**Cancer Classification Using PCA, Feature Selection, and Robustness Analysis to Label Noise**

**Part 1: Model Development with PCA and Feature Selection**

**1. Data Loading and Initial Exploration**

* Loaded gene expression data and corresponding class labels.
* Verified no missing values in features or labels.

**2. Dimensionality Reduction with PCA**

* Scaled training data using StandardScaler.
* Explored number of principal components (PCs) by examining explained variance ratio and scree plot elbow method.
* Selected 10 to 20 PCs as candidates for dimensionality reduction.
* Implemented PCA within pipelines combined with classifiers for cross-validation.

**3. Classification Models with PCA**

* Tested three classifiers within pipelines with PCA:
  + K-Nearest Neighbors (KNN)
  + Logistic Regression (multi-class, multinomial)
  + Quadratic Discriminant Analysis (QDA)
* **Results:**
  + Logistic Regression generally performed best.
  + KNN was flexible with minimal assumptions.
  + QDA modeled class-specific covariances but showed some warning of collinearity.
  + PCA reduced dimensions effectively while preserving high accuracy.

**4. Logistic Regression Without PCA**

* Also trained logistic regression on raw scaled data.
* Observed slightly better performance than with PCA but difference was minimal.
* PCA remains beneficial for larger datasets to reduce computation and handle spurious correlations.

**5. Feature Selection Approaches**

* Explored feature selection methods applied to logistic regression without PCA:
  + Variance Thresholding: Removed low variance and constant features.
  + ANOVA F-score Selection: Selected features most correlated with target labels.
* **Findings:**
  + Removing only constant features (variance threshold=0) gave best results among variance threshold experiments.
  + Selecting top features by F-score (especially ~80%) improved classification performance slightly beyond raw logistic regression.
  + Even selecting as low as 5% of features by F-score sustained high accuracy, simplifying the model and reducing dimensionality.

**6. Summary of Part 1**

* Logistic regression with F-score based feature selection achieved best overall results.
* PCA and variance thresholding helped reduce data complexity with minor impact on accuracy.
* Models showed excellent precision, recall, and f1-score across multiple cancer types, with slightly lower performance on low-sample classes.

**Part 2: Robustness Analysis to Label Misclassification Noise**

**1. Introduction**

* Investigated effects of label noise by randomly mislabeling 5%, 30%, and 70% of training labels across six cancer classes.
* Tested on three training set sizes: 80%, 65%, and 50% to observe effect of training data quantity.

**2. Methodology**

* Created mislabeled training datasets at three noise levels (5%, 30%, 70%).
* Trained classifiers with and without PCA, as well as feature selection.
* Evaluated on clean test data using balanced accuracy and recall metrics.

**3. Classifier Performance Under Mislabeling**

* **Logistic Regression without PCA:**
  + Performance degrades gracefully at 5% mislabels.
  + Severe degradation at 30% and especially 70% mislabels.
* **Logistic Regression with PCA:**
  + Performance dropped sharply even at 5% mislabels.
  + Poor resistance to label noise compared to no PCA.
* **QDA with PCA:**
  + Similar drop in performance with increasing label noise.
  + More sensitive than logistic regression.
* **KNN with PCA:**
  + Also shows considerable performance decay with mislabels.
  + Best hyperparameters varied with noise level.

**4. Feature Selection vs No Feature Selection Under Mislabeling**

* Logistic regression combined with feature selection (variance thresholding and F-score) showed better robustness to label noise.
* Feature selection retained more consistent recall scores across varying noise levels and training sizes.
* Visualized using plots showing average recall decay over mislabeling percentages.

**5. Visualizations**

* Plots illustrating average recall vs mislabeling for different classifiers and training sizes.
* Comparisons of logistic regression with and without feature selection methods under noisy labels.
* Heatmaps highlighting recall and support across classes for selected models.

**6. Conclusions from Part 2**

* Label noise substantially harms classifier performance.
* Feature selection methods improve robustness to mislabels.
* Larger training sizes help maintain accuracy despite noise.
* PCA can reduce performance robustness to label noise in this domain.

**Overall Project Conclusion**

This project comprehensively investigated cancer subtype classification from gene expression data through dimensionality reduction, feature selection, and analysis of noise robustness:

* PCA and feature selection effectively reduce dimensionality while retaining classification power.
* Logistic regression with feature selection (particularly F-score based) achieves the best predictive accuracy.
* Classifiers’ resilience to label noise varies significantly; feature selection is crucial for robust performance.
* Visual and quantitative analyses reveal trade-offs in training size, noise level, and model complexity.

These findings can guide practical model development for biological datasets prone to noise and high dimensionality, supporting accurate and robust cancer classification.

If you need, I can help prepare this full project in a formatted report or presentation form.

1. <https://ppl-ai-file-upload.s3.amazonaws.com/web/direct-files/attachments/31249425/64ea461a-f248-42ce-a63f-c6927068f49e/paste.txt>