

Biomedical Data Science 2026: Homework Assignment 2

Due: March 29th, 11:59pm

Choose to do either MCDB & MBB (non-programming) or CBB & CPSC & S&DS (programming) assignment, depending on your academic affiliation. No late submissions will be accepted. Submission should be done in Canvas.

1 MCDB & MBB (Non-Programming)

1.1 Supervised Learning in Genomics: Logistic Regression (25pt)

You are building a clinical classifier to predict whether a cancer patient will respond to a specific immunotherapy based on the normalized expression levels of two biomarker genes: Gene X_1 and Gene X_2 . You use a Logistic Regression model:

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + e^{-(\mathbf{w}^T \mathbf{x} + b)}}$$

where $y = 1$ indicates a "Responder" and $y = 0$ indicates a "Non-Responder". After training on a large dataset, your learned parameters are:

1. Weight for Gene 1 (w_1) = 1.5
2. Weight for Gene 2 (w_2) = -1.0
3. Bias (b) = -0.5

You receive RNA-seq data for three new patients:

1. Patient A: $X_1 = 2.0, X_2 = 1.0$
2. Patient B: $X_1 = 0.0, X_2 = -2.0$
3. Patient C: $X_1 = 1.0, X_2 = 2.0$

When providing the final probability values for Patient A, Patient B, and Patient C in 1.1.1, please enclose the final numerical answer in double curly braces, for example, $\{\{0.40\}\}$.

1.1.1 Prediction & Classification (4pt)

Calculate the predicted probability of response $P(y = 1|\mathbf{x})$ for all three patients. If your clinical decision threshold is 0.5, how would you classify each patient? State the predicted class for each patient. (You may leave intermediate exponential terms like $e^{-1.5}$ in your work, but provide the final probability to two decimal places).

1.1.2 Loss Function (4pt)

Suppose the clinical trial concludes, and the true outcomes are revealed: Patient A responded ($y = 1$), Patient B responded ($y = 1$), and Patient C did not respond ($y = 0$). Calculate the mean binary cross-entropy loss specifically for this batch of 3 patients using natural logarithms. Report the final mean loss to two decimal places.

1.1.3 Decision Boundary (4pt)

Derive the algebraic equation for the decision boundary in the 2D feature space (X_1 vs. X_2) corresponding to the 0.5 probability threshold. What is the slope of this line?

1.1.4 Regularization Strategy (6pt)

In a real-world scenario, you may use ($\sim 20,000$) gene-expression features. If you want the model to select a sparse panel (e.g., 10–20 genes) by driving many irrelevant coefficients to be exactly zero, should you use L1 (Lasso) or L2 (Ridge) regularization? Briefly justify your choice by referencing the form of the penalty term ($|\mathbf{w}|_1 = \sum_j |w_j|$, $|\mathbf{w}|_2^2 = \sum_j w_j^2$) and how it affects sparsity of the learned weights.

1.1.5 Cost-Sensitive Clinical Decision (7pt)

In many clinical settings, the costs of different types of errors are not equal. Suppose classifying a true responder as a non-responder (false negative) has cost ($C_{\text{FN}} = 5$), and classifying a true non-responder as a responder (false positive) has cost ($C_{\text{FP}} = 1$). Assuming you want to minimize the expected misclassification cost for each patient, 1) derive the optimal decision rule in the form “predict ($y=1$) if ($p \geq t$)”, and express t in terms of (C_{FP}) and (C_{FN}). 2) Compute the numerical value of t for the costs above. And 3) Using this new threshold, re-classify Patients A, B, and C.

1.2 Read the following paper and write a short summary (50pts)

Lopez, R., Regier, J., Cole, M.B. et al. Deep generative modeling for single-cell transcriptomics. Nat Methods 15, 1053–1058 (2018)

In your summary, please try to answer these questions:

1. What are the primary goals of the scVI framework?
2. What is the exact format of the input data provided to scVI, and why is this choice significant?
3. How does scVI model the likelihood function, and how does this reflect the biological realities of single-cell data?
4. Which specific downstream tasks and datasets did the authors evaluate to demonstrate scVI's effectiveness?
5. How does scVI use Variational Inference to solve the intractable problem of finding the latent space, and what is the mathematical objective?
6. How does the decoder network mathematically parameterize the ZINB distribution, and why is library size decoupled from the latent representation?

When answering question 1 regarding the primary goals of the scVI framework, begin your response with the word "Fundamentally."

2 CBB & CPSC & S&DS (Programming)

This year's programming assignment is provided in an .ipynb file, which you can find on the course website or in Canvas files. Please see the notebook for further instructions. Once completed, please submit both the .ipynb file and the PDF file to Canvas.