Scientific Computing in Python / 2027 Lecture 4

Regression, classification, and multivariate data

Lecture notes

Synthetic experiments record temperature, duration, and measured yield. Regression estimates yield as a number. Classification estimates a score for reaching a stated yield threshold. Principal components and clustering describe the predictor geometry. We keep the statistical targets and the fitted transformations distinct throughout.

1 The design matrix

Slides for this section

Let T_i and d_i be the temperature and duration of experiment i, and y_i its measured yield. An affine model uses an intercept and one coefficient for each predictor. The columns of the design matrix have different units; the corresponding coefficients supply the units needed to produce yield.

$$A = \begin{pmatrix} 1 & T_1 & d_1 \\ \vdots & \vdots & \vdots \\ 1 & T_n & d_n \end{pmatrix}, \qquad \widehat{y} = A\beta.$$

$$\widehat{y}_i = \beta_0 + \beta_T T_i + \beta_d d_i.$$

Least-squares equations

$$\widehat{\beta} = \arg\min_{\beta} \|A\beta - y\|_2^2, \qquad A^{\mathsf{T}}(A\widehat{\beta} - y) = 0.$$

Use a QR/SVD-based solve. An explicit inverse is unnecessary.

2 A success event

Slides for this section

Define the actual label from measured yield: $z_i = 1$ when $y_i \geq 8$ mg. This threshold defines the event to predict. The model's score threshold is a separate choice applied to a fitted probability-like score. Observed yield must not be placed among predictors when the task is to predict success before yield is measured.

$$z_i = \mathbf{1}_{\{y_i \geq 8 \text{ mg}\}}, \qquad p_i = \sigma(w_0 + w_T T_i + w_d d_i).$$

$$\widehat{z}_i = \mathbf{1}_{\{p_i \geq 0.5\}}.$$

Two distinct thresholds

The 8 mg threshold defines the actual scientific outcome.

The 0.5 score threshold defines the decision rule. Moving the second threshold changes false-positive and false-negative counts without changing the measured labels.

3 Training-only standardization

Slides for this section

A predictor with larger numerical units can dominate Euclidean distances and affect gradient-descent scaling. Standardize using training means and standard deviations. The transformation is then frozen for held-out observations. A constant predictor must receive an explicit treatment because its standard deviation is zero.

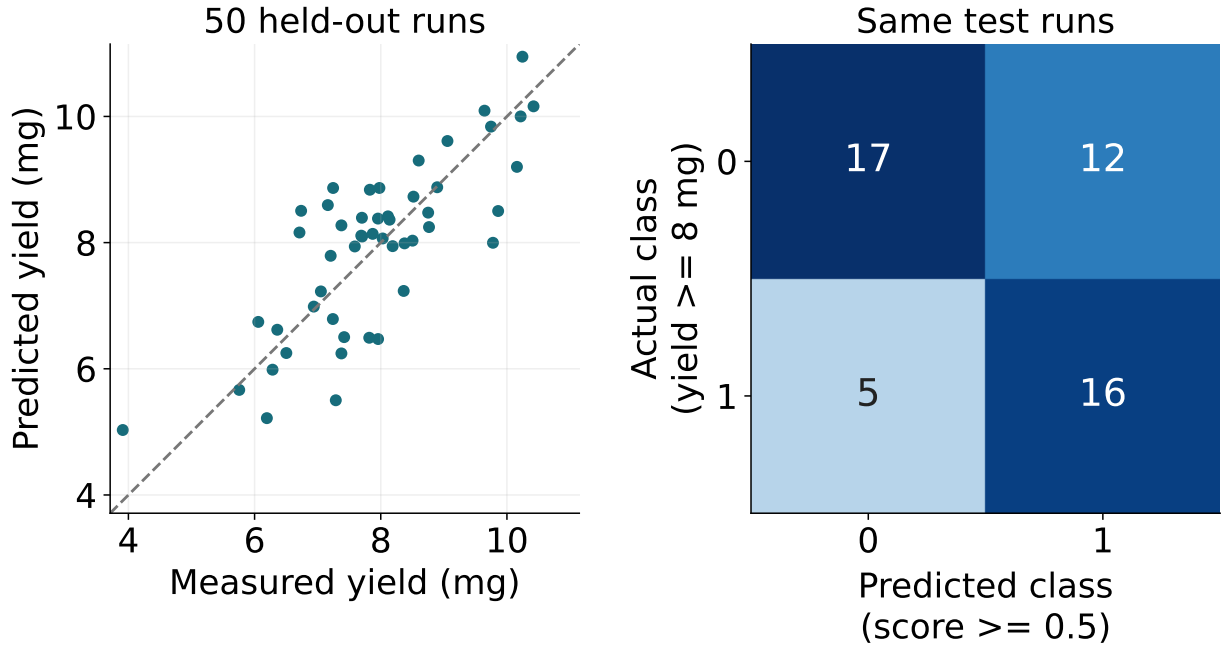
$$x'_{ij} = \frac{x_{ij} - \mu_{j,\text{train}}}{s_{j,\text{train}}}.$$

