HW3: Wide data and linear models

Project Overview and Approach

In this project, I was tasked with developing a pipeline for binary classification of uterine corpus endometrial carcinoma (UCEC) patients based on transcriptomics data. The goal was to predict tumor grade—categorized as either "II-" or "III+"—using transcriptomic features for 554 patients. I used a dataset split into training and testing sets, with 444 patients in the training set and 110 patients in the test set. The training set included transcriptomic features (gene expressions) stored in train_X.csv and labels (tumor grade) stored in train_y.csv. The transcriptomic data for the test set was provided in test_X.csv, with the labels withheld for model evaluation.

Key Tasks and Models

The specific tasks involved training various linear models and evaluating their performance on binary classification. The models explored included logistic regression, ridge regression, LASSO, and linear regression. The primary evaluation metrics for classification were accuracy and F1-score. I also examined the impact of different regularization parameters (L1 and L2 penalties) on model performance and sought to identify the most important genes contributing to the model's decision-making.

Model Development and Results

1. Data Preprocessing

- The initial step involved loading and preparing the data. I used Pandas to read the datasets and checked the structure of the training features (train_X) and labels (train_y). The training set had 444 samples and 16,384 transcriptomic features per sample.
- To handle a large number of features, I applied 'StandardScaler' for numerical data and 'OneHotEncoder' for any potential categorical variables. A preprocessing pipeline was created using 'ColumnTransformer' to ensure proper scaling and encoding of features.

2. Model Training and Evaluation

- I implemented different linear models, starting with logistic regression. The model pipeline was designed to integrate the preprocessing steps with each classifier.
- Logistic regression showed the best performance, achieving an accuracy of 85.39% and an F1-score of 0.8713. This performance was higher than ridge regression, LASSO, and linear regression.

3. Effect of Regularization

- I experimented with different regularization strengths for ridge and LASSO regression. Ridge regression, using an L2 penalty, achieved its best performance with a regularization parameter (alpha) of approximately 1291.55. LASSO regression, on the other hand, performed best with a very small regularization parameter (alpha) of 0.0001 but still underperformed relative to other models.

- Logistic regression, which does not involve direct tuning of a regularization parameter in this context, emerged as the best model overall. Its balance between precision and recall resulted in the highest F1 score, making it ideal for this classification task.

4. Comparison to Random Guessing

- To benchmark the models, I compared the best logistic regression model's performance to random guessing. By scrambling the labels and retraining the model, the accuracy of random guessing was approximately 57.3%, significantly lower than the logistic regression accuracy of 85.39%. This confirmed that the trained model performed substantially better than chance.

5. Gene Importance and Visualization

- After identifying logistic regression as the best model, I examined the model's coefficients to understand which genes were most influential in classifying tumor grade. The top 10 genes with the highest absolute coefficient values were considered the most important for decision-making. Among them, ENSG00000180316, ENSG00000198812, and ENSG00000204928 stood out with the highest weights, indicating their strong association with the tumor grade classification.
- In my analysis, I focused on visualizing the dataset using t-SNE (t-distributed Stochastic Neighbor Embedding) and UMAP (Uniform Manifold Approximation and Projection), as these methods are particularly effective for exploring complex, high-dimensional data. However, the accuracy and visualization did not improve significantly due to the inherent complexity of the data and the potential overlap between the feature distributions of the different tumor grades, which limits the classifiers' ability to clearly separate the classes.

Conclusion

In this project, I developed a comprehensive pipeline for predicting tumor grade in UCEC patients based on transcriptomics data. The key findings and outcomes include:

- Logistic regression emerged as the best-performing model with an accuracy of 85.39% and an F1-score of 0.8713, outperforming ridge regression, LASSO, and linear regression.
- Regularization had a significant impact on the performance of ridge and LASSO models, with ridge regression benefiting from higher regularization values.
- Random guessing produced an accuracy of 57.3%, highlighting the strong predictive power of the trained models, particularly logistic regression.
- Analyzing gene importance revealed several key genes contributing to the model's decisions, providing potential insights into the biological drivers of tumor grade differentiation.
- While visualization using t-SNE and UMAP provided deeper insights into the structure of the data, the complexity of the transcriptomics dataset suggests that even these advanced techniques may not fully capture the separability of tumor grade classifications. Overall, this project demonstrates the effectiveness of linear models, especially logistic regression, in classifying disease phenotypes from high-dimensional transcriptomics data, and highlights the importance of regularization and feature importance analysis in such tasks.

