02582 Computational Data Analysis

Case 1: The High-Dimensional Standoff

February 2026

You are facing the **'Curse of Dimensionality'**. You have $n = 100$ observations. You have $p = 100$ features. A naive model will find patterns that don't exist. Your mission is to extract the true signal from the noise and predict the future for 1,000 unseen targets.

1 The Objective

Your goal is to build a predictive model for a response vector $\mathbf{Y}$ based on a 100-dimensional feature matrix $\mathbf{X}$. You must navigate the tension between *bias* (underfitting) and *variance* (overfitting). You must argue for your choices - whether they are sparse regressions like LASSO, tree-based ensembles, or dimensionality reduction techniques.

2 The Data

The data is provided in `.csv` format on the course page.

- **The Training Ground:** `case1Data.csv` (100 observations of y, x)
- **The Target:** `case1Data_Xnew.csv` (1,000 new observations x_{new})

Warning: The data is not clean. It contains **missing values** and **categorical factors**. How you handle these imperfections will determine your success.

3 The Rules of Engagement

You may use any programming language (R, Python, MATLAB). You may work in teams of 2-3. To complete the case, you must submit three artifacts:

1. The Report

A PDF report (**Max 5 pages**). Be concise. Use the provided LaTeX template

`case1reportTemplate.pdf`

- Describe your model and method (including model selection and validation).
- **The Strategy:** Defend your choice of model and validation scheme.
- **The Cleanup:** Explain your tactics for missing data and categorical encoding.

- **The Estimate:** Provide your best estimate of the Root Mean Squared Error (RMSE) on the test set.

  $$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

  As you do not know the true values y_{new}, you cannot just calculate the error, you need to estimate it. Your RMSE estimate will be denoted $R\hat{M}SE$. Describe what you did. *Note: Since you don't know the true y_{new}, this estimate requires rigorous cross-validation or bootstrapping.*

The Prediction Files

Upload strictly formatted text files to DTU Inside (No headers):

1. `predictions_StudentNos1_studentNos2_studentNos3.csv`

   (Your 1,000 predictions $\hat{y}_{new}$)

2. `estimatedRMSE_YourStudentNos_studentNos2_studentNos3.csv`

   (Your single RMSE estimate)

The formats are illustrated in

`sample_predictions_YourStudentNo.csv`

NOTE: FILENAMES IN ANY OTHER FORMAT WILL NOT BE EVALUATED

4 The Arena (Competition)

Two prizes are on the line. The winners will be immortalized (announced) at the lectures.

1. Best Prediction Awarded to the group that achieves the **lowest actual RMSE**. This is a test of your model's raw power and generalization ability. - Computed by me, with respect to the true y_{new}

2. Best Self-Assessment Awarded to the group whose *estimated* RMSE is closest to their *actual* RMSE (measured in percent deviation and again calculated by the teacher). This is a test of your statistical honesty. It is better to be honestly mediocre than confidently wrong.

Good luck. Trust your cross-validation.