Recap 4

Classification

ISLP 4, ESL 4

Jilles Vreeken
Krikamol Muandet

UNIVERSITÄT DES SAARLANDES

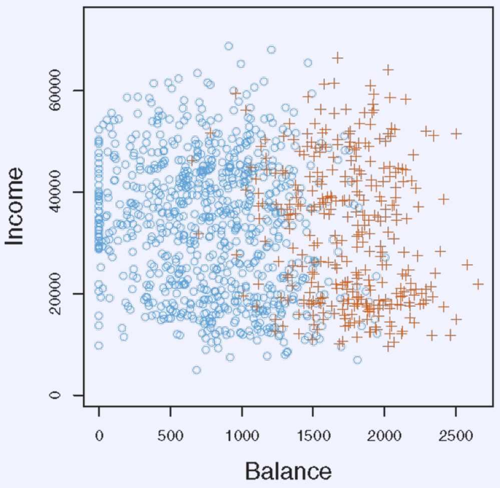
CISPA HELMHOLTZ CENTER FOR INFORMATION SECURITY

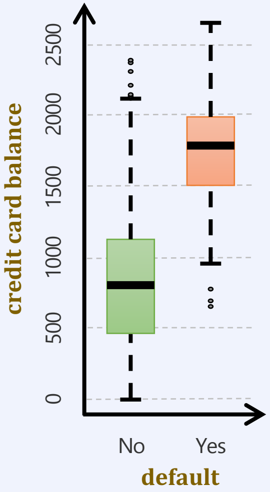
Classification Overview

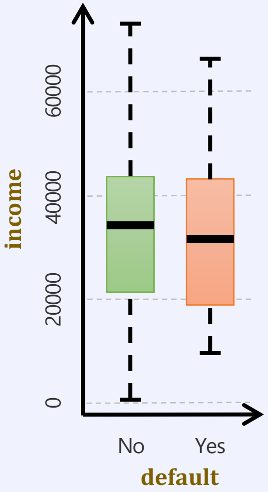
In classification, we want to predict categorical outputs

Example will someone pay back their loan? yes or no?

- inputs: annual **income**, monthly **balance, student** status



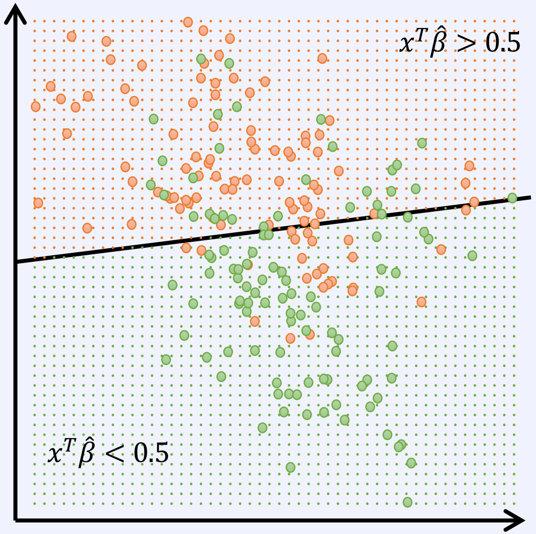




Why not just do linear regression?

Linear regression can actually work for binary classification

- simply code $Y = \begin{cases} 0 & \text{if } \mathbf{green} \\ 1 & \text{if } \mathbf{red} \end{cases}$



Problems:

- Does not generalize to more than two classes
- $Y = \begin{cases} 0 & \text{if } \mathbf{green} \\ 1 & \text{if } \mathbf{red} \\ 2 & \text{if } \mathbf{blue} \end{cases}$ or $Y = \begin{cases} 0 & \text{if } \mathbf{red} \\ 1 & \text{if } \mathbf{blue} \\ 2 & \text{if } \mathbf{green} \end{cases}$
- each imposes a different ordering, and different distances between classes

