Statistics and Data Analysis
Unit 04 – Lecture 08 Notes

Tofik Ali

February 14, 2026

Topic

Train/test split, k-fold cross-validation, and hyper-parameter tuning (grid/random search).

Learning Outcomes

- Explain train/validation/test split roles
- Describe k-fold cross-validation
- Explain grid search vs random search
- Avoid data leakage using pipelines

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.

Cross-validation

- CV estimates generalization more stably than one split
- k-fold repeats train/validate across folds
- Average score guides selection

Hyper-parameter Tuning

- Grid search tries all combos
- Random search samples combos efficiently
- Never tune on the test set

Exercises (with Solutions)

Exercise 1: Grid size

3 parameters with 4 values each: how many combinations?

Solution

- $4^3 = 64$

Exercise 2: Leakage

Is scaling on full dataset before split leakage?

Solution

- Yes; fit preprocessing on training only.

Exercise 3: Why CV

Why is a single train-test split misleading sometimes?

Solution

- Performance depends on split; CV reduces variance.

Exit Question

Why must you never use the test set to choose hyperparameters?

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.