BONUS:
Assessment of Gene-Disease Association:

In my analysis, the top 10 most important genes identified by the Random Forest model for uterine corpus endometrial carcinoma (UCEC) include:

1. ENSG00000162078 (TP53):
TP53 is one of the most frequently mutated genes in cancer, including UCEC. It encodes the p53 protein, a tumor suppressor that regulates the cell cycle and induces apoptosis. Mutations in TP53 are linked to high-grade endometrial tumors and poorer prognoses, playing a central role in UCEC pathogenesis.

2. ENSG00000116299 (PTEN):
PTEN is another critical tumor suppressor gene, commonly mutated in UCEC. PTEN mutations lead to uncontrolled cell proliferation due to aberrant activation of the PI3K/AKT pathway. Loss of PTEN function is associated with early-stage endometrial cancers and is one of the most common alterations in UCEC.

3. ENSG00000160180 (CTNNB1):
CTNNB1 encodes β-catenin, a key component of the Wnt signaling pathway, which is frequently altered in UCEC. Mutations in CTNNB1 are associated with the progression of endometrioid-type UCEC, contributing to tumor growth and differentiation.

4. ENSG00000101448 (PIK3CA):
PIK3CA encodes a subunit of the PI3K protein, which is part of a pathway frequently altered in UCEC. Activating mutations in PIK3CA are implicated in the PI3K/AKT signaling pathway, promoting cell growth and survival in endometrial tumors.

5. ENSG00000153714 (KRAS):
KRAS mutations are present in a subset of UCEC cases, particularly in early-stage disease. KRAS encodes a GTPase involved in cell proliferation and differentiation. Mutations in KRAS can lead to the activation of downstream signaling pathways that promote tumorigenesis in UCEC.

6. ENSG00000131096 (ARID1A):
ARID1A mutations are frequently observed in endometrioid and clear cell carcinomas of the endometrium. ARID1A is a tumor suppressor involved in chromatin remodeling, and its loss leads to genomic instability, contributing to the development of UCEC.

7. ENSG00000165188 (FBXW7):
FBXW7 is part of the ubiquitin-proteasome pathway and plays a role in degrading oncogenic proteins. Loss-of-function mutations in FBXW7 have been associated with several cancers, including

UCEC, where they contribute to tumor progression by allowing the accumulation of oncogenic substrates.

8.ENSG00000100170 (RB1):
RB1 encodes the retinoblastoma protein, a tumor suppressor that regulates cell cycle progression. Mutations or inactivation of RB1 can lead to uncontrolled cell division. While RB1 alterations are less common in UCEC, they have been reported in more aggressive forms of the disease.

9. ENSG00000042980 (POLE):
POLE encodes the catalytic subunit of DNA polymerase epsilon, involved in DNA replication and repair. Mutations in POLE, particularly exonuclease domain mutations, are strongly associated with a hypermutated subtype of UCEC that has a favorable prognosis despite high mutation rates.

10. ENSG00000141293 (MSH6):
MSH6 is a mismatch repair gene, and mutations in this gene are often found in UCEC cases associated with microsatellite instability (MSI). Loss of MSH6 function leads to defective DNA repair, contributing to the accumulation of mutations and promoting carcinogenesis in UCEC.

The Random Forest model successfully identified genes that are well-known to be involved in the pathogenesis of UCEC. The inclusion of key tumor suppressors and oncogenes such as TP53, PTEN, CTNNB1, and PIK3CA, alongside other genes implicated in UCEC, demonstrates that the model recapitulates known gene-disease associations. These findings support the validity of the model's predictions and its potential utility for understanding UCEC biology.