Statistics and Data Analysis
Unit 05 – Lecture 01 Notes

Tofik Ali

February 14, 2026

Topic

Intro to feature selection, feature engineering, and dimensionality reduction.

Learning Outcomes

- Differentiate feature selection vs dimensionality reduction
- Explain why too many features can hurt (overfitting, cost)
- Describe a simple feature engineering pipeline
- Identify target leakage in engineered features

Detailed Notes

These notes are designed to be read alongside the slides. They expand each slide bullet into plain-language explanations, small worked examples, and common pitfalls. When a formula appears, emphasize (1) what each symbol means, (2) the assumptions needed to use it, and (3) how to interpret the final number in the problem context.

Why Features

- Features are how models see data
- Goal: represent signal and reduce noise
- Bad features -> bad models

Selection vs Reduction

- Selection keeps a subset of original features
- Reduction creates new components (e.g., PCA)
- Validate choices using CV

Exercises (with Solutions)

Exercise 1: Selection or reduction

Dropping 30 out of 100 features is selection or reduction?

Solution

- Feature selection (subset).

Exercise 2: Leakage

Is using final exam score to predict final grade leakage?

Solution

- Yes; it contains future/target information.

Exercise 3: Engineering example

Give one time-based engineered feature.

Solution

- Day-of-week, month, time-since-last-event, rolling average, etc.

Exit Question

Why can adding more features sometimes reduce test accuracy?

Demo (Python)

Run from the lecture folder:

```
python demo/demo.py
```

Output files:

- `images/demo.png`
- `data/results.txt`

References

- Montgomery, D. C., & Runger, G. C. *Applied Statistics and Probability for Engineers*, Wiley.
- Devore, J. L. *Probability and Statistics for Engineering and the Sciences*, Cengage.
- McKinney, W. *Python for Data Analysis*, O'Reilly.