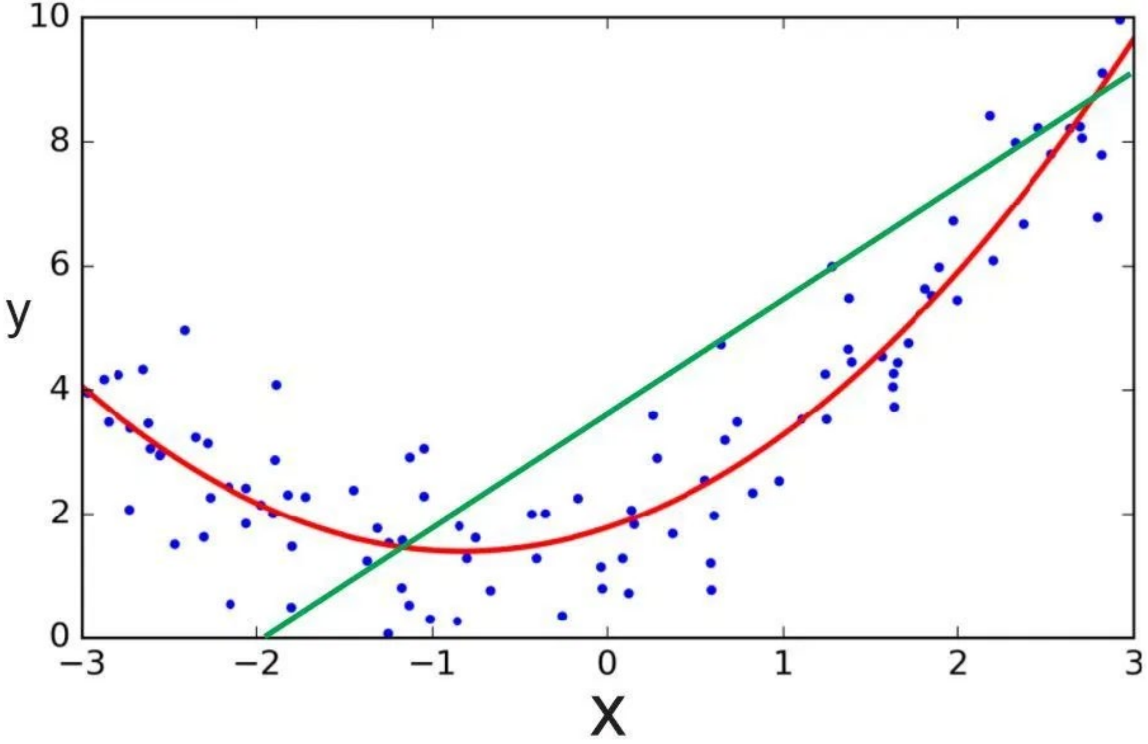
Polynomial regression

- Polynomial Regression is a form of linear regression in which the relationship between the independent variable x and dependent variable y is modelled as an *nth* degree polynomial. Polynomial regression fits a nonlinear relationship between the value of x and the corresponding conditional mean of y, denoted E(y |x)

$$y_i = \beta_0 + \beta_1\tilde{X}_i + \beta_2 Z_i + \beta_3\tilde{X}_i Z_i + \beta_4\tilde{X}_i^2 + \beta_5\tilde{X}_i^2 Z_i + e_i$$

where:

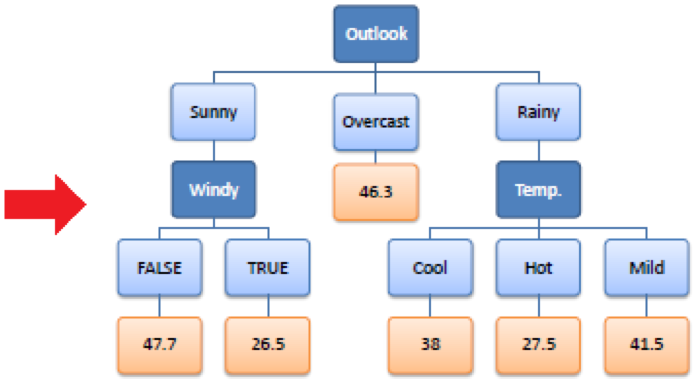
y_i = outcome score for the i^{th} unit
β_0 = coefficient for the *intercept*
β_1 = linear pretest coefficient
β_2 = mean difference for treatment
β_3 = linear interaction
β_4 = quadratic pretest coefficient
β_5 = quadratic interaction
X_i = transformed pretest
Z_i = dummy variable for treatment(0 = control, 1= treatment)
e_i = residual for the i^{th} unit


10
8
6
4
2
0
y
−3
−2
−1
0
1
2
3
X

Decision tree regression

Decision tree builds regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node (e.g., Outlook) has two or more branches (e.g., Sunny, Overcast and Rainy), each representing values for the attribute tested. Leaf node (e.g., Hours Played) represents a decision on the numerical target. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

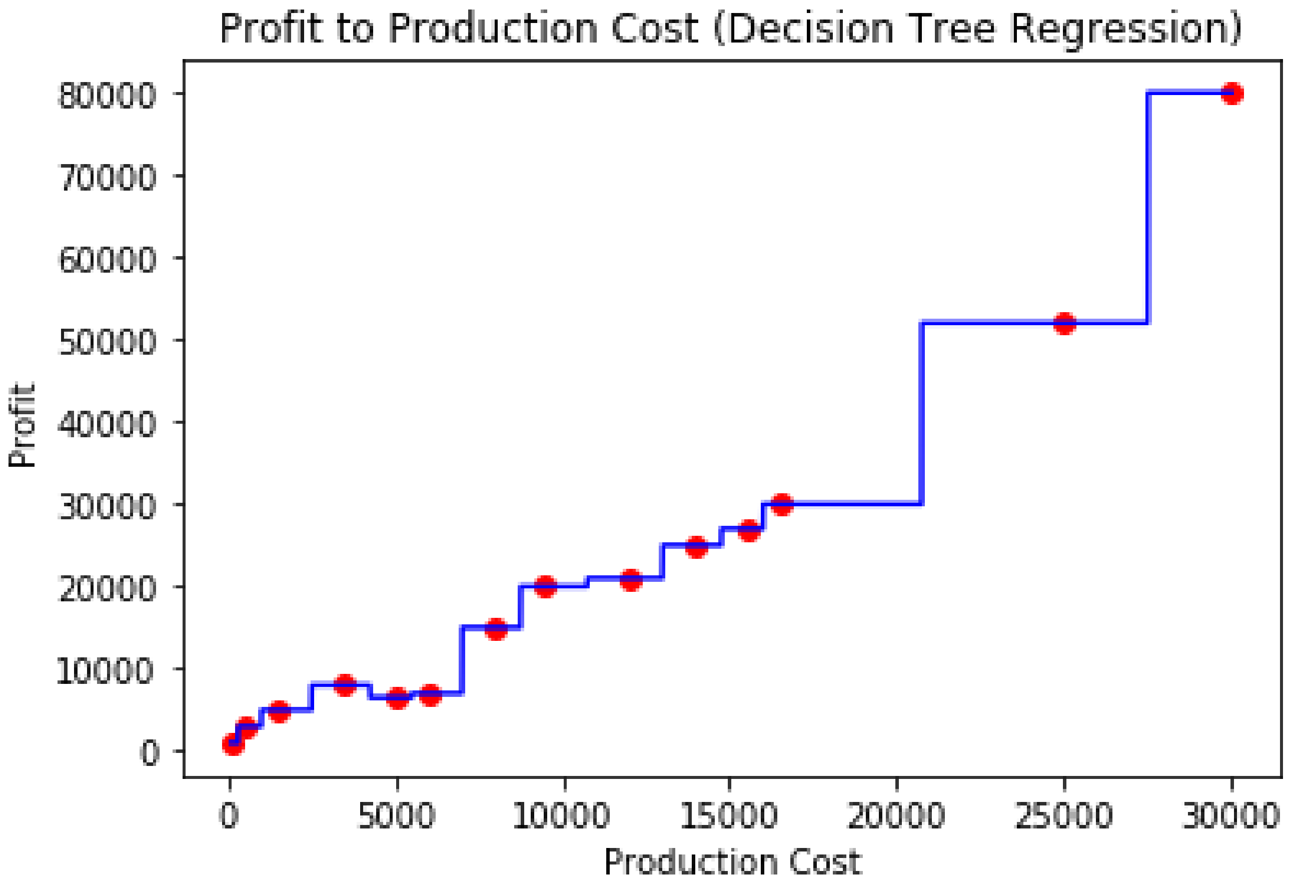
Outlook	Temp.	Humidity	Windy	Hours Played
Rainy	Hot	High	False	25
Rainy	Hot	High	True	30
Overcast	Hot	High	False	46
Sunny	Mild	High	False	45
Sunny	Cool	Normal	False	52
Sunny	Cool	Normal	True	23
Overcast	Cool	Normal	True	43
Rainy	Mild	High	False	35
Rainy	Cool	Normal	False	38
Sunny	Mild	Normal	False	46
Rainy	Mild	Normal	True	48
Overcast	Mild	High	True	52
Overcast	Hot	Normal	False	44
Sunny	Mild	High	True	30



Decision tree regression observes features of an object and trains a model in the structure of a tree to predict data in the future to produce meaningful continuous output. Continuous output means that the output/result is not discrete, i.e., it is not represented just by a discrete, known set of numbers or values.