Logistic Regression

Example Credit **default** data

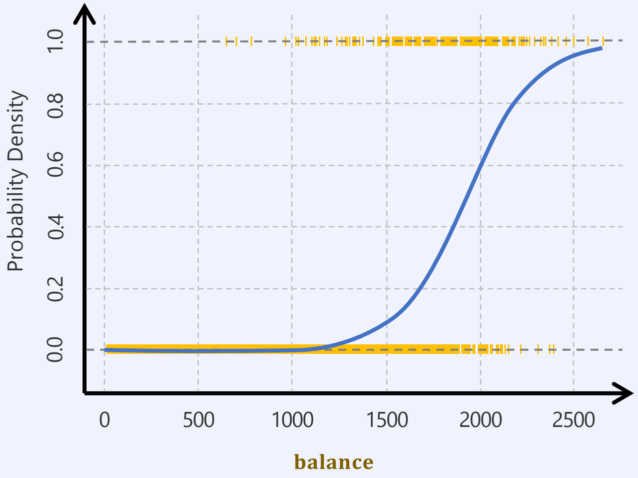
- univariate model, e.g.

$$\Pr(\textbf{default} = \text{yes} \mid \textbf{balance})$$

- simple linear regression models this as

$$f(X) = \beta_0 + \beta_1 X_1$$

- which leads to values outside $[0,1]$



We can map these into $[0,1]$ using the logistic function

probability that $Y = \textbf{yes} = 1$ →

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}$$

- not only are all values now sensible, we also have the

  odds ratio as $\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X}$, and the log-odds (logit) as $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$

Example Single Continuous Predictor

If we increasing X by one unit, we

- Add β_1 to the log-odds --> multiply the odds by e^{β_1}
- If $\beta_1 > 0$, adding X increases $p(X)$
- If $\beta_1 < 0$, adding X decreases $p(X)$

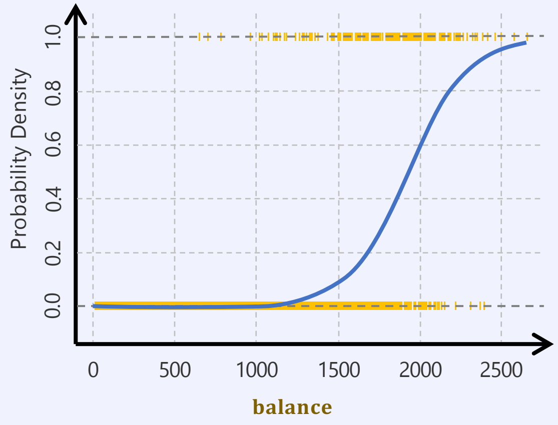
$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \qquad \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \qquad p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Probabilities of **default** given **balance**

- For $\beta_0 = -10.653$ and $\beta_1 = 0.0055$ (balance)

$$\hat{p}(2000) = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

- If we increase **balance** by 1 EUR, this
- Increases the log odds of defaulting by 0.0055
- Multiplies the odds of defaulting by $e^{0.0055} = 1.0055\%$



Example Single Binary Predictor

If we increasing X by one unit, we

- Add β_1 to the log-odds --> multiply the odds by e^{β_1}
- If $\beta_1 > 0$, adding X increases $p(X)$
- If $\beta_1 < 0$, adding X decreases $p(X)$

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \qquad \frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \qquad p(X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

Probabilities of **default** given **student**

$$\hat{p}(\textbf{student} = \text{yes}) = \frac{e^{-3.5041+0.40409\times 1}}{1+e^{-3.5041+0.40409\times 1}} = 0.00431$$

$$\hat{p}(\textbf{student} = \text{no}) = \frac{e^{-3.5041+0.40409\times 0}}{1+e^{-3.5041+0.40409\times 0}} = 0.00292$$

- For $\beta_0 = -3.5041$ and $\beta_1 = 0.4049$
- → Being a student yields a higher prob. of defaulting!

