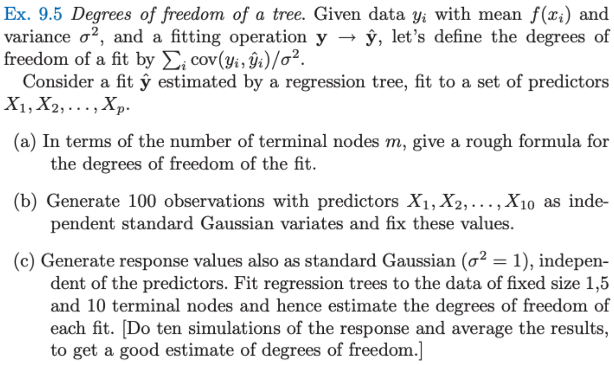
Statistical Learning: HW3

Deadline: 2023-03-10

- Yu-Ching Liao ycliao3@illinois.edu



Ex. 9.5 *Degrees of freedom of a tree.* Given data y_i with mean $f(x_i)$ and variance σ^2, and a fitting operation $\mathbf{y} \to \hat{\mathbf{y}}$, let's define the degrees of freedom of a fit by $\sum_i \operatorname{cov}(y_i, \hat{y}_i)/\sigma^2$.

Consider a fit $\hat{\mathbf{y}}$ estimated by a regression tree, fit to a set of predictors $X_1, X_2, \dots, X_p$.

(a) In terms of the number of terminal nodes m, give a rough formula for the degrees of freedom of the fit.

(b) Generate 100 observations with predictors $X_1, X_2, \dots, X_{10}$ as independent standard Gaussian variates and fix these values.

(c) Generate response values also as standard Gaussian ($\sigma^2 = 1$), independent of the predictors. Fit regression trees to the data of fixed size 1,5 and 10 terminal nodes and hence estimate the degrees of freedom of each fit. [Do ten simulations of the response and average the results, to get a good estimate of degrees of freedom.]

In [17]:

```python
import numpy as np
from sklearn.tree import DecisionTreeRegressor
import matplotlib.pyplot as plt
from sklearn.tree import plot_tree
import pandas as pd
```

In [45]:

```python
import numpy as np
from sklearn.tree import DecisionTreeRegressor

np.random.seed(123)  # for reproducibility

# Generate predictor variables
X = np.random.normal(size=(100, 10))

# Define function to generate response values
def generate_response(n):
    return np.random.normal(size=n)

# Define function to fit regression tree and estimate degrees of freedom
def fit_and_estimate_df(m):
    df = np.zeros(10)
    for i in range(10):
        y = generate_response(100)
        if m != 1:
            tree = DecisionTreeRegressor(max_leaf_nodes=m)
            tree.fit(X, y)
            y_hat = tree.predict(X)
            #print(y_hat)
        else:
            y_hat = X.mean(axis=1)
            #print(y_hat)

        df[i] = np.sum(np.cov(y, y_hat) / np.var(X))
    return np.mean(df)

# Fit regression trees with 2, 5, and 10 terminal nodes and estimate degrees
df_1= fit_and_estimate_df(1)
df_5 = fit_and_estimate_df(5)
df_10 = fit_and_estimate_df(10)

# Print estimated degrees of freedom for each tree size
print("Degrees of freedom for 1 terminal nodes:", df_1)
print("Degrees of freedom for 5 terminal nodes:", df_5)
print("Degrees of freedom for 10 terminal nodes:", df_10)
```

```
Degrees of freedom for 1 terminal nodes: 0.9923880977088825
Degrees of freedom for 5 terminal nodes: 1.9398998680614525
Degrees of freedom for 10 terminal nodes: 2.5879486307482087
```

In [ ]: