Statistics and Data Analysis
Unit 05 – Lecture 02 Notes
Feature Selection Methods (Filter/Wrapper/Embedded)

Tofik Ali

February 17, 2026

Topic

Filter, wrapper, and embedded feature selection methods (overview).

How to Use These Notes

These notes are written for students who are seeing the topic for the first time. They follow the slide order, but add the missing 'why', interpretation, and common mistakes. If you get stuck, look at the worked exercises and then run the Python demo.

Course repository (slides, demos, datasets): `https://github.com/tali7c/Statistics-and-Data-Analysis`

Time Plan (55 minutes)

- 0–10 min: Attendance + recap of previous lecture
- 10–35 min: Core concepts (this lecture's sections)
- 35–45 min: Exercises (solve 1–2 in class, rest as practice)
- 45–50 min: Mini demo + interpretation of output
- 50–55 min: Buffer / wrap-up (leave 5 minutes early)

Slide-by-slide Notes

Title Slide

State the lecture title clearly and connect it to what students already know. Tell students what they will be able to do by the end (not just what you will cover).

Quick Links / Agenda

Explain the structure of the lecture and where the exercises and demo appear.

- Overview

- Filter Methods
- Wrapper/Embedded
- Exercises
- Demo
- Summary

Learning Outcomes

- Explain filter methods (variance, correlation, mutual information)
- Explain wrapper methods (RFE) at a high level
- Explain embedded methods (lasso, tree importance) at a high level
- Discuss pros/cons of each approach

Why these outcomes matter. **Correlation** measures the strength of a linear association between two variables. It is symmetric (no X/Y direction) and does not imply causation. Outliers can inflate or hide correlation, so always look at the scatter plot. **Lasso (L1)** can shrink some coefficients exactly to zero, acting like automatic feature selection. This can improve interpretability, but it can be unstable when predictors are highly correlated.

Filter Methods: Key Points

- Fast scoring without training many models
- Examples: variance threshold, correlation with target
- May miss interactions

Explanation. **Correlation** measures the strength of a linear association between two variables. It is symmetric (no X/Y direction) and does not imply causation. Outliers can inflate or hide correlation, so always look at the scatter plot.

Wrapper/Embedded: Key Points

- Wrapper: search subsets using a model (slow)
- Embedded: selection during training (lasso, trees)

Explanation. **Lasso (L1)** can shrink some coefficients exactly to zero, acting like automatic feature selection. This can improve interpretability, but it can be unstable when predictors are highly correlated.

Exercises (with Solutions)

Attempt the exercise first, then compare with the solution. Focus on interpretation, not only arithmetic.

Exercise 1: Low variance

If a feature is almost constant, keep it?

Solution

- Usually no; low variance adds little information.

Exercise 2: Redundant features

Two features have corr=0.99. What might you do?

Solution

- Drop one or use regularization/PCA.

Walkthrough. PCA finds new axes (principal components) that capture maximum variance. It is a rotation of the feature space. Because PCA is variance-based, it is sensitive to scaling: standardize features first unless all features are already comparable.

Exercise 3: Wrapper trade-off

Why is RFE slower than filters?

Solution

- It trains many models on many subsets.

Mini Demo (Python)

Run from the lecture folder:

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

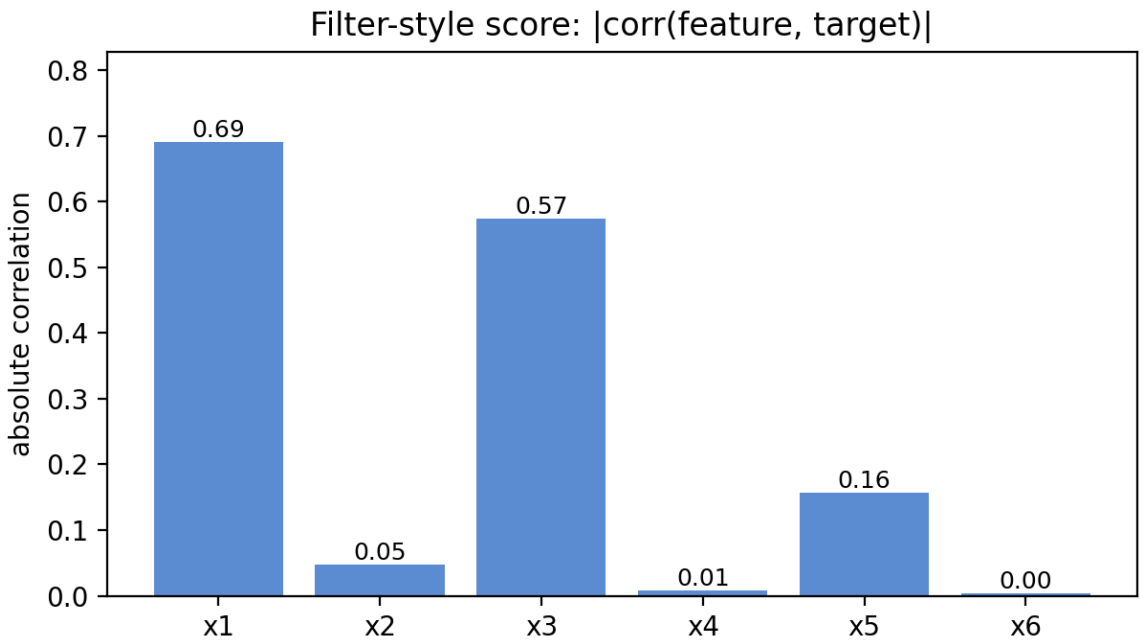
Output files:

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

What to show and say.

