Least squares, principal components, and clustering — review sheet

Use this after attempting the worked examples and practice. Explain each claim in words before calculating.

Key ideas

- Choose a scientific output before selecting a method.
- Regression uses a target; PCA and clustering here do not.
- Report test error with its units, split and simple baseline.
- Centering and scaling are fitted transformations.
- Projection can discard information, while a full basis change need not.
- Cluster labels are arbitrary; metric and number of centers are choices.

Common errors and repairs

Fitting preprocessing on all rows before a test split. Split first and fit transformations on training rows.

Reading explained variance as prediction accuracy. It describes feature variation under the chosen metric.

Treating raw cluster numbers as class predictions. Group identities require external meaning; label permutations are equivalent.

Claiming a unit conversion changed the specimens. It changed raw numerical geometry; state the metric or standardization.

Retrieve without your notes

- What are the units of a response residual?
- Why can a PCA direction change sign without changing reconstruction?
- What does a rank-deficient line fit reveal?
- What is needed to interpret a cluster scientifically?

Readiness check

- Fit a straight-line baseline with an explicit response and coefficient units.
- Calculate held-out RMSE and compare it with a training-mean baseline.

- Center and project two-feature data while explaining what PCA preserves.
- Trace one two-center assignment/update and distinguish cluster labels from known classes.
- Fit transformations on training rows and test the effect of changing a feature unit.