

BIOS226 - Topic 5 - Supervised Learning (Part 2)

Student Companion Guide

The Tumour Subtype Classification Pipeline

This document is a **companion guide to Part 2 of the Supervised Learning lecture**.

Use it alongside the lecture slides and code walkthrough to reinforce the main ideas and decisions in the modelling pipeline.

From Data to Decision: The Pipeline

Here, we focus on applying supervised learning to a synthetic breast cancer gene-expression dataset. We will discuss the origin of this dataset in this week's workshops.

- 120 patients
- 10,000 gene features (**Gene_1** to **Gene_10000**)
- Two tumour subtypes: **Luminal_A** and **Basal_like**
- Logistic regression classifier

The core objective is to predict tumour subtype from gene-expression patterns while avoiding leakage and overfitting. We will discuss leakage/overfitting in more detail in Part 2.

1. Define the Prediction Problem

This is a binary classification task.

- **Input (X):** gene-expression measurements
- **Output (Y):** probability that a tumour is **Basal_like**

In this pipeline, **Basal_like** is treated as the positive class.

2. Validate the Data Structure

Before any model training, check:

- Required columns exist (**Patient_ID**, **Subtype**, **Gene_***)
- Exactly two classes are present
- Gene columns are numeric
- No schema inconsistencies are present

Why this matters: model outputs are only meaningful when the input structure is valid.

3. Stratified Train/Test Split

Split the dataset into:

- 80% training set
- 20% held-out test set

Use **stratification** so class balance is preserved in both splits.

The test set is not used during training. It represents unseen patients and should only be used once for final evaluation.

4. Feature Selection (Train Only)

Rank genes by absolute difference in class-wise mean expression, then keep the top K genes (for example, 25).

Important rule: feature selection must be done on the training set only.

If test data influences feature selection, leakage occurs and performance estimates become optimistic.

5. Scaling & Normalisation (Train Only)

After selecting features, standardise each one:

- subtract training mean
- divide by training standard deviation

Then apply those same training-derived parameters to the test set.

Aside: RNA-seq Normalization vs ML Scaling

These are different steps.

- **RNA-seq normalisation** adjusts for sequencing depth/composition (for example CPM, TPM, DE-Seq2 size factors, TMM).
- **Machine-learning scaling** standardises feature magnitude for stable model fitting.

For this synthetic dataset, values behave like already-normalised expression values, so the teaching focus is on ML scaling and leakage-safe evaluation.

6. Cross-Validation on Training Data

Use k-fold cross-validation on the training split only.

Example (10-fold):

1. Split training data into 10 folds
2. Train on 9 folds
3. Validate on 1 fold
4. Repeat until each fold has been the validation fold once

Track metrics per fold (for example AUC and precision) to estimate stability before touching the test set.

7. Fit the Logistic Regression Model

Logistic regression models log-odds:

$$\log(p / (1 - p)) = \text{beta0} + \text{beta1}x_1 + \text{beta2}x_2 + \dots$$

The model output is a probability between 0 and 1 for each sample.

8. Confusion Matrix: Understanding Errors

Choose a threshold (for example 0.85) to convert probabilities to class predictions.

Confusion matrix components:

- True Positives (TP)
- False Positives (FP)
- True Negatives (TN)
- False Negatives (FN)

Derived metrics:

- **Precision** = $TP / (TP + FP)$
- **Sensitivity (Recall)** = $TP / (TP + FN)$
- **Specificity** = $TN / (TN + FP)$

These describe performance at one specific operating point.

9. ROC Curve & AUC

The ROC curve evaluates performance across all thresholds.

- X-axis: False Positive Rate
- Y-axis: True Positive Rate (Sensitivity)

The AUC summarises ranking quality:

- 0.5: random ranking
- 1.0: perfect ranking

AUC asks whether positive cases are generally scored above negative cases.

Aside: What Is an ROC Curve?

If you are seeing ROC for the first time, this is the quick intuition:

- **ROC** stands for **Receiver Operating Characteristic**.
- The name comes from radar/signal-detection work in the 1940s, where “receiver operators” distinguished true signals from noise.
- In machine learning, the same idea is used to evaluate how well a model separates positive and negative classes.

How to interpret an ROC curve:

- Each point on the curve corresponds to one classification threshold.
- X-axis (False Positive Rate): how often negatives are incorrectly called positive.
- Y-axis (True Positive Rate): how often positives are correctly detected.
- Moving along the curve means changing the threshold from strict to lenient.

What the shape tells you about model success:

- A curve that bends strongly toward the **top-left corner** indicates good discrimination.
- A curve close to the **diagonal line** ($AUC \sim 0.5$) indicates near-random performance.
- A curve mostly **below the diagonal** suggests predictions may be systematically reversed (for example, labels or score direction flipped).

In short: the closer the curve is to top-left, the better the model is at ranking true positives above true negatives.

10. Final Test Evaluation and Discipline

After all model decisions are fixed, evaluate once on the held-out test set.

Report:

- confusion matrix
- precision
- ROC AUC
- MSE (probability-based)
- R-squared (probability-based)

This gives the most honest estimate of how the model may perform on new data.

Aside: Threshold Choice and Clinical Trade-Offs

A higher threshold (for example 0.85) usually:

- increases certainty when calling positive
- reduces false positives
- can miss more true positives

A lower threshold (for example 0.4 to 0.5) usually:

- increases sensitivity
- catches more true positives
- increases false positives

The right threshold depends on the clinical consequences of false negatives versus false positives.

Pipeline Summary

- Keep train and test roles strictly separate.
 - Perform feature selection and scaling inside the training workflow.
 - Use CV for model stability and test set for final confirmation.
 - Interpret confusion-matrix metrics and ROC/AUC together.
 - Treat threshold selection as a domain decision, not just a technical default.
-

Reference

Perou, C. M., Sorlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature*, 406, 747-752.