Multiple Logistic Regression

The multivariate logistic regression model is defined as

- $\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X + \cdots \beta_p X_p$ with $p(X) = \frac{e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X+\cdots+\beta_p X_p}}$

Example predicting **default** based on **balance**, **income**, and **student**

$$\hat{p}(\textbf{student} = \text{yes}, \textbf{balance} = 1{,}500, \textbf{income} = 40) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 1}} = 0.058$$

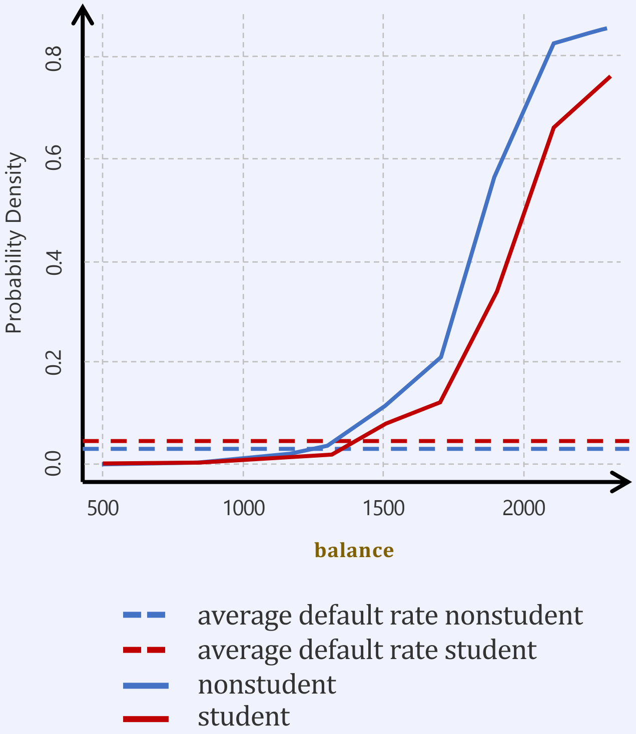
$$\hat{p}(\textbf{student} = \text{no}, \textbf{balance} = 1{,}500, \textbf{income} = 40) = \frac{e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1{,}500+0.003\times 40-0.6468\times 0}} = 0.105$$

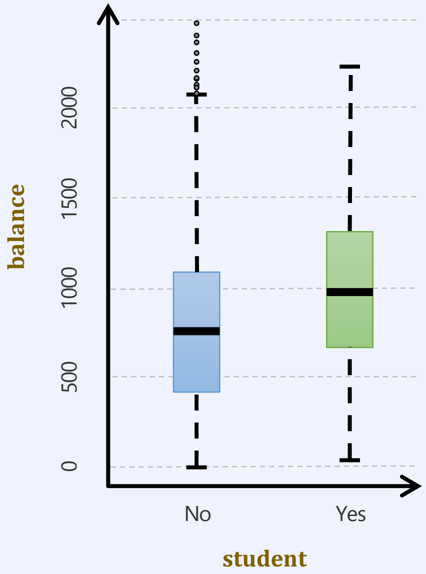
Why is the **student** coefficient positive in the univariate and negative in the multivariate model?

Example Confounding in Logistic Regression

Why is the **student** coefficient positive in the univariate and negative in the multivariate model?

- confounding!
- students have higher **balance**
- students **default** at higher **balance**
- for a fixed value of **balance** and **income**, a **student** is less likely to default than a nonstudent!





Fitting Logistic Regression Models

We usually fit a logistic regression model by maximum likelihood

- log-likelihood function $\ell(\theta) = \sum_{i=1}^{n} \log p_{g_i}(x_i; \theta)$ and density function $p_k(x_i, \theta) = \Pr(\, G = k \mid X = x_i; \theta\,)$
- for a binary problem, class coding $y_i = \begin{cases} 1 \mid g_i = 1 \\ 0 \mid g_i = 0 \end{cases}$ gives us $p_1(x; \theta) = p(x; \theta)$ and $p_2(x; \theta) = 1 - p(x; \theta)$

The log-likelihood then becomes

$$\ell(\beta) = \sum_{i=1}^{n} \{ y_i \log p(x_i; \theta) + (1 - y_i) \log\big(1 - p(x_i; \theta)\big) \} = \sum_{i=1}^{n} \left\{ y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right\}$$

- where $\beta = \{\beta_0, \beta_1, \ldots\}$ and x_i a vector of the input values padded with a constant term $X_0 = 1$