BIOS226 -- Supervised Learning & Biological Model Evaluation

Part 1 -- Foundations

A Student Guide to Understanding Supervised Learning in Biology

---

Introduction

In earlier sessions, you explored biological datasets using techniques such as Principal Component Analysis (PCA). PCA helps us *visualise structure* in high-dimensional data. It tells us whether samples cluster or separate naturally.

But PCA does not answer questions like:

- Can we predict whether a tumour will return?
- Can we predict whether a patient will respond to a drug?
- Can we estimate someone's genetic risk of disease?

To answer those questions, we use **supervised learning**.

This document explains what supervised learning is, how it works in biology, and why it has transformed medicine and research.

---

1. From Exploration to Prediction

Unsupervised methods (like PCA) look for patterns **without knowing the answer in advance**.

Supervised learning is different.\
It learns from data where the correct outcome is already known.

For example:

- Scientists collect tumour samples from patients.
- They measure the activity (expression) of thousands of genes.
- They already know which patients experienced cancer recurrence.

A supervised model can learn from these labelled examples and then predict recurrence risk for a new patient.

Real-world impact: Breast cancer recurrence prediction

Gene expression tests such as Oncotype DX measure the activity of a small panel of genes and generate a recurrence score. That score helps doctors decide whether chemotherapy is necessary. This approach has influenced treatment decisions for hundreds of thousands of patients (Sparano et al., 2015).

Supervised learning moves us from description to decision-making.

---

2. Unsupervised vs Supervised Learning

Let's clarify the difference.

Unsupervised learning:

- Finds patterns without outcome labels.

- Example: PCA groups samples based on similarity.

Supervised learning:
- Uses known labels during training. -
- Learns relationships between features and outcomes.

Cancer subtype classification

Historically, cancers were classified by how they looked under a microscope (histology). However, cancers that look similar can behave very differently.

Large projects such as **The Cancer Genome Atlas** (TCGA) analysed thousands of tumours using gene expression and mutation data. Supervised models helped identify molecular subtypes that predict prognosis and therapy response (Hoadley et al., 2018).

This changed how some cancers are defined and treated.

---

3. Features (X) and Outcomes (Y)?

Supervised learning involves two components:

X = Input features\
Y = Outcome (label)

In biology, features might include:

- Gene expression levels (how active each gene is)
- SNPs (genetic variants)
- Clinical variables (age, blood pressure)

The outcome might be:

- Disease vs healthy
- Tumour subtype
- Survival time
- Drug response level

What is a SNP?

DNA is made of four chemical bases: A, T, C, and G.

A **Single Nucleotide Polymorphism (SNP)** is a position in DNA where individuals differ by one base.

For example:

At a specific position in the genome:

- Person A has A
- Person B has G

If having G slightly increases disease risk, that SNP becomes a useful predictive feature.

SNPs are usually encoded numerically:

- 0 = no copies of risk variant
- 1 = one copy

- 2 = two copies

Individually, most SNPs have tiny effects. But thousands combined can produce meaningful predictions.

Real-world impact: Polygenic Risk Scores

Polygenic Risk Scores combine thousands of SNPs to estimate a person's genetic risk of diseases such as coronary artery disease. Recent research shows that these scores can identify individuals at risk levels comparable to single-gene mutations (Khera et al., 2018; Natarajan et al., 2021).

This influences screening and preventative medicine.

---

4. Classification vs Regression

Supervised problems fall into two categories.

Classification: The outcome is categorical. Example: Tumour subtype A or B.

Regression: The outcome is continuous. Example: Predicted drug response level.

Drug response prediction

Researchers use regression models to predict how sensitive a cancer cell line is to a chemotherapy drug (Geeleher et al., 2014).

Survival analysis is more complex because not all patients experience the event during observation (this is called censoring).

The key point: The type of outcome determines how we measure model performance.

---

5. What Is a Model?

A model is a mathematical function that connects inputs (X) to outputs (Y).

One of the simplest classification models is **logistic regression**.

Logistic regression:

1. Combines features using weights.
2. Applies a sigmoid function.
3. Outputs a probability between 0 and 1.

Example:

Probability of disease = 0.82

If we choose a threshold of 0.5 → classify as disease.\
If we choose 0.8 → only high-risk cases are classified as disease.

Threshold choice changes sensitivity and specificity.

Real-world impact: Sepsis prediction

Hospitals use supervised machine learning systems trained on electronic health records to predict sepsis hours before clinical deterioration (Seymour et al., 2016).

Early prediction can save lives.

Important principle: Models estimate probability, not certainty.

---

6. The High-Dimensional Problem (p \>\> n)

In biology:

p = number of features (e.g., 20,000 genes)\
n = number of samples (e.g., 100 patients)

When p is much larger than n, models can easily overfit.

Overfitting occurs when a model memorises noise instead of learning true patterns.

In early gene expression studies, researchers reported near-perfect cancer classification accuracy. Later studies failed to replicate those results because the original models had overfit small datasets (Simon et al., 2003; Varma & Simon, 2006).

With enough variables, it is almost always possible to find some pattern that separates two small groups --- even if the pattern is random.

High accuracy does not guarantee scientific truth.

---

7. Why Biological Data Is Challenging

Biological data is noisy and complex.

Sources of variation include:

- Technical variation (machine differences)
- Batch effects (processing samples at different times or labs)
- Biological heterogeneity (natural differences between individuals)

A model might accidentally learn:

- Which sequencing machine was used
- Which laboratory processed the sample\
- Differences in RNA concentration

Instead of learning disease biology.

Batch effects are a major cause of non-reproducible genomic findings (Leek et al., 2010).

Supervised learning is powerful, but fragile.

Key Takeaways

- Supervised learning predicts outcomes from biological data.\
- It has transformed cancer diagnosis, genetic risk prediction, and hospital monitoring.\

- High-dimensional data increases overfitting risk.\
- Biological noise and batch effects can mislead models.\
- Prediction does not equal causation.

References

Geeleher, P., Cox, N., & Huang, R. (2014). Clinical drug response prediction using baseline gene expression levels and in vitro drug sensitivity data. *Genome Biology*, 15, R47.

Hoadley, K. A., Yau, C., Hinoue, T., et al. (2018). Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types. *Cell*, 173(2), 291--304.

Khera, A. V., et al. (2018). Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nature Genetics*, 50, 1219--1224.

Leek, J. T., et al. (2010). Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11, 733--739.

Natarajan, P., et al. (2021). Polygenic risk score identifies individuals at risk for coronary artery disease. *Nature Medicine*, 27, 103--111.

Seymour, C. W., et al. (2016). Assessment of clinical criteria for sepsis. *JAMA*, 315(8), 762--774.

Simon, R., Radmacher, M. D., Dobbin, K., & McShane, L. (2003). Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute*, 95(1), 14--18.

Varma, S., & Simon, R. (2006). Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7, 91.