Model assessment

23 Feb 2023

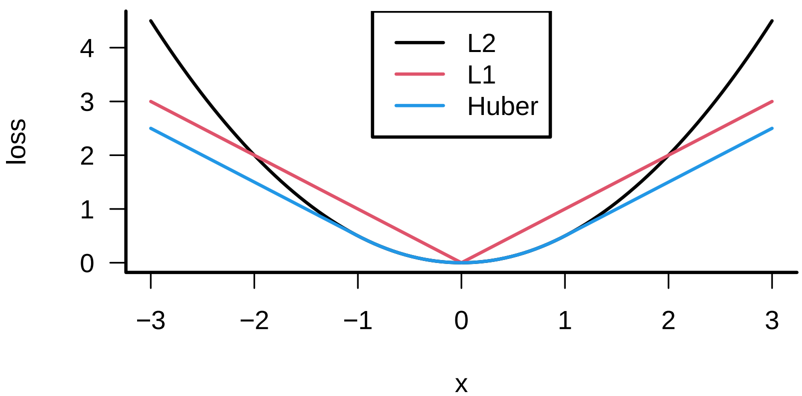
Table of contents


```r
## use help("image-methods", "Matrix")
## lattice graphics: ?lattice:xyplot for details on scales
ifun <- function(x, title = "", ck = FALSE, raster = TRUE) {
    image(Matrix(x),
          sub = "", xlab = "", ylab = "",
          colorkey = ck,
          aspect = "fill",
          scales = list(x = list(draw = FALSE),
                        y = list(draw = FALSE)),
          main = title,
          useRaster = raster
          )
}
```

loss functions (regression/quantitative outcome)

- continuous: L2, L1, **Huber loss**:

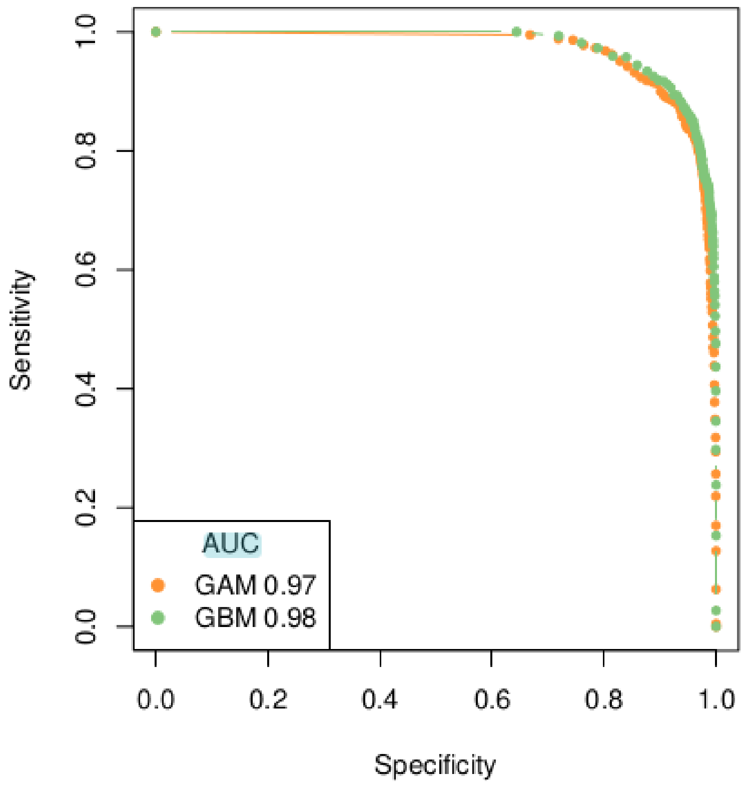
```r
par(las = 1, bty = "l", lwd = 2)
huber <- function(x, d) ifelse(abs(x)<d, x^2/2, d*abs(x)-d/2)
curve(x^2/2, from = -3, to = 3, ylab = "loss")
curve(abs(x), add = TRUE, col = 2)
curve(huber(x, 1), add = TRUE, col = 4)
legend("top", c("L2", "L1", "Huber"), col = c(1, 2, 4), lty = 1)
```



loss functions (classification)

- 0-1
- **deviance**: $-2 \sum I(G = k) \log \hat{p}_k$ = -2 log-likelihood
- deviance generalizes to other distributions
- not a loss function, but worth introducing AUC (area under the curve)
  - may be problematic in terms of implied misclassification costs? (Hand 2009)

Hand, David J. 2009. "Measuring Classifier Performance: A Coherent Alternative to the Area Under the ROC Curve." *Machine Learning* 77 (1): 103–23. https://doi.org/10.1007/s10994-009-5119-5.



