Date __/__/__
Page

Lab 3: ID3 algorithm

What is ID3 algorithm?

→ The ID3 is a popular decision tree algorithm use in machine learning
→ It aims to build a decision tree by iteratively selecting the best attribute to split the data based on the info gain
→ It's a greedy algorithm and best for features that categorial compared to continuous

ID3 Metrics

The metrics are entropy and information gain.

Entropy

- also referred to as gini impurity, it measures randomness in the datasets
- high entropy data sets are evenly distributed across all categories where low entropy ones have data concentrated at few particular points

$$H(S) = \sum -(P_i \cdot \log_2(P_i))$$

S ← current dataset
i → set of classes in S

Information gain

- it assesses how much value info an attribute can provide.
- We select attribute with highest info gain

$$IG(A, D) = H(S) - \sum_{i=1}^{v} \frac{|S_v|}{|S|} * H(S_v)$$

Pseudocode

```
def ID3(D, A):
    if D is pure. or A is empty:
        return a leaf node with the
        majority class in D
    else:
        A_best = argmax(info(D, A))
        root = Node(A_best)
        for v in values(A_best)
            D_v = subset(D, A_best, v)
            child = ID3(D_v, A - {A_best})
            root.add_child(v, child)
    return root
```

Code

```
import pandas as pd
import numpy as np
import seaborn as sb
from go...

df = pd.read_csv('play_tennis.csv')

# label encoding

from sklearn.preprocessing import
LabelEncoder

le = LabelEncoder()

for col in df.columns
    if col != 'day':
        df[col] = le.fit_transform
                    (df[col])

X = df.drop(['day', 'play'], axis=1)
y = df.play

from sklearn.model_selection
import train_test_split
X_train, X_test, y_train, y_test
= train_test_split(X, y, test_size=0.
```

Date __/__/__
Page

```
from sklearn.tree import
DecisionTreeClassifier.

tree = DecisionTreeClassifier()
tree.fit(x_train, y_train)

tree.score(x_test, y_test)
→ 0.668
```