The same μ_j, s_j are used for all later predictions.

Prediction on a new experiment

```
X_test_scaled = (X_test - train_mean) / train_scale
prediction = model(X_test_scaled)
```

Re-estimating the scale from test rows would change the fitted procedure.



Both panels use the same 50 held-out synthetic runs. The regression diagonal marks exact agreement. Classification uses measured yield ≥8 mg for the actual label and model score ≥0.5 for prediction.

4 Regression and classification errors

Slides for this section

RMSE measures yield discrepancies in milligrams. Classification metrics count threshold decisions. A prediction can have a small yield error yet cross the success boundary, especially near 8 mg. Conversely a large yield error can leave the class unchanged when both values lie far on the same side.

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i}(y_i - \widehat{y}_i)^2}.$$

$$\mathrm{precision} = \frac{TP}{TP + FP}, \qquad \mathrm{recall} = \frac{TP}{TP + FN}.$$

Near-boundary observations

An actual yield of 8.1 mg and a prediction of 7.9 mg differ by only 0.2 mg, but fall on opposite sides of the success threshold.

Report the numerical target and decision target separately.

5 Principal components

Slides for this section

Center the training predictor matrix. A first principal direction maximizes projected predictor variance under a unit-norm constraint. It does not use the response labels. Its usefulness for prediction therefore cannot be concluded from the fraction of predictor variance retained.

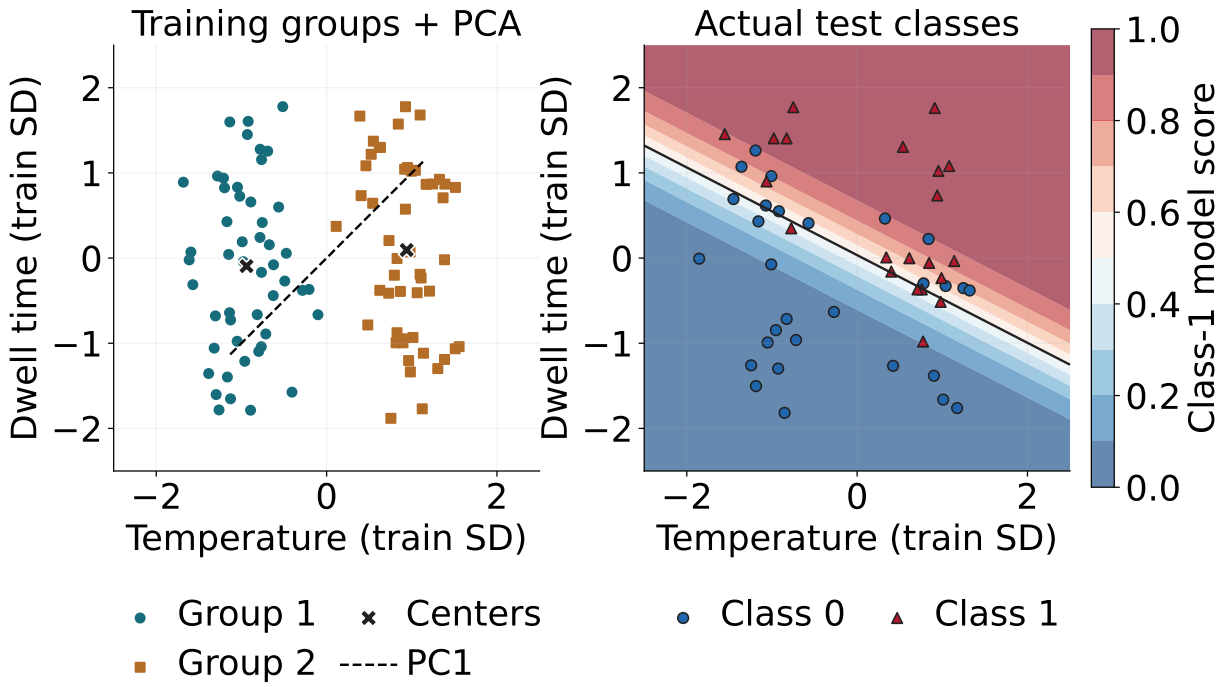
$$\max_{\|v\|_2=1} v^{\mathsf{T}} X_c^{\mathsf{T}} X_c v, \qquad X_c^{\mathsf{T}} X_c v = \lambda v.$$

