Predictive Modelling

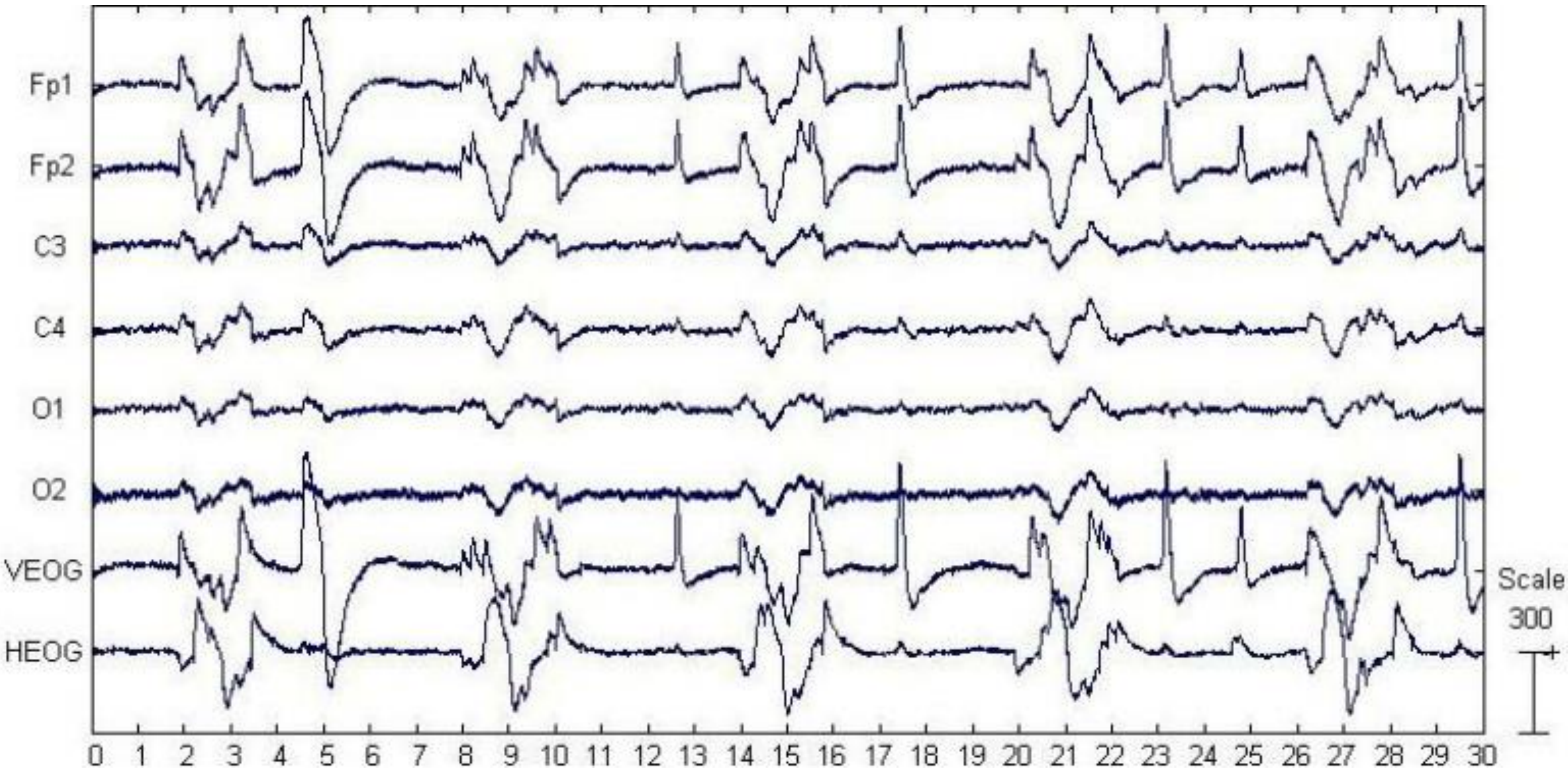
Lecture 4a: Linear Methods for Classification

