

COP528 Applied Machine Learning (LABs)

LAB Day 05. Feature selection & Decision tree

Introduction

The aim of this session is to understand decision tree method. The experiments include 4 tasks.

Task 01. Manually build a decision tree

A training set of data is provided, please calculate the MI of each split and build a decision tree manually.

(1) Please calculate the MI of each part of the given data set.

Outlook	Temperature	Humidity	Wind	PlayTennis
sunny	hot	high	weak	No
sunny	hot	high	strong	No
overcast	hot	high	weak	Yes
rain	mild	high	weak	Yes
rain	cool	normal	weak	Yes
rain	cool	normal	strong	No
overcast	cool	normal	strong	Yes
sunny	mild	high	weak	No
sunny	cool	normal	weak	Yes
rain	mild	normal	weak	Yes
sunny	mild	normal	strong	Yes
overcast	mild	high	strong	Yes
overcast	hot	normal	weak	Yes
rain	mild	high	strong	No

(2) Please manually build a decision tree based on this dataset.

Task 02. Use the decision tree classifier in Sklearn to verify the tree you build from the previous practice.

Hint: The tree built by Sklearn might be different from the one built manually.

Think about why and compare them. The display function 'plot_tree()' in Sklearn and 'savefig()' in matplotlib maybe helpful.

An example of using decision tree in sklearn is given as:

```
from sklearn.tree import DecisionTreeClassifier
model = DecisionTreeClassifier(criterion='entropy', random_state=0, max_depth=3, min_samples_leaf=2)
model.fit(X,Y) #train your classifier using preprocessed data, saying attributes (X) and label (Y)
```

The data can be imported as follows

```
import pandas as pd
tennis = pd.read_csv('tennis.csv', header=0)
```

Task 03. Decision tree on Wine dataset

In the lab session of Day 03, you have made some data processing on the Wine dataset. Your task today is to use decision tree to make classification models to classify wines into 3 classes. Compare the results with a decision tree classifier based on the evaluation methods learned at the first day.

Here is the code to load the data and convert them into Pandas. You could use the numpy array with sklearn directly:

```
from sklearn.datasets import load_wine
import pandas as pd
wine = load_wine()
winedata = pd.DataFrame(data= wine.data, columns=wine.feature_names)
winedata['target'] = pd.Series(wine.target)
winedata.head(5)
```

Task 04. Decision tree on breast cancer dataset

The breast cancer set is from the Institute of Oncology, University Medical Centre, Ljubljana, Yugoslavia, and was provided by M. Zwitter and M. Soklic, both with thanks. The dataset is a classical dichotomous dataset. Use this to build a decision tree and evaluate it.

```
from sklearn.datasets import load_breast_cancer
dataset = load_breast_cancer()
df= pd.DataFrame(data= dataset.data, columns=dataset.feature_names)
df['target'] = pd.Series(dataset.target)
df.head(5)
features, target = load_breast_cancer(return_X_y=True)
```