Computational Cardiology Lecture 15

Repeated observations, group leakage and predictive baselines

Zan Ahmad

Fall 2024

Repeated rows share a latent group

$$x_{gr} = s(y_g) + a_g + \epsilon_{gr}.$$

Repeated rows share a latent group

$$x_{gr} = s(y_g) + a_g + \epsilon_{gr}.$$

A stable fingerprint a_g can identify a known group even when it is independent of its outcome across the population.

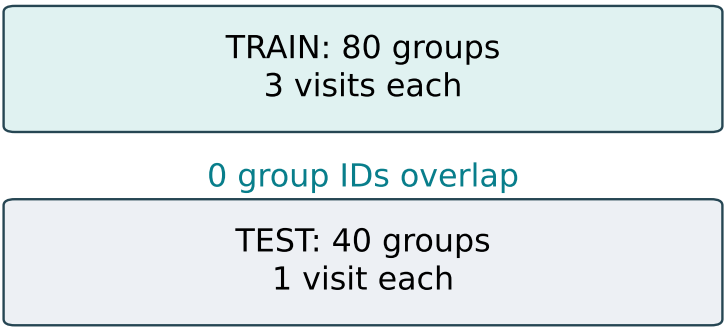
Two prediction problems

Another row from a known group and the first row from a new group contain different prior information.

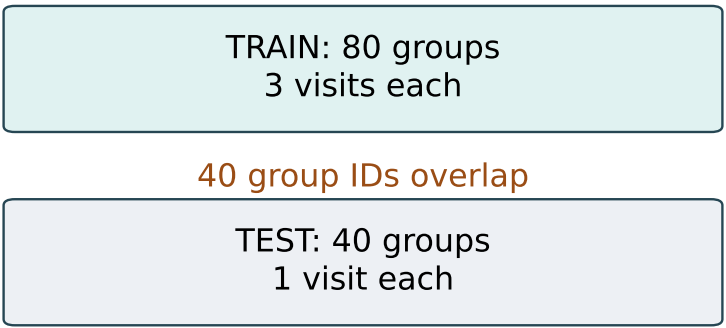
Two prediction problems

Another row from a known group and the first row from a new group contain different prior information. The evaluation split determines which problem is measured.

Unseen-group protocol



Seen-group contamination



The new-group condition

$$G_{train} \cap G_{test} = \varnothing.$$

The new-group condition

$$G_{train} \cap G_{test} = \varnothing.$$

Disjoint row indices do not imply this stronger condition.

A controlled contamination

The same forty test rows are retained while training includes earlier repeats from their groups.

A controlled contamination

The same forty test rows are retained while training includes earlier repeats from their groups. Training and test rows differ, but their fingerprints reveal shared identity.

The constant baseline

$$\widehat{p} = \frac{1}{N_{train}} \sum_i y_i.$$

The constant baseline

$$\widehat{p} = \frac{1}{N_{train}} \sum_{i} y_i.$$

Predict the training majority class, with the stated tie rule. No feature information enters.

Signal-only logistic prediction

$$p = \sigma(w\widetilde{x}_1 + b).$$

Signal-only logistic prediction

$$p = \sigma(w\widetilde{x}_1 + b).$$

This baseline uses the declared population signal without the random fingerprint coordinates.

The penalized objective

$$J = \frac{1}{N} \sum_i [\log(1 + e^{z_i}) - y_i z_i] + \frac{\lambda}{2} w^2.$$

The penalized objective

$$J = \frac{1}{N} \sum_i [\log(1 + e^{z_i}) - y_i z_i] + \frac{\lambda}{2} w^2.$$

The fixed penalty is selected without test outcomes.

A nearest-neighbor rule

$$j_* = \arg \min_{j \in train} \|\widetilde{x} - \widetilde{x}_j\|^2, \qquad \widehat{y} = y_{j_*}.$$

A nearest-neighbor rule

$$j_* = \arg\min_{j \in train} \|\widetilde{x} - \widetilde{x}_j\|^2, \qquad \widehat{y} = y_{j_*}.$$

Near-duplicate fingerprints make recognition easy.

Fit the coordinate transformation on training data

$$\widetilde{x}_k = \frac{x_k - \mu_k^{train}}{s_k^{train}}.$$

Fit the coordinate transformation on training data

$$\widetilde{x}_k = \frac{x_k - \mu_k^{train}}{s_k^{train}}.$$

Test cases reuse the same means and scales.

Standardization defines a distance

$$\|\widetilde{x} - \widetilde{x}_j\|^2 = \sum_k \frac{(x_k - x_{jk})^2}{(s_k^{train})^2}.$$

Standardization defines a distance

$$\|\widetilde{x} - \widetilde{x}_j\|^2 = \sum_k \frac{(x_k - x_{jk})^2}{(s_k^{train})^2}.$$

Refitting a scaler changes the nearest-neighbor metric.

Read the confusion matrix

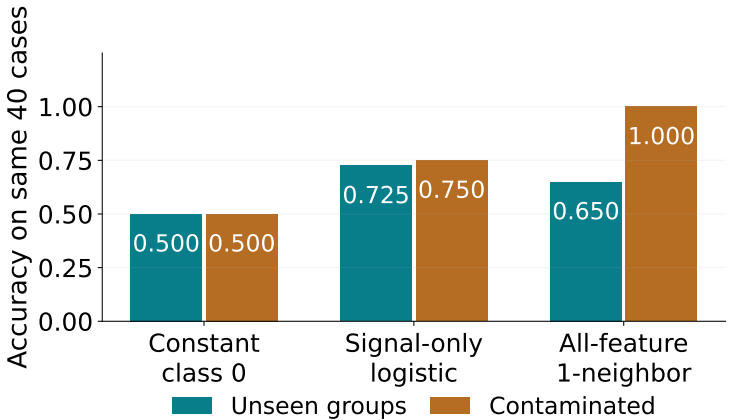
$$\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}.$$

Read the confusion matrix

$$\begin{pmatrix} TN & FP \\ FN & TP \end{pmatrix}.$$

Rows are true classes and columns are predicted classes. Each recall uses its true-class row total.

Accuracy on same 40 cases



Balanced accuracy

$$\mathrm{BA} = \frac{1}{2}\left(\frac{TN}{TN+FP} + \frac{TP}{TP+FN}\right).$$

Balanced accuracy

$$\mathrm{BA} = \frac{1}{2}\left(\frac{TN}{TN+FP} + \frac{TP}{TP+FN}\right).$$

Both classes must be present.

Prevalence-weighted accuracy

$$\text{Accuracy} = (1 - \pi)\text{TNR} + \pi\text{TPR}.$$

Prevalence-weighted accuracy

$$\text{Accuracy} = (1 - \pi)\text{TNR} + \pi\text{TPR}.$$

Balanced accuracy uses equal class weights instead of prevalence weights.

Recognition is not new-group prediction

The contaminated nearest neighbor attains perfect accuracy on this constructed sample.

Recognition is not new-group prediction

The contaminated nearest neighbor attains perfect accuracy on this constructed sample. The result reflects access to the same group fingerprints, not perfect population discrimination.

The target distribution remains an assumption

Holding out people does not automatically hold out sites, calendar periods or acquisition devices.

The target distribution remains an assumption

Holding out people does not automatically hold out sites, calendar periods or acquisition devices. Prediction-time availability and the group-generating distribution remain part of the statistical problem.