Assignment 8

(Clustering and Feature Selection)

Due: 16, June

1. [20pts] Short answers

(1) [5pts] Could cross validation prevent overfitting? State your reason.

(2) [5pts] What are the potential purposes of applying LASSO in regression problems?

(3) [5pts] If you are training a linear regression model, and it behaves well on training data but poorly on test data, what will happen to the bias and variance of your model when you decide to use ridge regression? State your reason using knowledge about bias-variance tradeoff.

(4) [5pts] Bagging with decision tree as base learner, and Random Forest, which one has the lower computational cost? And state the reason. (Assume they are trained the same number of iterations)

(1) No, cross validation alone cannot prevent overfitting. Cross validation is a technique used to estimate the performance of a model on unseen data by splitting the available data into multiple subsets for training and testing. It helps in assessing the generalization ability of a model, but it does not directly address the issue of overfitting. To prevent overfitting, additional techniques such as regularization or feature selection should be applied.

(2) The potential purposes of applying LASSO (Least Absolute Shrinkage and Selection Operator) in regression problems include feature selection and regularization. LASSO performs both variable selection and regularization by adding a penalty term to the ordinary least squares objective function. This penalty term encourages sparsity in the model coefficients, effectively selecting a subset of important features and shrinking the coefficients of less important features towards zero.

(3) When a linear regression model behaves well on training data but poorly on test data, it indicates that the model is likely overfitting the training data. By using ridge regression, which adds a regularization term to the ordinary least squares objective function, the bias of the model may increase. The regularization term in ridge regression shrinks the

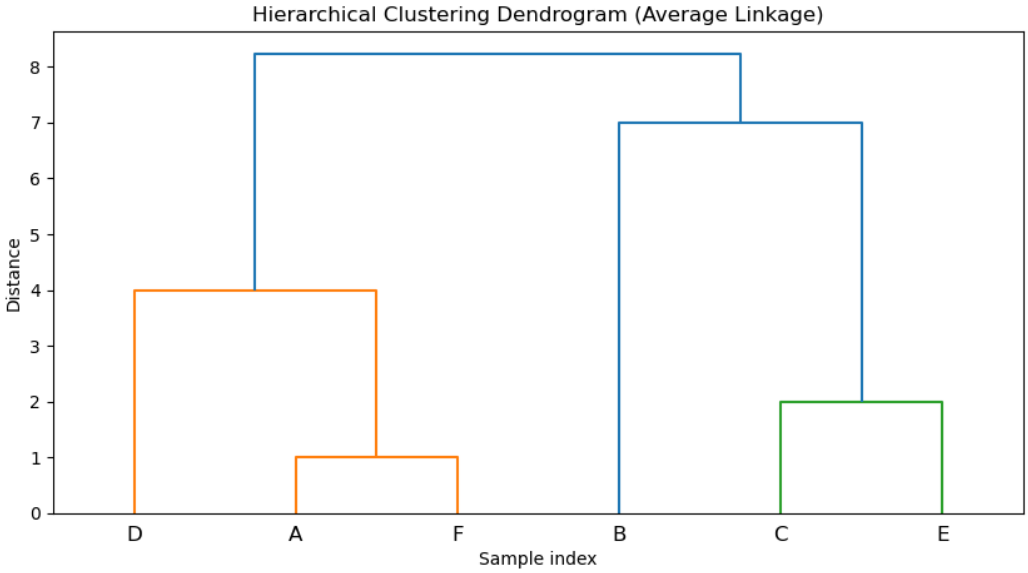
coefficients, reducing the model's flexibility and bringing down the complexity. This increased bias helps to address overfitting and improve the model's performance on unseen test data. The variance of the model may decrease as well, as ridge regression can reduce the sensitivity of the model to the specific training data.

(4) Bagging with decision tree as the base learner has a lower computational cost compared to Random Forest. Bagging involves training multiple decision trees independently on different subsets of the training data, and then averaging their predictions. Each decision tree is constructed without considering the others, which makes the process computationally efficient. On the other hand, Random Forest also uses bagging, but with an additional random feature selection process at each split. This feature selection adds computational complexity to the training process of Random Forest, making it relatively more computationally expensive compared to bagging with decision trees alone.

2. [25pts] Hierarchical Clustering

Apply the Hierarchical Clustering to the following distance matrix with average linkage. Write down each step of your clustering procedure. And give the dendrogram (树状图, 只给出层次结构即可，树状图中无需显示簇之间的距离).

	A	B	C	D	E	F
A	0					
B	12	0				
C	6	8	0			
D	2	7	9	0		
E	3	6	2	7	0	
F	1	8	20	6	2	0


Hierarchical Clustering Dendrogram (Average Linkage)
Distance
0
1
2
3
4
5
6
7
8
D
A
F
B
C
E
Sample index