Discrete output example: A weather prediction model that predicts whether or not there'll be rain in a particular day.

Continuous output example: A profit prediction model that states the probable profit that can be generated from the sale of a product.


Profit to Production Cost (Decision Tree Regression)
Profit
Production Cost
80000
70000
60000
50000
40000
30000
20000
10000
0
5000
10000
15000
20000
25000
30000

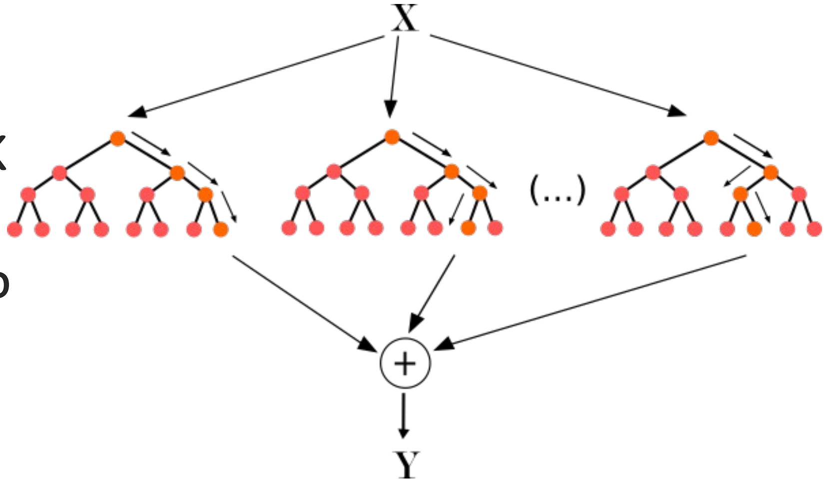
Random forest regression

The Random Forest is one of the most effective machine learning models for predictive analytics, making it an industrial workhorse for machine learning.

The random forest model is a type of additive model that makes predictions by combining decisions from a sequence of base models. Here, each base classifier is a simple decision tree. This broad technique of using multiple models to obtain better predictive performance is called model ensembling. In random forests, all the base models are constructed independently using a different subsample of the data

Approach

- Pick at random K data points from the training set.
- Build the decision tree associated with those K data points.
- Choose the number Ntree of trees you want to build and repeat step 1 & 2.
- For a new data point, make each one of your Ntree trees predict the value of Y for the data point, and assign the new data point the average across all of the predicted Y values.



Pros and cons

Regression model	**Pros**	**Cons**
Linear regression	Works on any size of dataset, gives information about features.	The Linear regression assumptions.
Polynomial regression	Works on any size of dataset, works very well on non linear problems	Need to choose right polynomial degree for Good bias and trade off.
Decision tree recession	Interpretability, no need for feature scaling, works on both linear and non linear problems	Poor results on small datasets, overfitting can easily occur
Random forest regression	Powerful and accurate, good performance many problems, including non linear	No Interpretability , overfitting can easily occur, need to choose number of trees

Logistic regression
Not for regression, it’s for Classification

In statistics, the logistic model is used to model the probability of a certain class or event existing such as pass/fail, win/lose, alive/dead or healthy/sick. This can be extended to model several classes of events such as determining whether an image contains a cat, dog, lion, etc

Based on the number of categories, Logistic regression can be classified as:

binomial: Target variable can have only 2 possible types: “0” or “1” which may represent “win” vs “loss”, “pass” vs “fail”, “dead” vs “alive”, etc.

multinomial: Target variable can have 3 or more possible types which are not ordered(i.e. types have no quantitative significance) like “disease A” vs “disease B” vs “disease C”.

ordinal: It deals with target variables with ordered categories. For example, a test score can be categorized as: “very poor”, “poor”, “good”, “very good”. Here, each category can be given a score like 0, 1, 2, 3.