

Methods for classification

- introduction
- linear regression
- discriminant analysis (LDA, QDA)
- logistic regression (+ penalised version)
- support vector machines
- classification version of many methods that we saw for regression (KNN, trees, random forests, boosting, GART, ...)

Idea: allocate a statistical unit to a category or class based on its characteristics (values of its covariates)

$$y_i = G_i \in \{0, 1\} \quad \text{or} \quad G_i \in \{0, 1, \dots, K-1\}$$

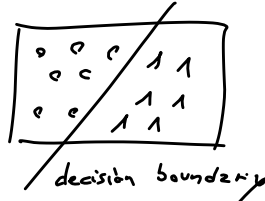
↑ categories or classes

binary response (by far the most used)

examples: healthy / sick
reliable / unreliable
receptive / not receptive (e.g., to a marketing campaign)

Goal of a predictive model (for classification): find a predictive rule for that allocates the new statistical unit to the right class

→ often based on a decision boundary



Typical loss function

0/1 loss, misclassification error $\frac{1}{n} \sum_{i=1}^n \mathbb{1}(\hat{G}_i \neq G_i)$

Example with binary response:

		true value	
		0	1
prediction (by a model)	0	true negative	false negative
	1	false positive	true positive
		all negatives	all positive

predicted positive

NB: the 0/1 loss gives the same weight to false negatives and false positive

sensitivity: proportion of correctly predicted positives = $\frac{\text{true positive}}{\text{all positive}}$

specificity: proportion of correctly predicted negatives = $\frac{\text{true negative}}{\text{all negative}}$

false discovery rate (FDR) = $\frac{\text{false positive}}{\text{predicted positive}}$

ideally as small as possible

ideally 1

ROC curve (Receiver Operating Characteristic)

plot sensitivity vs (1-specificity) as a function of a threshold

AUC = Area under the (ROC) curve

0.5 ≤ AUC ≤ 1
random chance → perfect classification

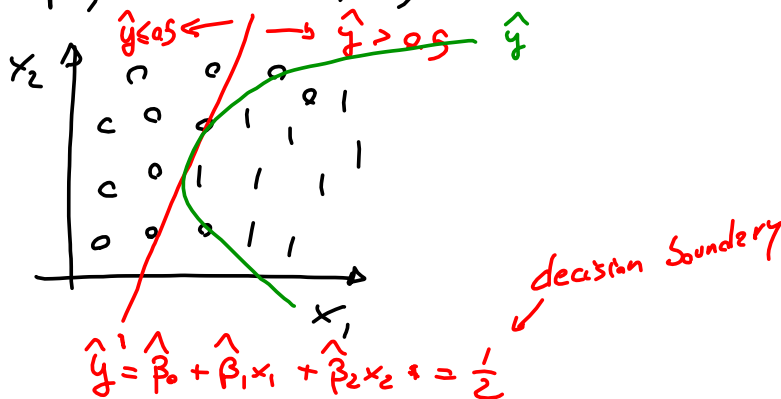
Use linear regression for classification

$$G \in \{0, 1\}$$

$$y = \beta_0 + \beta^T x + \varepsilon$$

$$G = \mathbb{1}(y > 0.5)$$

Example, with $x = (x_1, x_2)$



$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2 = \frac{1}{2}$$

Note:

- often we get good results, but it seems unnatural to use linear regression models for classification ($Y \in \{0, 1\}$ not $\sim \mathcal{N}(\mu, \sigma^2)$);
- all inferential aspects do not work;
- homoscedasticity is also violated;
- possible masking effect