Machine Learning Assignment 5

(Ensemble, Clustering & Feature Selection)

Due: June 12
ID:
Name:

1. [20pts] Bagging, Random Forest & Feature selection, Short answers

(1) Bagging with decision trees as base learners, and Random Forest, which one has the lower computational cost? And state the reason. (Assume they are trained the same number of iterations)

Solution.
Random Forest has a lower computational cost compared to Bagging with decision trees as base learners because it uses a random subset of features for splitting nodes, reducing the complexity of each tree. This leads to faster training and less overfitting, while Bagging uses all features.

(2) How do bagging and Random Forest introduce randomness to generate diverse individual learners?

Solution.
Bagging introduces randomness by creating multiple subsets of the training data through random sampling with replacement. Random Forest further introduces randomness by selecting a random subset of features for each split in the decision trees, ensuring diverse individual learners.

(3) From the perspective of bias-variance decomposition, what does bagging and Random Forest reduce respectively?

Solution.
Bagging reduces variance by averaging predictions from multiple models trained on different samples. Random Forest reduces both bias and variance by averaging multiple decision trees trained on varied features and samples.

(4) What are the potential purposes of applying LASSO in regression problems?

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.

(5) What's the difference between using L1 norm and L2 norm as regularization term in liner regression model?

In linear regression, using L1 norm (LASSO) leads to sparsity by potentially reducing some coefficients to zero, aiding in feature selection. Conversely, L2 norm (Ridge Regression) shrinks coefficients towards zero but does not set any to zero, which is beneficial for handling multicollinearity without eliminating variables.

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 draw the dendrogram.

	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

Solution.

Step 1 Initially, each element is its own cluster:

Cluster 0: 0 Cluster 1: 1 Cluster 2: 2 Cluster 3: 3 Cluster 4: 4 Cluster 5: 5

Step 2: Find Closest Clusters

Look for the smallest distance in the matrix that isn't zero. In this matrix, the smallest distance is 1 between clusters 0 and 5.

Step 3: Merge Clusters

Merge clusters 0 and 5 into a new cluster (0,5). Now update the cluster list:

Cluster (0,5) Cluster 1: 1 Cluster 2: 2 Cluster 3: 3 Cluster 4: 4

Step 4: We need to update the distances from the new cluster (0,5) to all other clusters using the average linkage method. Here's how we calculate it:

Distance to Cluster 1 = (Distance(0,1) + Distance(5,1))/2 = (12 + 8)/2 = 10

Distance to Cluster 2 = (Distance(0,2) + Distance(5,2))/2 = (6 + 20)/2 = 13

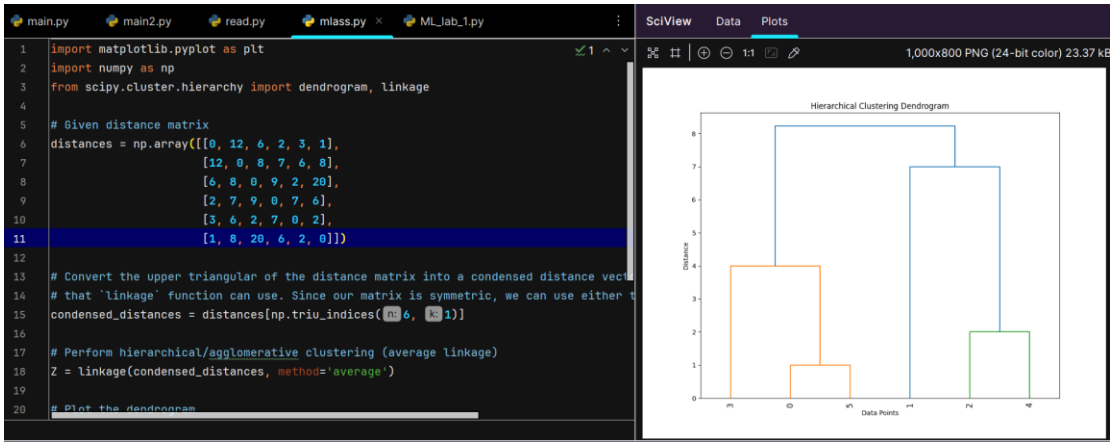
Distance to Cluster 3 = (Distance(0,3) + Distance(5,3))/2 = (2 + 6)/2 = 4

