Machine Learning

58122204 Xie xing *Instructor: Xu-Ying Liu*

SEU — 2024 Spring

Assignment 5

§ Ensemble, Clustering & Feature Selection §

Problem 1: Bagging, Random Forest & Feature Selection, Short Answer

(1) (1) Computational Cost Random Forest has a lower computational cost compared to Bagging with decision trees as base learners. This is because Random Forests use feature bagging, selecting a random subset of features for each tree, leading to smaller and less complex trees.

(2) (2) Introducing Randomness

- **Bagging:** Introduces randomness by bootstrapping, creating multiple subsets of the training data through sampling with replacement. Each subset is used to train a different base learner.
- **Random Forest:** In addition to bootstrapping, it randomly selects a subset of features at each split in the decision tree, ensuring more diversity among the trees.

(3) (3) Bias-Variance Decomposition

- **Bagging:** Primarily reduces variance by averaging the predictions of multiple base learners trained on different subsets of the data.
- **Random Forest:** Reduces variance similarly to Bagging but also reduces correlation between the trees by introducing feature randomness, enhancing variance reduction.

(4) (4) Purposes of LASSO in Regression

- **Feature Selection:** LASSO shrinks some coefficients to exactly zero, selecting a simpler model with fewer features.
- **Regularization:** LASSO adds an L1 norm penalty to the regression loss function, helping to prevent overfitting.

(5) (5) L1 Norm vs. L2 Norm

- **L1 Norm (LASSO):** Encourages sparsity in the model by shrinking some coefficients to zero, effectively performing feature selection.
- **L2 Norm (Ridge):** Distributes the regularization effect more evenly across all coefficients, typically leading to smaller, non-zero coefficients without feature selection.

Problem 2: Hierarchical Clustering

(1) Steps for Hierarchical Clustering

1. **Initialization:**
    - Start with each data point as its own cluster.

- The given distance matrix is:

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

2. **Step 1: Find the Closest Clusters**
   - Clusters A and F are closest with a distance of 1.
3. **Step 2: Merge Clusters A and F**
   - Form a new cluster AF.
   - Update the distance matrix using average linkage:

$$d(AF, X) = \frac{d(A, X) + d(F, X)}{2}$$

   - New distances:

$$d(AF, B) = \frac{12+8}{2} = 10, \quad d(AF, C) = \frac{6+20}{2} = 13$$
$$d(AF, D) = \frac{2+6}{2} = 4, \quad d(AF, E) = \frac{3+2}{2} = 2.5$$

4. **Step 3: Find the Closest Clusters**
   - Clusters AF and E are closest with a distance of 2.5.
5. **Step 4: Merge Clusters AF and E**
   - Form a new cluster AFE.
   - Update the distance matrix:

$$d(AFE, B) = \frac{10+6}{2} = 8, \quad d(AFE, C) = \frac{13+2}{2} = 7.5$$
$$d(AFE, D) = \frac{4+7}{2} = 5.5$$

6. **Step 5: Find the Closest Clusters**
   - Clusters AFE and D are closest with a distance of 5.5.
7. **Step 6: Merge Clusters AFE and D**
   - Form a new cluster $AFED$.
   - Update the distance matrix:

$$d(AFED, B) = \frac{8+7}{2} = 7.5, \quad d(AFED, C) = \frac{7.5+9}{2} = 8.25$$

8. **Step 7: Find the Closest Clusters**
   - Clusters $AFED$ and B are closest with a distance of 7.5.
9. **Step 8: Merge Clusters $AFED$ and B**
   - Form a new cluster $AFEDB$.
   - Update the distance matrix:

$$d(AFEDB, C) = \frac{8.25+8}{2} = 8.125$$

10. **Step 9: Merge the Remaining Clusters**
    - Merge clusters $AFEDB$ and C with a distance of 8.125.

