Data Science

Lab 5

Ensemble Learning
& Clustering Algorithm

Ok-Ran Jeong

Contents

1. Ensemble Learning
   - Bagging method with decision tree algorithm
   - Aggregation Method: majority voting 결정할 땐 다수결로
   - Evaluation Metric: Confusion Matrix

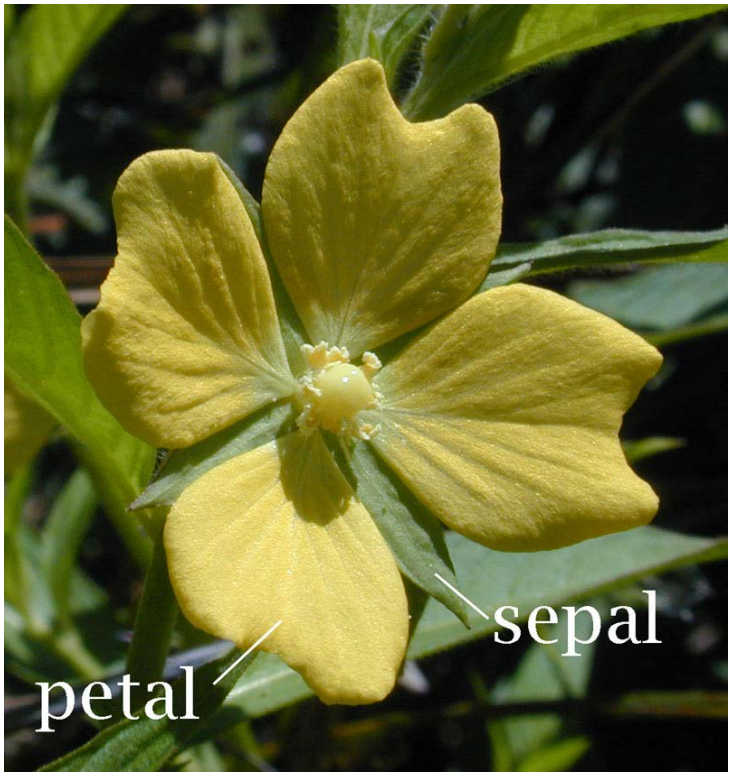
2. Clustering algorithm
   - K-Means Clustering
   - Measure: Euclidean Distance

Problem 1: Ensemble Learning

- Using bagging method with decision tree algorithm, predict with voting method and calculate the accuracy using confusion matrix.
- Dataset
  - Iris-bagging dataset with 30 recodes
  - Attributes: sepal length, sepal width, petal length, petal width
- Bagging and Evaluation
  - Generate decision tree model with DecisionTreeClassifier function
  - Run 10 bagging rounds
  - Predict the label using voting
  - Calculate the accuracy using confusion matrix

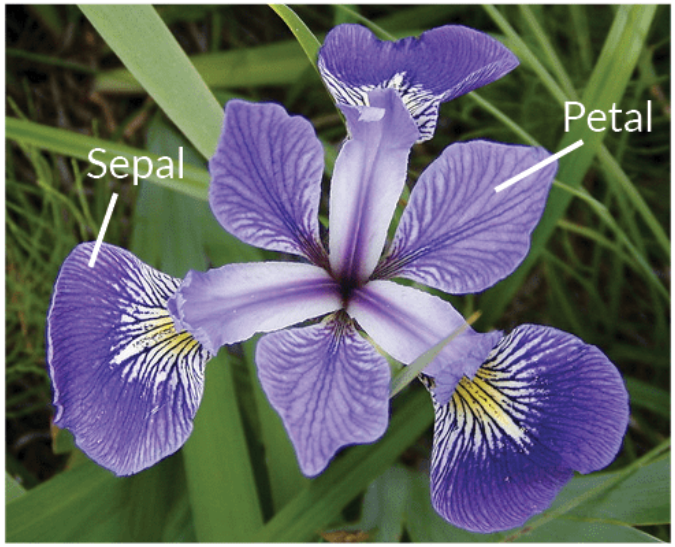
Fun Facts (1/2)

- The word **Iris** originates from the Greek word for **rainbow**.
- Iris is the flower of the Greek goddess Iris who is the messenger of love.



Fun Facts (2/2)

- Three of the species of iris
  - Versicolor, Setosa, Virginica
- The iris dataset was first used for multivariate discriminant analysis by Ronald Fisher in 1936.



Iris Versicolor

Iris Setosa

Iris Virginica

Import libraries and data file

```
# Bagging
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings(action='ignore')
```

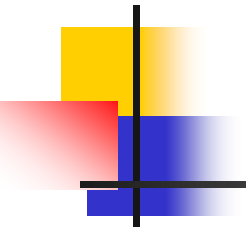
```
for j in range(len(compare)):
  if (compare['labels'][j] == 0): compare['labels'][j] = 2
  elif (compare['labels'][j] == 1): compare['labels'][j] = 0
  else: compare['labels'][j] = 1
```

```
iris = pd.read_csv('Iris.csv', encoding='utf-8')
labels = iris['Species']
iris.head()
```

	Id	SepalLengthCm	SepalWidthCm	PetalLengthCm	PetalWidthCm	Species
0	1	5.1	3.5	1.4	0.2	Iris-setosa
1	2	4.9	3.0	1.4	0.2	Iris-setosa
2	3	4.7	3.2	1.3	0.2	Iris-setosa
3	4	4.6	3.1	1.5	0.2	Iris-setosa
4	5	5.0	3.6	1.4	0.2	Iris-setosa

10 data sets for bagging rounds

```
# Load iris dataset samples
samples = []
sample = pd.read_csv('Iris_bagging_dataset (1).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (2).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (3).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (4).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (5).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (6).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (7).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (8).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (9).csv', encoding='utf-8')
samples.append(sample)
sample = pd.read_csv('Iris_bagging_dataset (10).csv', encoding='utf-8')
samples.append(sample)
```



Hints

- Import sklearn modules (tree, metrics)
- DecisionTreeClassifier
- confusion_matrix, classification_report

Problem 2: Clustering algorithm

- Using K-Means clustering algorithm, group the given dataset into 3 clusters and evaluate the accuracy.
- Dataset
  - Iris dataset with 150 recodes
  - Features: sepal length, sepal width, petal length, petal width
- Computing and Evaluation
  - Compute the Euclidean Distance
  - Compare the actual cluster label with example cluster labels

결과와 비교해보기

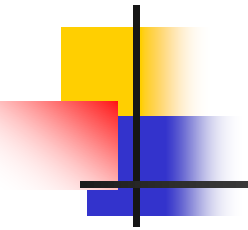
Import libraries and data file

```
# k-means
import pandas as pd
import numpy as np
import warnings
warnings.filterwarnings(action='ignore')

from sklearn import datasets
iris = datasets.load_iris()
# setosa, versicolor, virginica
```

```
labels = pd.DataFrame(iris.target)
labels.columns=['labels']

data = pd.DataFrame(iris.data)
data.columns=['Sepal length','Sepal width','Petal length','Petal width']
data = pd.concat([data,labels],axis=1)
```



End of lab