a short rant about loss functions

- 0-1 scoring dichotomizes prematurely
- leads to lots of confusing discussion about balancing data sets
- lots of discussion of what to do about imbalanced data sets (SMOTE etc.) (Goorbergh et al. 2022)
- when **should** we balance?
  - when we have to use 0-1 scoring for some technical reason
  - when we have too **much** data (downsampling, i.e., throw away majority class)
- (cf. discussion of variable selection)

Goorbergh, Ruben van den, Maarten van Smeden, Dirk Timmerman, and Ben Van Calster. 2022. "The Harm of Class Imbalance Corrections for Risk Prediction Models: Illustration and Simulation Using Logistic Regression." *Journal of the American Medical Informatics Association*, June, ocac093. https://doi.org/10.1093/jamia/ocac093.

Journal of the American Medical Informatics Association, 2022, Vol. 00, No. 0

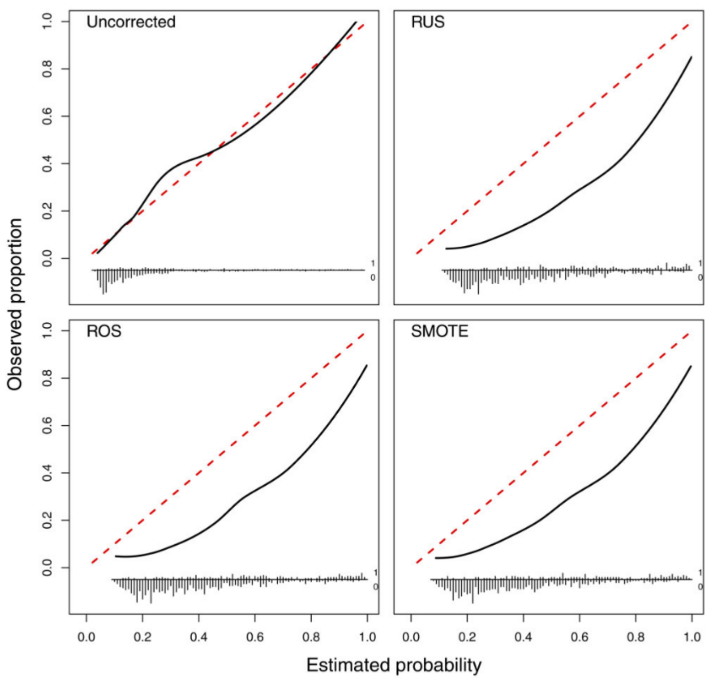


Figure 2. Flexible calibration curves on the test set for the Ridge models to diagnose ovarian cancer.

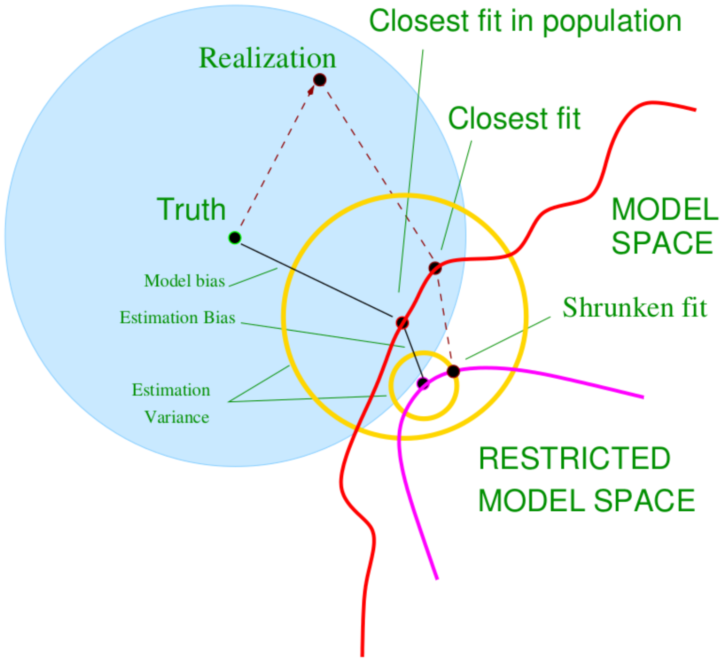
- test error (generalization error): prediction error over a **fixed** independent sample
- **expected** prediction error: test error averaged over test sets
- training error (within-sample): expectation

selection vs assessment

train-validation-test

$$E[f(x_0) - x_0^\top \beta^*]^2 + E[x_0^\top \beta^* - E x_0^\top \hat{\beta}_\alpha]^2$$

- estimation bias = 0 for linear regression etc., positive for ridge etc.



- in-sample error:
  - $C_p = \text{err} + 2\; d/N\sigma_\epsilon^2$
- leakage:
  - non-independence
  - data-dependence of training
- jackknife, bootstrap etc.

calibration