- Creates many features and demonstrates a simple selection approach.
- Shows how removing noisy/redundant features can improve validation score.
- Use it to compare filter vs wrapper vs embedded ideas conceptually.

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

When would you prefer a fast filter method over a wrapper method?

Suggested answer (for revision). Use fast filters when you have many features and limited compute/time, or as an initial screening before more expensive wrappers.

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.

Appendix: Slide Deck Content (Reference)

The material below is a reference copy of the slide deck content. Exercise solutions are explained in the main notes where applicable.

Title Slide

https://github.com/tali7c/Statistics-and-Data-Analysis

Quick Links

Overview **Filter Methods** **Wrapper/Embedded** **Exercises** **Demo** **Summary**

Agenda

- Overview
- Filter Methods
- Wrapper/Embedded
- Exercises
- Demo
- Summary

Learning Outcomes

- Explain filter methods (variance, correlation, mutual information)
- Explain wrapper methods (RFE) at a high level
- Explain embedded methods (lasso, tree importance) at a high level
- Discuss pros/cons of each approach

Filter Methods: Key Points

- Fast scoring without training many models
- Examples: variance threshold, correlation with target
- May miss interactions

Wrapper/Embedded: Key Points

- Wrapper: search subsets using a model (slow)
- Embedded: selection during training (lasso, trees)

Exercise 1: Low variance

If a feature is almost constant, keep it?

Solution 1

- Usually no; low variance adds little information.

Exercise 2: Redundant features

Two features have corr=0.99. What might you do?

Solution 2

- Drop one or use regularization/PCA.

Exercise 3: Wrapper trade-off

Why is RFE slower than filters?

Solution 3

- It trains many models on many subsets.

Mini Demo (Python)

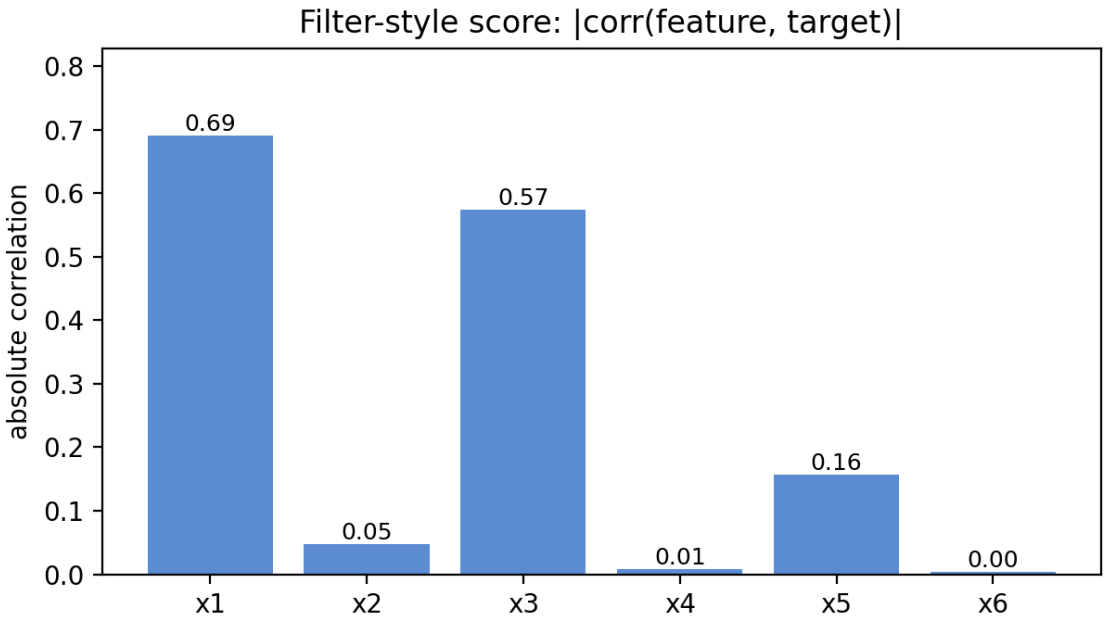
Run from the lecture folder:

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

Outputs:

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

Demo Output (Example)



Summary

- Key definitions and the main formula.
- How to interpret results in context.
- How the demo connects to the theory.

Exit Question

When would you prefer a fast filter method over a wrapper method?