Crash course - Data Science

Extraction of knowledge from large volume of d
that might be structured or unstructured.

Steps in Datascience

1. Descriptive statistics.
   clustering and basic data summery.
2. Inferance.
   Making conclusions about populations from samples.
   Estimation
   confidence intervales
   hypothesis test
   Variability.
3. Prediction / classification
   Cross validation.
4. Experimental Design
   Randomization / sample testing
   To get better generalizability of the result to population.

Page:
Date: / /

Machine learning

1) Unsupervised learning

- Take input and return a prediction.

Nothing to train the algorithm.

Supervised learning

Predictions based on data observed

=> observed outcome and predict when it's not observed.

- Evaluate ~~performance~~ results based on prediction performance.

=> Traditional statistics

Sample generalized to superpopulation

No predictions.

Software Engineering

=> Procedure for standardizing / building models or procedure.

=> Interface between I/p & O/p.

I/p or O/p are not that important

Interface is important.

Page:
Date: / /

Random Forest Regression

- Ensemble learning
  Same algorithm multiple times and put them together

- Take prediction from N Decision trees and make them predict together.
  → Average - stable (change in one tree won't impact much of trees)
  - More accurate prediction.

Average of all predictions.

Regression trees.

1) Algorithm will create a split in information entropy plot
   ↓
   if split is increasing amount of information we have about the plot.
   less than 5% of data in each leaf and then algorithm will stop.
   ↓
   Taking averages of values from each leaf after splits.
   And assigning dependent value as the average.

→ Multiple trees hence more accurate value.

Information Entropy

$$H = \sum_{i=1}^{n} p_i \log_2 \left(\frac{1}{p_i}\right)$$

outputs of datascience experiment.

1. Reports
2. Presentations
3. Interactive web pages
4. Apps

—

least square error function

$$\sum_{t=1}^{n} (y_t - \hat{y}_t)^2$$

Gradient descent.

— ID3 algorithm. ← To construct random forest

Measure the purity of split.

— Information Gain.

$$= \underbrace{H(S)}_{\text{old}} - \sum \frac{|S_v|}{|S|} \underbrace{H(S_v)}_{\text{new}}$$

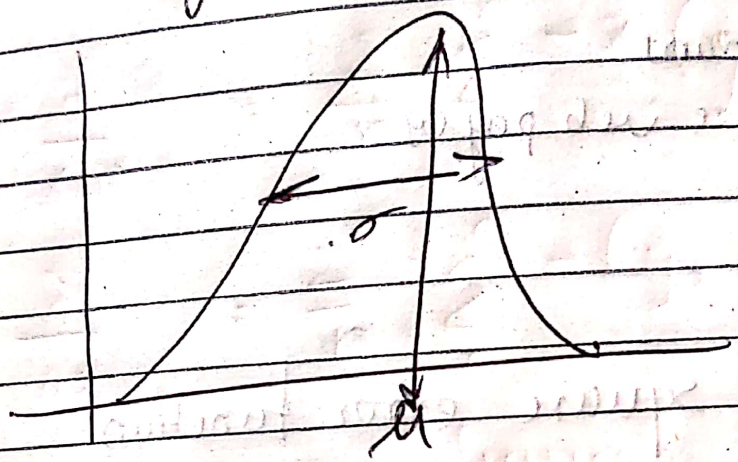
Most clarity with highest gain.

Date: / /

Gain Ratio = $\frac{\text{Gain}}{\text{Split Entropy}(S, A)}$

Split Entropy $(S, A) = -\sum \frac{|S_v|}{|S|} \log \frac{|S_v|}{|S|}$

Linear Algebra



$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{ \frac{-(x-\mu)^2}{2\sigma^2} \right\}$$

Approximation of goodness

↓

prediction.

- Add attribute as vector.

Common way of measuring information gain:

decrease in C

Gini impurity

$$I(f) = \sum_{k=1}^{K} f_k (1 - f_k)$$

$$f_k = \frac{1}{|S|} \sum_{i \in S} \mathbb{I}[y_i = c_k]$$

set of observations

$$S_\theta^k = \{ i : x_i^j > \tau, \forall i \in S \}$$

x_i^j ← value of j^{th} feature for i^{th} observation

* Optimization for the split

$$\theta^* = \arg\max \; n_S I(S) - n_L I(S_\theta^L) - n_R I(S_\theta^R)$$

- Binary Space partitioning trees.
- ABT

Gradient Descent & Cost function

1) Prediction function:
Finding a function with the use of I/p & O/p values.
Best fit line of scatter plot.

Mean square error $\rightarrow \frac{1}{n} \sum_{i=1}^{n} (\Delta i)^2$

n ← number of Data points

↑ To reduce -ve problem

↑ Cost function.

$$= \frac{1}{n} \sum_{i=1}^{n} (y_i - (mx_i + b))^2$$

2) By iteration you can find a proper value of function.

- learning rate in conjunction with the slope to reach minimum point on cost function

$$\text{Cost}_{\text{fun}^n} = \frac{1}{n} \sum_{1}^{n} (y_i - (mx_i + b))^2$$

$$\frac{d}{dm} = \frac{2}{n} \sum_{1}^{n} -x_i (y_i - (mx_i + b))$$

$$\frac{d}{db} = \frac{2}{n} \sum_{1}^{n} -(y_i - (mx_i + b))$$

$m = m - \text{learning rate} * \frac{d}{dm}$ ← to get next point

np.array ← faster than usual.

- Algorithm to perform split
  - Linear discriminant analysis (LDA)
  - canonical correlation analysis (CCA)
  - logistic regression.

Supervised learning - overfit noise
reduce diversity.

- Carver strategy.

Hyperparameter is a parameter whose value is used to control the learning process.

CCF canonical correlations forest

&

- Clustering

Flat - Group number devided

Hierchial - Number of grops devided by machine

Given feature set → machine decides the clustering.

-> k means. — Flat

-> Mean shift — Hierchial

1) Find centroids

2) find feature distance from the centroid

3) take mean of all the centroids And get new centroids

4) keep repeating process untill centroids are not moving anymore.

↓

Thn supervised learning