37:07

Decision_ Tree_ Handson Theory

* In Supervised learning → Regression

- Regression
- Logistic Regression.

* if output variable is yes or not mean two only It is Call binary classification.

→ if output is more than two its called multi classification.

→ The benefit Of Decision Tree is it Can handle both numerical and categorical data as.

For example

Age → Numerical

Salary → Numerical & continuous value.

Gender → Classification.

Occupation → Classification

→ Note !-

Output here is binary Classification.

formula Of entropy:-

$$= \sum_{i=1}^{2} P_i \log_2 P_i$$

how to calculate entropy values.

$$-[0.5 \log_2 0.5 + 0.5 \log_2 0.5]$$

→ In Machine learning entropy exist as a mesures of uncertainity

for Classification data

↓

ML Model

→ (i) Logistic Regression

→ (ii) Decision Tree

→ (iii) Decision Tree Classifier.

1.5cy [illegible]

HOU:U7. teaching MCP

→

Calculation Entropy

→ Decision Tree works on the principal or homogenuitney or purity.

→ Decision ~~one~~ tree overcome the limitation of Logistic Regression.

* It, is Create only one decision boundary So Decision trees come into the picture. there are Several data which need many decision boundary.

→ Graph of data is for Algorithm, and tree is for us.

→ ID3 (Iterative Dichomizer 3) → old → best for classification data.

→ if the less entrophy is there means in data more homogenity is there

→ if high them less homegenity is in the data

→ entropy is the measure of Randomness or Chaus, mixture.

53.54

→ One drawback of decision tree is it 90% Stuck with overfitting.

And we have to Control it.

→ Stracity is very usefue in Case of inbalanced data.

Byy Degree M.com

→ Sections/ Box/ Decision boundary

→ overlearning is also known as overfitting.

→ We should always ensure that our training accuracy never ever be 100%.

→ Whenever extra node added into the tree the it will be always bottom of tree

→ Confusion matrix

PREDICTED

ACTUAL

165	NO	Yes.
NO	TN (50)	FP (10)
yes	FN (5)	TP (100)

Accuracy Score $= \frac{TP + TN}{Total}$

$= \frac{50+100}{165}$

$= 0.91$

Error Rate $= 1 -$ accuracy

$= 1 - 0.91$

$= 0.09$

or $= \frac{FP + FN}{TN+FP+FN+TP}$ (Total)

$= \frac{10+5}{165}$

$= 0.09$

Recall :- $\frac{TP}{\text{actual yes}}$

$= \frac{100}{105}$

$= 0.95$

precision $= \frac{TP}{\text{predicted yes}}$

$= \frac{100}{FP+TP}$

$= \frac{100}{10+100}$

$= \frac{100}{110}$

$= 0.64$

To develop such a model, the computed information gain (C, pitch) with respect to target is ....

(rounded off to two decimal places).

Match No	Pitch	Format	Winner (Target)
1	S	T	Green
2	S	T	Blue
3	F	O	Blue
4	S	O	~~Blue~~ Blue
5	F	T	~~Blue~~ Green
6	F	O	~~Green~~ Blue
7	S	O	~~Blue~~ Green
8	F	T	~~Green~~ Blue
9	F	O	~~Blue~~ Blue
10	S	O	~~Blue~~ Green.
			~~G~~

Entropy Calculation

$$\text{Entropy}(S) = -P(\text{yes}) \log_2 P(\text{yes}) - P(\text{no}) \log_2 (P\text{no})$$

where;

— S is the total sample space

P (yes) is probability of yes.

if number of yes = number of no ie $P(S) = 0.5$

⇒ Entropy (S) = 1

if it contains all yes or all no ie $(P(S) = 1$ or 0

⇒ Entropy (S) = 0

$E(S) = -P(ye)\log_2 P(yes) - P(no)\log_2 (P(no))$

$= -0.5\log_2 0.5 - 0.5\log_2 0.5$

$= -0.5(\log_2 0.5 - \log_2 0.5)$

$=$

$= 1$

- Step 1: Entropy of entire datasets

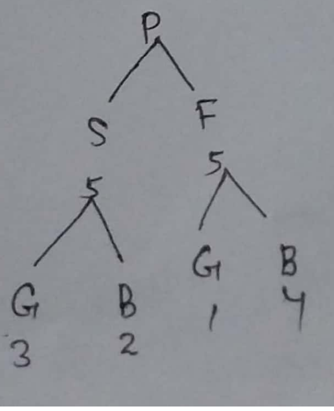
$C\{4,6\} = -\frac{4}{10}\log_2\frac{4}{10} - \frac{6}{10}\log_2\frac{6}{10} = 0.97$

Step 2: Entropy of all attributes in pitch

$\{3,2\} = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0.97$

Entropy of Spin (S)

IG (P) = entropy (whole)



BARBU U MANG
0 7 8 601000
16603

Progres SQL

Micrsoft aue traing, Dont but
free. → Check now

Entropy of pace (F)

$\{1, 4\} = -\frac{1}{5}\log_2\frac{1}{5} - \frac{4}{5}\log_2\frac{4}{5} = 0.72$

Information Gain (C, pitch) = $\text{Entropy}(C) - \frac{5}{10}\text{Ent}(S) - \frac{5}{10}\text{Ent}(F)$

$= 0.97 - 0.485 - 0.360$

$= 0.13$

* what is gini- Index

Ginin idex measures the impurity of the data.

1:38:09

higher Gini meas more impurity.