The score of a new row is $s = (x - \mu)v_1$.

Reconstruction error

$$\widehat{x} = \mu + s v_1^{\mathsf{T}}.$$

The lost component $(x - \mu) - s v_1^{\mathsf{T}}$ is orthogonal to v_1. It may still contain information about yield.



Left colors show training-only two-center memberships and the dashed line marks PC1. Right points show actual test success labels over the fitted logistic score field. The color meanings differ between panels.

6 Distance-based groups

Slides for this section

Two-center clustering partitions standardized predictor rows by proximity to fitted centers. The assignments are unsupervised and need not agree with success labels. A graph colored by clusters and a graph colored by actual classes therefore use different meanings of color even if their point locations are identical.

$$\min_{\mu_1,\mu_2,c_i} \sum_i \|x_i - \mu_{c_i}\|_2^2.$$

Cluster assignments depend on predictor geometry, not on the success event.

Evaluation after fitting

Fit scaling, regression, classification, projection and cluster centers using training data.

Evaluate each operation on the held-out experiment using its own stated target. Do not select the best-looking method from repeated test-set inspection.

Exercises

P27-4-1. Define the prediction time

A different synthetic experiment records pressure before a run and final concentration afterward. Success is final concentration ≥6 mg/L. Total: 10 points.
a Identify a regression target, a binary label and one legitimate pre-run feature.

b Explain why final concentration cannot also be a legitimate input to a pre-run success predictor.

c If repeated runs share the same specimen, why might a random row split mislead?

P27-4-2. Compute and compare regression error

Held-out yields are [4,7,9] mg; model predictions are [5,6,8] mg. A predeclared constant baseline predicts 6 mg for each. Total: 10 points.
a Compute the model residuals as observed minus predicted and its RMSE.

b Compute the baseline RMSE and state the bounded comparison.

c What goes wrong if the baseline constant is chosen to minimize error on these same test rows?

P27-4-3. Read counts before scores

For a new classifier, actual-by-predicted counts are [[18,2],[6,14]]. Total: 10 points.
a Name TN, FP, FN and TP, then compute accuracy.

b Compute precision and recall for class 1 and explain each denominator.

c What is precision if the classifier predicts no positives?

P27-4-4. Projection and grouping are not labels

Two standardized input features are projected onto PC1 and grouped by a two-center algorithm. Total: 10 points.
a Does PCA need success labels, and what does its first direction optimize?

b A retained component explains 55% of variance. How much is omitted, and does this quantify prediction accuracy?

c Why can swapping cluster IDs leave the result unchanged? Give one reason the groups may differ from success classes.

References

Original Python Lecture 5. PDF pages 5–18.
https://zanahmad.com/files/scientific-python/Lecture_5__Intro_to_Python.pdf

NumPy least-squares documentation. numpy.linalg.lstsq: minimum residual, rank and singular values.
https://numpy.org/doc/stable/reference/generated/numpy.linalg.lstsq.html

Scikit-learn decomposition guide. Section 2.5.1.1: PCA.
https://scikit-learn.org/stable/modules/decomposition.html#pca

Scikit-learn clustering guide. Section 2.3.2: K-means and inertia.
https://scikit-learn.org/stable/modules/clustering.html#k-means

Scikit-learn model evaluation guide. Binary confusion matrix, precision and recall definitions.
https://scikit-learn.org/stable/modules/model_evaluation.html

Scikit-learn common pitfalls guide. Sections 12.1–12.2.1: preprocessing and data leakage.
https://scikit-learn.org/stable/common_pitfalls.html#data-leakage

Python source materials: Phillip Kerger (2023). Course adaptation and additional exposition: Zan Ahmad.