Mutual Information Scores (Feature Selection)

UTKARSH GAIKWAD

CLASS STARTING SHARP AT 6:03 PM

Calculating Mutual Info scores

Study Hours	Pass/Fail
2	Pass
4	Pass
1	Fail
3	Pass
5	Fail
2	Fail
6	Pass
3	Fail
2	Fail
4	Fail

P(Pass)	0.4
P(Fail)	0.6

P(H=1)	0.1
P(H=2)	0.3
P(H=3)	0.2
P(H=4)	0.2
P(H=5)	0.1
P(H=6)	0.1

P(H=1, Pass)	0
P(H=2,Pass)	0.1
P(H=3,Pass)	0.1
P(H=4,Pass)	0.1
P(H=5, Pass)	0
P(H=6,Pass)	0.1

P(H=1, Fail)	0.1
P(H=2, Fail)	0.2
P(H=3, Fail)	0.1
P(H=4, Fail)	0.1
P(H=5, Fail)	0.1
P(H=6, Fail)	0

Mutual Info Score = 0.2954

MI(Study Hours, Pass/Fail) = Σ P(Study Hours = x, Pass/Fail = y) * log2(P(Study Hours = x, Pass/Fail = y) / (P(Study Hours = x) * P(Pass/Fail = y)))

Sklearn mutual information scores

```
from sklearn.feature_selection import mutual_info_classif
mf = mutual_info_classif(X_pre,Y)
```

```
from sklearn.feature_selection import mutual_info_regression
mf = mutual_info_regression(X_pre,Y)
```

Two separate pipelines

- Feature Selection Pipeline : Categorical variables apply Ordinal Encoder
- Final Pipeline : Categorical variables apply OneHotEncoder

Thank you

PING ME ON SKYPE FOR ANY QUERIES