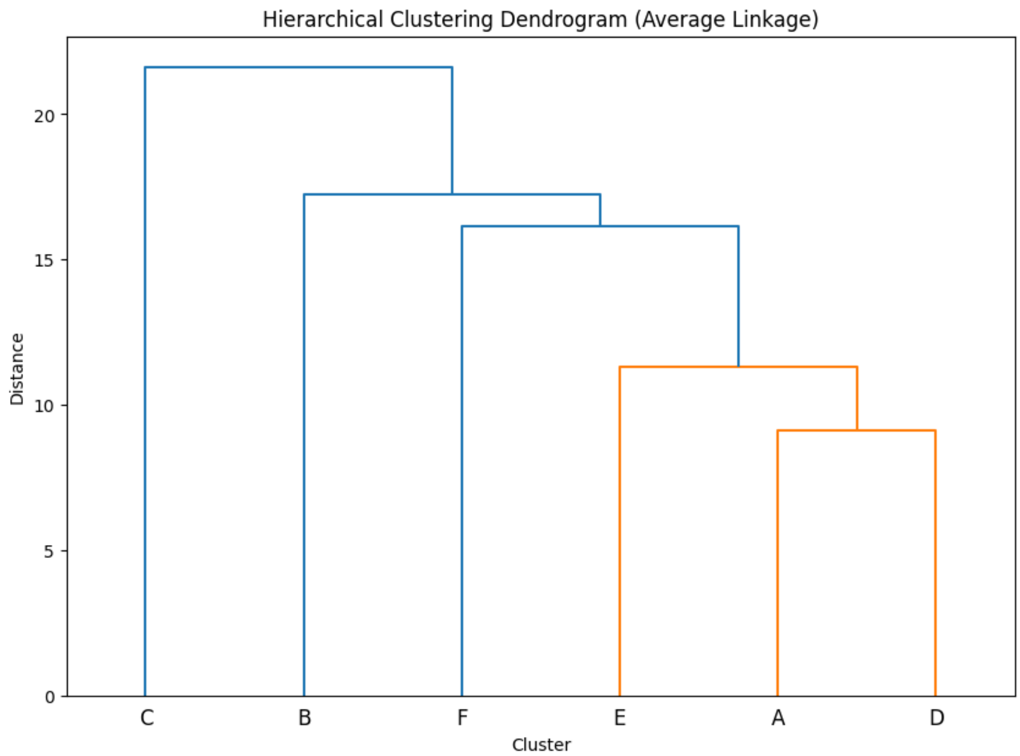

Figure 1: Hierarchical Clustering Dendrogram (Average Linkage)

(2) Dendrogram The resulting dendrogram is shown in the Fig. 1.

Problem 3: Logistic Regression and Regularization

(1) Optimization Problem The logistic regression optimization problem with L2 regularization is given by:

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

(2) Gradient Derivation The gradient of the objective function with respect to $\boldsymbol{\beta}$ is:

$$\nabla_{\boldsymbol{\beta}} = \sum_{i=1}^{m} \left(y_i \hat{\boldsymbol{x}}_i - \frac{e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}}{1 + e^{\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i}} \hat{\boldsymbol{x}}_i \right) + 2\lambda \boldsymbol{\beta}$$

Simplifying the expression inside the sum:

$$\nabla_{\boldsymbol{\beta}} = \sum_{i=1}^{m} \left(y_i - \sigma(\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i) \right) \hat{\boldsymbol{x}}_i + 2\lambda \boldsymbol{\beta}$$

where $\sigma(z) = \frac{1}{1+e^{-z}}$ is the sigmoid function.

(3) Parameter Update Formula Using gradient descent, the parameter update rule is:

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha \nabla_{\boldsymbol{\beta}}$$

where α is the learning rate. Substituting the gradient:

$$\boldsymbol{\beta} \leftarrow \boldsymbol{\beta} - \alpha \left(\sum_{i=1}^{m} \left(y_i - \sigma(\boldsymbol{\beta}^T \hat{\boldsymbol{x}}_i) \right) \hat{\boldsymbol{x}}_i + 2\lambda \boldsymbol{\beta} \right)$$