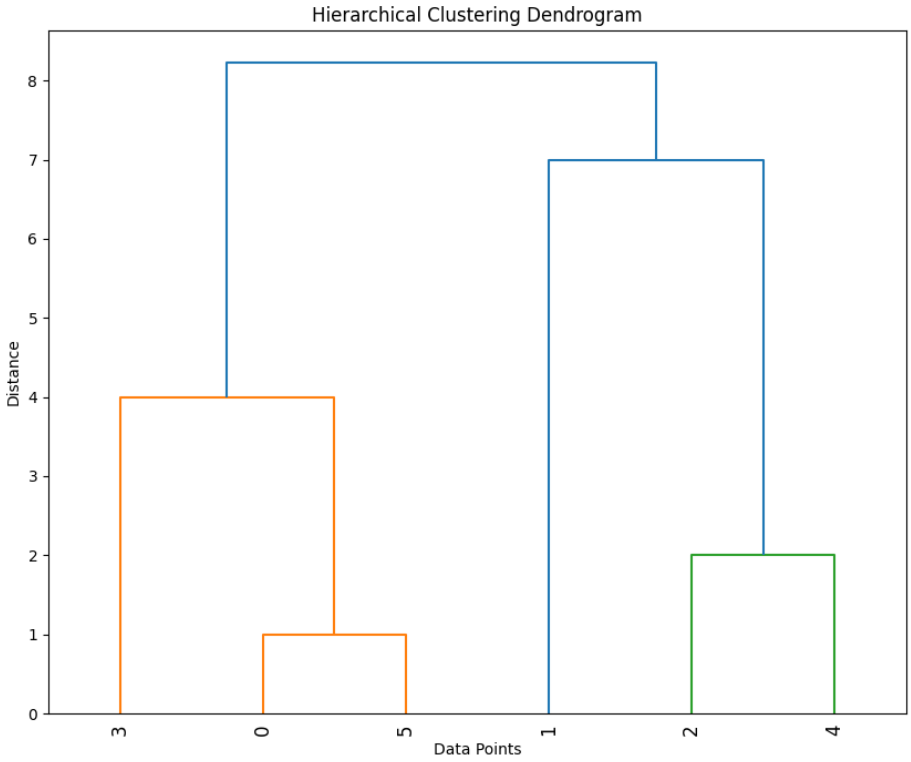
Distance to Cluster 4 = (Distance(0,4) + Distance(5,4))/2 = (3 + 2)/2 = 2.5

Step 5: Repeat steps 2-4 until all elements are in one cluster.

Drawing the Dendrogram

Using python to draw the dendrogram





3. [25pts] Logistic Regression and Regularization

Logistic regression 的优化问题为：

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \left[\, y_i \boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i - \log\left(1 + e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}\right) \right]$$

对目标函数加上 L2 正则化项以减小过拟合风险，则目标函数变为：

$$\min_{\boldsymbol{\beta}} \sum_{i=1}^{m} \left[\, y_i \boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i - \log\left(1 + e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}\right) \right] + \lambda \|\boldsymbol{\beta}\|_2^2$$

使用梯度下降法推导新目标函数的梯度，并写出参数更新公式。

Solution.

目标函数为：

$$J(\beta) = \sum_{i=1}^{m} \left[y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right] + \lambda \left|\left|\beta\right|\right|_2^2$$

对目标函数求β的偏导：

$$\frac{\partial J(\beta)}{\partial \beta}=\sum_{i=1}^{m}(y_i x_i-\frac{x_i e^{\beta^T x_i}}{1+e^{\beta^T x_i}})+2\lambda\beta=\sum_{i=1}^{m}(y_i x_i-x_i\sigma(\beta^T x_i))+2\lambda\beta$$

梯度下降法的参数更新公式：

$$\beta\leftarrow\beta-\alpha\frac{\partial J(\beta)}{\partial \beta}$$

将梯度带入，得到更新公式：

$$\beta\leftarrow\beta-\alpha(\sum_{i=1}^{m}\big(y_i x_i-x_i\sigma(\beta^T x_i)\big)+2\lambda\beta)$$