BIOS226 - Topic 5 - Supervised Learning (Part 1)

Foundations for Biological Model Evaluation

Dr. Robert Treharne

By the end of this topic, you should be able to:

- Explain the difference between exploratory and supervised learning in biological data
- Describe a leakage-safe supervised learning pipeline from raw data to final evaluation
- Interpret a confusion matrix and ROC curve in a clinical context
- Recognise common failure modes (overfitting, leakage, imbalance, label errors)
- Critically evaluate whether a model is genuinely generalisable or silently flawed

From Exploration to Prediction

- PCA and clustering show structure in biological data.
- Supervised learning moves us from pattern finding to outcome prediction.
- Practical impact: treatment and risk decisions can be data-driven.

Example: Oncotype DX

- Gene expression from tumour tissue is used to estimate recurrence risk.
- That estimate supports chemotherapy decision-making.

Questions

- What does Oncotype DX measure?
- Is this classification or regression?

Unsupervised vs Supervised

Unsupervised learning

- Finds structure without known labels (for example PCA).

Supervised learning

- Learns from labelled examples to predict outcomes.

Example: The Cancer Genome Atlas (TCGA) subtype work

- Molecular profiles improved subtype definitions beyond histology.

What Are X and Y?

- **X**: measurable input features
- **Y**: outcome to predict

Biological examples

- SNPs -> disease risk
- Gene expression -> cancer subtype
- Clinical + genomic data -> prognosis

Questions

- What is a SNP in practical terms?
- Why can many small SNP effects still become predictive?

Classification vs Regression

Classification

- Predicts a category (for example subtype A vs B).

Regression

- Predicts a continuous value (for example drug response level).

Example

- Drug-response models estimate the strength of expected treatment effect.

What Is a Model?

- A model is a function that maps **X -> Y**.
- It learns from training data, then predicts on new samples.

Common models

- Logistic regression
- Random forest
- Support vector machine

Clinical example

- Early-warning systems can estimate sepsis risk before diagnosis.

The High-Dimensional Problem (p >> n)

- In omics, features (genes) can far exceed patients.
- This can make models unstable and easy to overfit.
- Apparent high accuracy may not replicate in new datasets.

Example

- Early microarray studies often reported strong results that later failed replication.

Questions

- Why does p >> n increase overfitting risk?
- How can noise create false signatures in small cohorts?

Why Biology Is Harder

Biological datasets are often:

- noisy
- small
- batch-affected
- biologically heterogeneous

Key risk

- A model may learn technical artefacts (for example batch or depth) instead of biology.

Take-home

- Good validation design is as important as model choice.

Common Non-Biological Applications

Supervised learning is widely used outside biology, for example:

- Email spam filtering
- Fraud detection in banking and payments
- Credit-risk scoring for lending
- Product and content recommendation systems
- Demand forecasting and inventory planning
- Predictive maintenance for machines and vehicles
- Speech recognition and language translation
- Computer vision for quality control in manufacturing

Biological Applications Beyond Omics

Other supervised-learning scenarios in biology include:

- Medical image diagnosis (for example chest X-ray, retinal imaging, MRI)
- Digital pathology slide classification
- ECG-based arrhythmia detection
- ICU deterioration and sepsis risk prediction from vital signs
- Microscopy image classification (cell type, morphology, localisation)
- Crop disease detection from plant images
- Wildlife species identification from camera traps or acoustic recordings