Scientific Computing in Python / 2027 Lecture 4

Regression, classification, and multivariate data

Zan Ahmad

Proposed 2027; calendar dates pending

The design matrix

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

The design matrix

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

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

Least-squares equations

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

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.

A success event

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

A success event

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

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

Two distinct thresholds

The 8 mg threshold defines the actual scientific outcome.

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.

Training-only standardization

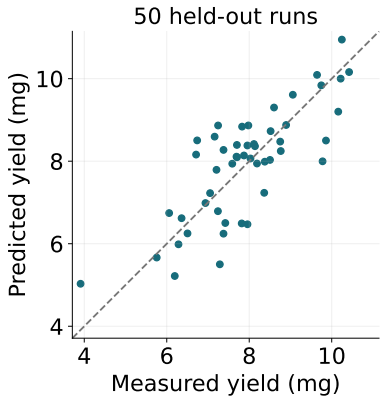
$$x'_{ij} = \frac{x_{ij} - \mu_{j,\text{train}}}{s_{j,\text{train}}}.$$

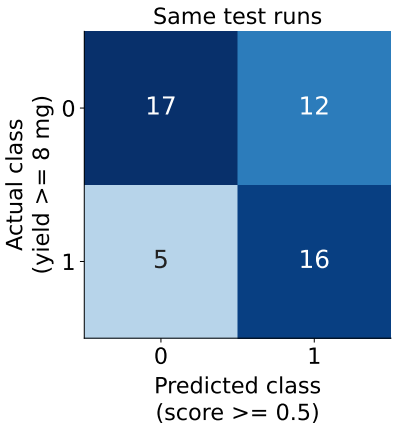
Training-only standardization

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

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

Heldout predictions





Prediction on a new experiment

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

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.

Regression and classification errors

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

Regression and classification errors

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

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

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.

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.

Principal components

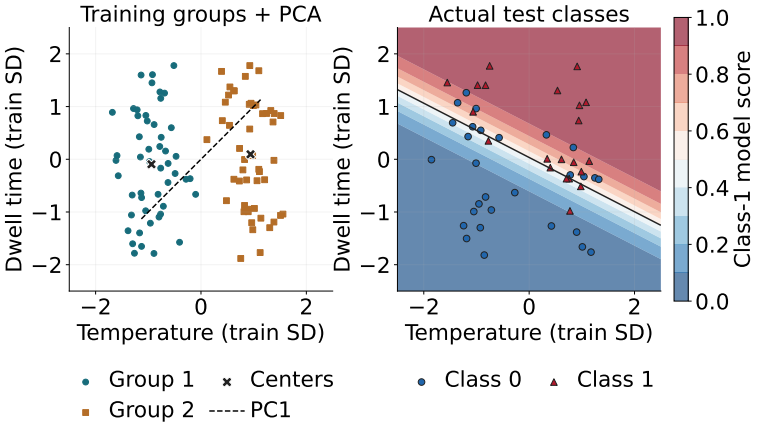
$$\max_{\|v\|_2=1} v^\mathsf{T} X_c^\mathsf{T} X_c v, \qquad X_c^\mathsf{T} X_c v = \lambda v.$$

Principal components

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

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

Projection groups classification



Reconstruction error

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

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.

Distance-based groups

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

Distance-based groups

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

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.

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.