0711239_HW3_code

Name: 李勝維

Student ID: 0711239

Homework 3

Part 1, Coding:

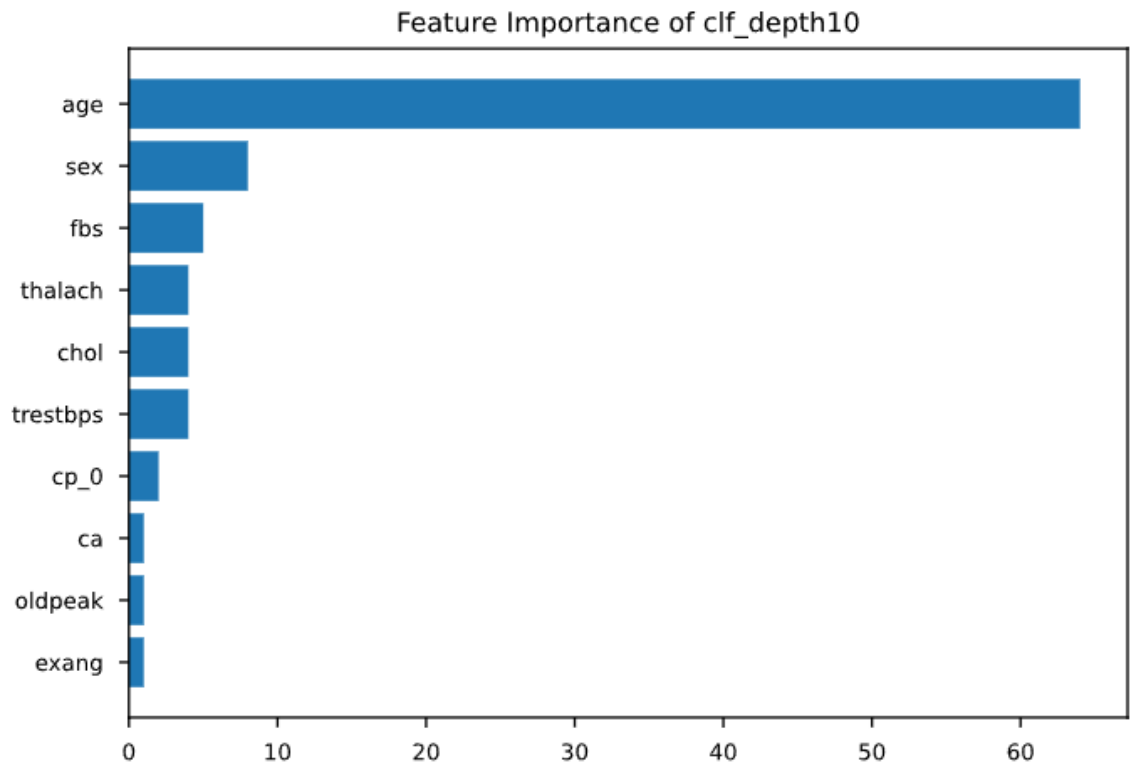
1. Given array: [1,2,1,1,1,1,2,2,1,1,2]
    a. Gini index = 0.4628099173553719
    b. Entropy = 0.9456603046006401
2. Implement CART with arguments Criterion & Max_depth
    a. Accuracy of different depths:

```
Q2.1
criterion=gini, max_depth=3: 0.77
criterion=gini, max_depth=10: 0.72
```

    b. Accuracy of different criterions:

```
Q2.2
criterion=gini, max_depth=3: 0.77
criterion=entropy, max_depth=3: 0.79
```

3. Feature importance (by counting each node) of clf_depth10:



4. Implement AdaBoost with argument N_estimators
    a. Accuracy of different n_estimators:
    (criterion='gini', max_depth=3)

```
Q4.1
n_estimators=10, max_depth=3: 0.79
n_estimators=100, max_depth=3: 0.78
```

5. Implement Random Forest with arguments N_estimators, Max_features, Bootstrap
    a. Accuracy of different n_estimators:
    (criterion='gini', max_depth=None, max_features=sqrt(n_features), Bootstrap=True)

```
Q5.1
n_estimators=10: 0.72
n_estimators=100: 0.77
```

b. Accuracy of different max_features:
(criterion='gini', max_depth=None, n_estimators=10, Bootstrap=True)

```
Q5.2
max_features=sqrt(n_features): 0.77
max_features=n_features: 0.74
```

6. Settings:
   a. Feature engineering:
      - Remove all catagorical features
   b. Result hyperparameters:
      - model: Random Forest
      - n_estimators: 100
      - max_features: sqrt(n_features)
      - boostrap: True
      - max_depth: 1
      - criterion: gini index

results in accuracy = 0.87

```
your_model = RandomForest(
    n_estimators=100,
    max_features=np.sqrt(x_train.shape[1]),
    max_depth=1
)
your_model.fit(x_train, y_train)
y_pred = your_model.predict(x_test)

print(f'Test-set accuarcy score: {accuracy_score(y_test, y_pred)}')

[209] ✓ 0.5s Python
··· Test-set accuarcy score: 0.87
```

Part 2, Questions:

Q1 ① Evaluate misclassification rate

Tree A : node 1 predicts C_1

⇒ 300 correct, 100 misclassified

node 2 predicts C_2

⇒ 300 correct, 100 misclassified

misclassification rate $= \frac{200}{800} = 0.25$

Tree B : node 1 predicts C_2

⇒ 400 correct, 200 misclassified

node 2 predicts C_1

⇒ 200 correct, 0 misclassified

misclassification rate $= \frac{200}{800} = 0.25$ equal

② Calculate cross-entropy :

Tree A : $P_1 = \frac{400}{800} = 0.5$, $P_2 = \frac{400}{800} = 0.5$

$E = -0.5 \log_2 0.5 - 0.5 \log_2 0.5 = 1$

Tree B : $P_1 = \frac{600}{800} = 0.75$, $P_2 = \frac{200}{800} = 0.25$

$E = -0.75 \log_2 0.75 - 0.25 \log_2 0.25 \approx 0.387$ lower

Q1 ③ Calculate gini index

Tree A: $G = 1 - 0.5^2 - 0.5^2 = 0.5$

Tree B: $G = 1 - 0.75^2 - 0.25^2 = 0.375$ lower

Tree A has higher cross-entropy/gini-index since it results in a "more" uniform distribution.

Q2 Let $E = \sum_t \int e^{-t y(x)} p(t|x) p(x) dx$

$$E = \int e^{-y(x)} p(t=1|x)\, p(x) dx + \int e^{y(x)} p(t=-1|x)\, p(x)\, dx$$

$$\frac{dE}{dy(x)} = \int -e^{-y(x)} p(t=1|x)\, p(x) dx + \int e^{y(x)} p(t=-1|x) p(x)\, dx = 0$$

$$\int e^{y(x)} p(t=-1|x) p(x)\, dx = \int e^{-y(x)} p(t=1|x) p(x)\, dx$$

$$e^{y(x)} p(t=-1|x) = e^{-y(x)} p(t=1|x)$$

↓ multiply by $e^{y(x)}$

$$e^{2y(x)} = \frac{p(t=1|x)}{p(t=-1|x)}, \quad y(x) = \frac{1}{2} \ln \frac{p(t=1|x)}{p(t=-1|x)}$$

#