DATA7202 : Assessment 3
Name : Peng Yu
Student ID : 46635884

Question 1

Answer:

Decision tree leads to different results according to different x_i and y_i, so the decision tree model is easy to overfit. Assume that when the first node is x_1, the result is y_1; if not, the depth of the tree is increased, and determine whether to x_2, if it is, the result is y_2, if not, to increase the depth of the tree again, and so on., when the depth of the decision tree is equal to the sample size of the training set, any training set can be fitted via a tree with zero training loss.

Question 2

Answer:

Suppose there is a c that minimizes the squared-error loss.
Then $\sum_{i=1}^{k}(y_i - c)^2 = \sum_{i=1}^{k}(y_i^2 - 2cy_i + c^2)$
$= nc^2 - 2c\sum_{i=1}^{k} y_i + \sum_{i=1}^{k} y_i^2$
The transformed expression can be equivalent to a quadratic equation of c of one variable, which is the minimum value for $c = -\frac{-2\sum_{i=1}^{k} y_i}{2n} = n^{-1}\sum_{i=1}^{,n} y_i = h^{\mathfrak{w}}(x)$.
Therefore $h^{\mathfrak{w}}(x)$ is minimizes the squared-error loss.

Question 3

Answer:

First there are 3 blue and 2 red data points in a certain tree region.
Then $P_b = \frac{3}{5}$, $P_r = \frac{2}{5}$.
Misclassification impurity:
$$Misclassification(t) = 1 - maxP(i|t) = 1 - \frac{3}{5} = \frac{2}{5}$$
Gini impurity:

$$Gini(z) = \frac{1}{2}\left(1 - \sum_{z=0}^{c-1} P_z^2\right) = \frac{1}{2} * (1 - ((\frac{3}{5})^2 + (\frac{2}{5})^2)) = \frac{6}{25}$$

Entropy impurity:

$$Entropy(z) = -\sum_{z=0}^{c-1} P_z log_2(P_z) = -\left(\frac{3}{5} * log_2\left(\frac{3}{5}\right) + \frac{2}{5} * log_2\left(\frac{2}{5}\right)\right) = 0.97095$$

Second for 2 blue and 3 red data points.

$P_b = \frac{2}{5}$, $P_r = \frac{3}{5}$

Misclassification impurity:

$$Misclassification(t) = 1 - maxP(i|t) = 1 - \frac{3}{5} = \frac{2}{5}$$

Gini impurity:

$$Gini(z) = \frac{1}{2}\left(1 - \sum_{z=0}^{c-1} P_z^2\right) = \frac{1}{2} * (1 - ((\frac{2}{5})^2 + (\frac{3}{5})^2)) = \frac{6}{25}$$

Entropy impurity:

$$Entropy(z) = -\sum_{z=0}^{c-1} P_z log_2(P_z) = -\left(\frac{2}{5} * log_2\left(\frac{2}{5}\right) + \frac{3}{5} * log_2\left(\frac{3}{5}\right)\right) = 0.97095$$

In summary, the results of “3 blue and 2 red” and “2 blue and 3 red” are the same.

Question 4

Answer:

Suppose p is that τ^* contains a fraction of the points from τ.

When n = 2, $p = \frac{1}{2} + \frac{2-1}{2} * \frac{1}{2}$

When n = 3, $p = \frac{1}{3} + \frac{3-1}{3} * \frac{1}{3} + \left(\frac{3-1}{3}\right)^2 * \frac{1}{3}$

When n = 4, $p = \frac{1}{4} + \frac{4-1}{4} * \frac{1}{4} + \left(\frac{4-1}{4}\right)^2 * \frac{1}{4} + \left(\frac{4-1}{4}\right)^3 * \frac{1}{4}$

When n = 5, $p = \frac{1}{5} + \frac{5-1}{5} * \frac{1}{5} + \left(\frac{5-1}{5}\right)^2 * \frac{1}{5} + \left(\frac{5-1}{5}\right)^3 * \frac{1}{5} + \left(\frac{5-1}{5}\right)^4 * \frac{1}{5}$

...

Therefore when n = n

$$p = \frac{1}{n} + \frac{1}{n}\left(\frac{n-1}{n}\right)^1 + \frac{1}{n}\left(\frac{n-1}{n}\right)^2 + \frac{1}{n}\left(\frac{n-1}{n}\right)^3 + \cdots + \frac{1}{n}\left(\frac{n-1}{n}\right)^{n-1}$$
$$= \frac{1}{n}\left(\left(\frac{n-1}{n}\right)^0 + \left(\frac{n-1}{n}\right)^1 + \left(\frac{n-1}{n}\right)^2 + \left(\frac{n-1}{n}\right)^3 + \cdots + \left(\frac{n-1}{n}\right)^{n-1}\right)$$
$$= \frac{1}{n} * \frac{\left(\frac{n-1}{n}\right)^0 * \left(1-\left(\frac{n-1}{n}\right)^n\right)}{1-\frac{n-1}{n}}$$
$$= 1 - \left(1 - \frac{1}{n}\right)^n$$

Therefore, (1 - p) is that τ^* does not contain a fraction of the points from τ.
When n is large

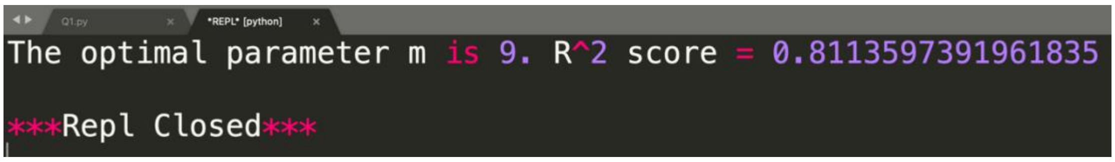
$$\lim_{n\to\infty} 1 - p = \lim_{n\to\infty} 1 - \left(1 - \left(1 - \frac{1}{n}\right)^n\right)$$
$$= \lim_{n\to\infty} \left(1 - \frac{1}{n}\right)^n$$
$$= \lim_{n\to\infty} \left(\frac{1}{1+\frac{1}{n-1}}\right)^n$$
$$= \lim_{n\to\infty} \frac{1}{\left(1+\frac{1}{n-1}\right)^{n-1}}\left(\frac{1}{1+\frac{1}{n-1}}\right) = \frac{1}{e} * 1 = e^{-1} \approx 0.37$$

From what has been discussed above that for large n, τ^* does not contain a fraction of about $e^{-1} \approx 0.37$ of the points from τ.

