**Analysis of Arcene Cancer Classification Data**

Group 37: Gaoli Lin, Haowei Yan, Haotian Liu, Yizhou Gu

1. **Instruction:**

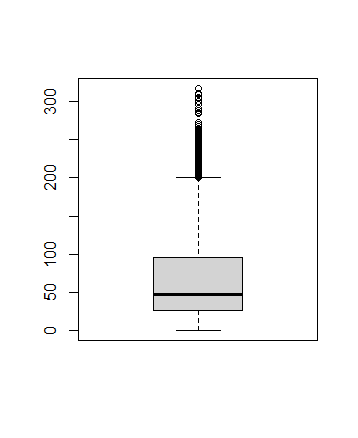
Modern mass spectrometry collects thousands of molecular features, but analyzing such high-dimensional data is challenging. Identifying key patterns can aid early diagnosis and improve treatment.

To explore the question, the Arcene dataset will be divided into a training set and a test set. A unique training set containing 100 samples with 5,000 randomly selected features. A fixed test set (100 samples with all 10,000 features) will be used for model evaluation. The study will employ various classification techniques to assess their accuracy in distinguishing between cancerous and normal tissue samples. Through the model analysis, to explore the following research questions:

* **Primary research question**: Can biochemical features accurately distinguish between cancerous and normal tissue samples?
* **Secondary research question**: Compare the results of different classification models to find out a best classification model.

1. **Exploratory Analysis:**

* **Initial rejection of potential probes**

Based on the dataset description, we performed an initial feature selection process to reduce the dimensionality of the data and eliminate variables that may not provide meaningful insights for the model. We focused on removing features with extremely low variance, as these are typically considered to be probes or noise sequences that do not contribute valuable information for predictive modeling. Such features often show minimal variability across observations and do not offer distinguishing power for classification or regression tasks. By removing these low-variance features, we reduced the data dimensionality, which facilitates more efficient analysis and model training, and helps to prevent overfitting by removing redundant or irrelevant information. This step is essential for improving the overall performance and interpretability of the model.

* **Data Dimension Reduction Processing**

PCA was further applied to reduce the dimensionality of the data. The Kaiser criterion and the cumulative variance contribution ratio were used to select the effective principal components. Based on these criteria, the original dataset was updated, leading to a significant reduction in its dimensionality.

1. **Method**
   1. SVM

Support Vector Machine (SVM) is a classification algorithm that finds the optimal boundary between different classes by maximizing the margin between data points. It identifies key data points, known as support vectors, that are closest to the decision boundary and uses them to define the classification rule. If the data is not linearly separable, SVM applies the kernel trick to transform it into a higher-dimensional space where separation is possible. One of the main reasons for using SVM is that it does not assume specific data distribution or variance requirements, unlike methods such as Discriminant Analysis (DA). SVM is effective for high-dimensional datasets and works well with small to medium-sized data. However, it can be computationally expensive for very large datasets and requires careful tuning of hyperparameters like C, gamma, and kernel type to achieve optimal performance.

The datasets contain features derived from different preprocessing methods:

* + Dataset 1: Original dataset with 5000 features.
  + Dataset 2: Feature selection reduced the number of features to 1824.
  + Dataset 3: Principal Component Analysis (PCA) reduced the feature count to 47.

For each dataset, SVM models were trained, tested, and optimized using GridSearchCV to identify the best hyperparameters. The datasets were loaded and divided into features (X) and target labels (y). Standardization was applied using StandardScaler to ensure the SVM operates effectively. An initial SVM model with default parameters was trained on each dataset. Model performance was evaluated using accuracy and Area Under the Curve (AUC) scores. Hyperparameter tuning was performed using GridSearchCV to optimize SVM parameters (C, gamma, and kernel). The best model from the grid search was selected and evaluated on the test set. The following hyperparameters were explored:

C: [0.01, 0.1, 0.5, 1, 2, 10, 100], gamma: ["scale", "auto", 0.01, 0.1, 1], kernel: ["linear", "rbf", "poly"]. The best hyperparameters were selected based on AUC scores.

1. **Results and Conclusions**

图表, 折线图

AI 生成的内容可能不正确。图表, 折线图

AI 生成的内容可能不正确。图表, 折线图

AI 生成的内容可能不正确。4.3 SVM

|  |  |  |  |
| --- | --- | --- | --- |
| Dataset | Features | Accuracy | AUC |
| Original | 5000 | 0.84 | 0.930 |
| Feature Selection | 1824 | 0.88 | 0.948 |
| PCA Reduction | 47 | 0.81 | 0.924 |

Reducing the number of features from 5000 to 1824 improved accuracy and AUC after tuning. PCA-reduced data (47 features) maintained good accuracy but performed slightly worse than feature selection. In most cases, GridSearchCV improved AUC over the default model. For Dataset 1, a linear kernel was chosen as optimal, while for Datasets 2 and 3, an RBF kernel was preferred.

Feature selection (1800 features) provided the best performance in terms of both accuracy and AUC. PCA (40 features) led to minor accuracy reduction but is still a viable option for dimensionality reduction.