Department of
Computer Science and Engineering

Title: Ensemble Learning

Green University of Bangladesh

1 Objective(s)

- To gather knowledge of ensemble learning.

2 Ensemble Learning

Ensemble means a group of elements viewed as a whole rather than individually. An Ensemble method creates multiple models and combines them to solve it. Ensemble methods help to improve the robustness/generalization of the model. In this article, we will discuss some methods with their implementation in Python.

3 Types of Ensemble Learning

Basic Ensemble Techniques

- Max Voting
- Averaging
- Weighted Average

Advanced Ensemble Techniques

- Bagging
- Stacking
- Blending
- Boosting

Algorithms based on Bagging and Boosting

- Bagging meta-estimator
- Random Forest
- AdaBoost
- GBM

3.1 Basic Ensemble Methods

Multiple predictions are made for each data point in averaging. In this method, we take an average of predictions from all the models and use it to make the final prediction. Averaging can be used for making predictions in regression problems or while calculating probabilities for classification problems.

3.2 Averaging Method and Max Voting Method:

```
1  #importlibrary
2  import pandas as pd
3  import numpy as np
4  from sklearn.model_selection import train_test_split
5  from sklearn.metrics import mean_squared_error
6  from sklearn import tree
7  from sklearn.linear_model import LinearRegression
8  from sklearn.linear_model import LogisticRegression
9  from sklearn.svm import  SVC
10
11 ----------------------
12 #import dataset
```

```
from google.colab import files
f=files.upload()

-----------------------
#read the dataframe
df=pd.read_csv("diabetes.csv")
df.head()
```

Output Top five rows of the data frame is shown:

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1

Figure 1: Top five rows of the dataset

```
#Defining feature and corresponding class label
x=df.drop("Outcome",axis="columns")
y=df.Outcome

-----------------------
#training set and testing set split
X_train, X_test, y_train, y_test = train_test_split(x,y, test_size=0.20)

# initializing all the model objects with default parameters
model_1 = LogisticRegression()
model_2 = SVC()
model_3 = tree.DecisionTreeClassifier()
-----------------------
## training all the model on the training dataset
model_1.fit(X_train, y_train)
model_2.fit(X_train, y_train)
model_3.fit(X_train, y_train)
-----------------------
# predicting the output on the validation dataset
pred_1 = model_1.predict(X_test)
pred_2 = model_2.predict(X_test)
pred_3 = model_3.predict(X_test)
-----------------------
#model score/accuracy
model_1.score(X_test,y_test)
model_2.score(X_test,y_test)
model_3.score(X_test,y_test)
-----------------------
# final prediction after averaging on the prediction of all 3 models
pred_final = (pred_1+pred_2+pred_3)/3.0

# printing the root mean squared error between real value and predicted value
print(mean_squared_error(y_test, pred_final))
-----------------------
```

```
# Making the final model using voting classifier
from sklearn.ensemble import VotingClassifier
final_model = VotingClassifier(estimators=[('lr', model_1), ('svm', model_2),('
    dt',model_3)], voting='hard')
-----------------------
# training all the model on the train dataset
final_model.fit(X_train, y_train)
-----------------------
# predicting the output on the test dataset
pred_final = final_model.predict(X_test)
-----------------------
# printing log loss between actual and predicted value
print(log_loss(y_test, pred_final))
```

3.3 Lab Task (Please implement yourself and show the output to the instructor)

- Implement weighted average ensemble technique on the same data set and analyse the output.

3.4 Advanced Ensemble Methods

Above are simple techniques, now let's take a look at advanced techniques for ensemble learning.

3.4.1 Bagging

It is also known as a bootstrapping method. Base models are run on bags to get a fair distribution of the whole dataset. A bag is a subset of the dataset along with a replacement to make the size of the bag the same as the whole dataset. The final output is formed after combining the output of all base models.

```
#importlibrary
import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error
from sklearn import tree
from sklearn.linear_model import LinearRegression
from sklearn.linear_model import LogisticRegression
from sklearn.svm import  SVC

-----------------------
#import dataset
from google.colab import files
f=files.upload()

-----------------------
#read the dataframe
df=pd.read_csv("diabetes.csv")

-----------------------
#Defining feature and corresponding class label
x=df.drop("Outcome",axis="columns")
y=df.Outcome

-----------------------
#training set and testing set split
X_train, X_test, y_train, y_test = train_test_split(x,y, test_size=0.20)

-----------------------
# initializing the bagging model using decision tree as base model
from sklearn.ensemble import BaggingClassifier
```

```
32
33 bag_model = BaggingClassifier(
34     base_estimator=tree.DecisionTreeClassifier(),
35     n_estimators=100,
36     max_samples=0.8,
37     oob_score=True,
38     random_state=0
39 )
40 bag_model.fit(X_train, y_train)
41 bag_model.oob_score_
42 ------------------------
43 bag_model.score(X_test, y_test)
44 ------------------------
45 #k-fold cross validation
46 from sklearn.model_selection import cross_val_score
47 scores = cross_val_score(tree.DecisionTreeClassifier(), x, y, cv=5)
48 scores
```

4 Conclusion

Based on the focused objective(s), we will learn about the different ensemble methods and comparison between ensemble methods.

5 Lab Exercise (Submit as a report)

- Implement stacking ensemble and AdaBoost ensemble on the same dataset

6 Policy

Copying from internet, classmate, seniors, or from any other source is strongly prohibited. 100% marks will be *deducted* if any such copying is detected.