Question 5

Answer:

The optimal parameter m is 9. R^2 score = 0.8113597391961835.



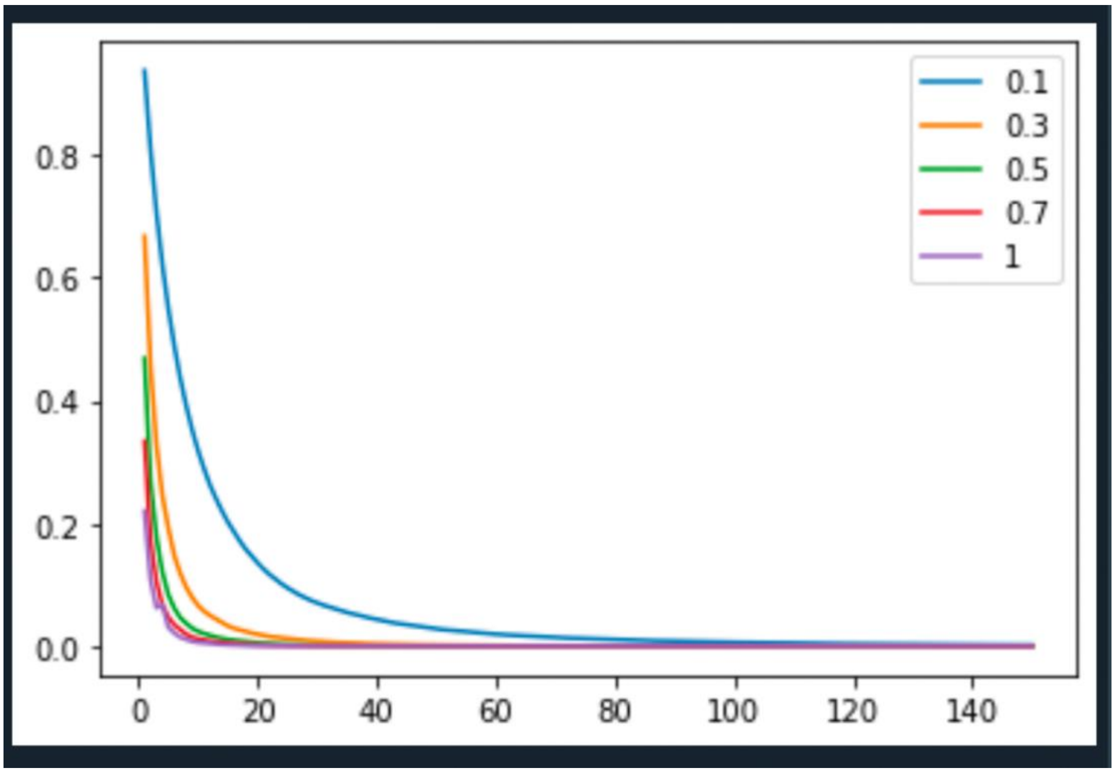
```
The optimal parameter m is 9. R^2 score = 0.8113597391961835

***Repl Closed***
```

Question 6

Answer:

As can be seen from the figure below, the smaller a is, the faster the slope of decreases with the increase of B. When the range of B is 0 to 20, the curve drops the fastest. When B is greater than 100, all the curves approach 0.



Code Appendix

Question 5

```
import numpy as np
from sklearn.datasets import make_friedman1
from sklearn.ensemble import RandomForestRegressor
from sklearn.model_selection import train_test_split
from sklearn.metrics import r2_score

# create regression problem
n_points = 1000 # points
scores=[]
x, y = make_friedman1(n_samples=n_points, n_features=15,
noise=1.0, random_state=100)
# split to train/test set
x_train, x_test, y_train, y_test = \
train_test_split(x, y, test_size=0.33, random_state=100)

#Answer:
for i in np.arange(1,15):
        rf = RandomForestRegressor(n_estimators=1000, oob_score = True,
max_features=i, random_state=100)
        rf.fit(x_train,y_train)
        yhatrf = rf.predict(x_test)
        scores.append(r2_score(y_test, yhatrf))
        print('m='+str(i)+'. r2='+str(scores[i-1]))
print('The optimal parameter m is '+ str(scores.index(max(scores))+1) + ". R^2 score =
"+ str(max(scores)))
```

Question 6

```
from sklearn.datasets import make_blobs
from sklearn.metrics import zero_one_loss
from sklearn.model_selection import train_test_split
import numpy as np
import matplotlib.pyplot as plt
from sklearn.ensemble import GradientBoostingClassifier
```

```
X_train, y_train = make_blobs(n_samples=1000, n_features=10, centers=3,
random_state=10, cluster_std=5)

#Answer:
gammas = [0.1, 0.3, 0.5, 0.7, 1]
for gamma in gammas:
    GBC = GradientBoostingClassifier(learning_rate=gamma, n_estimators = 150)
    GBC.fit(X_train, y_train)
    plt.plot(np.arange(1,151,1), GBC.train_score_, label = gamma)
plt.legend()
plt.show()
```