Outline of this lecture

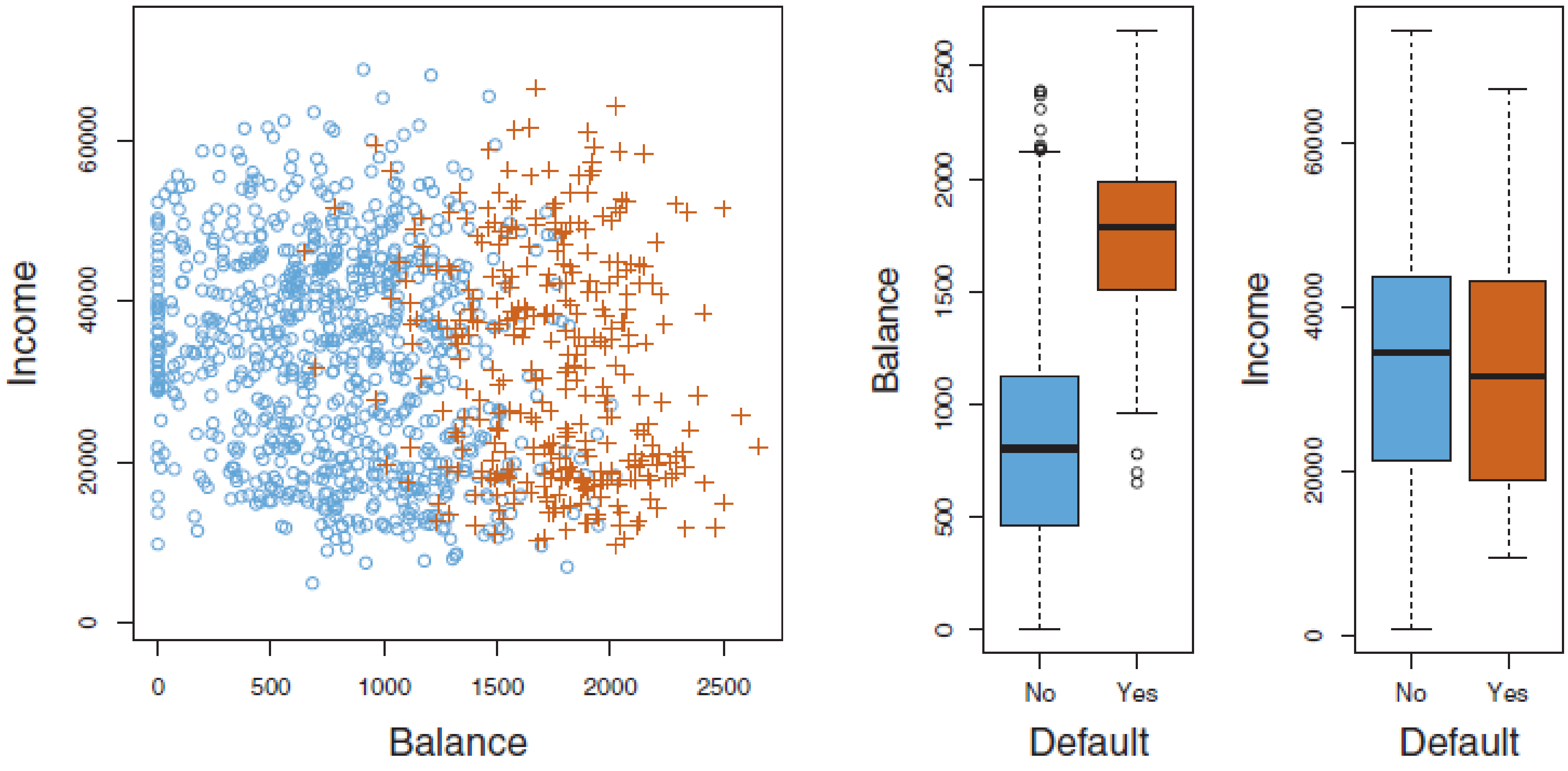
Chapter 4: Linear methods for classification

- 4.1: Introduction to classification
- 4.2: Least-squares for classification
- 4.3: Logistic regression
- From binary to multi-class classification

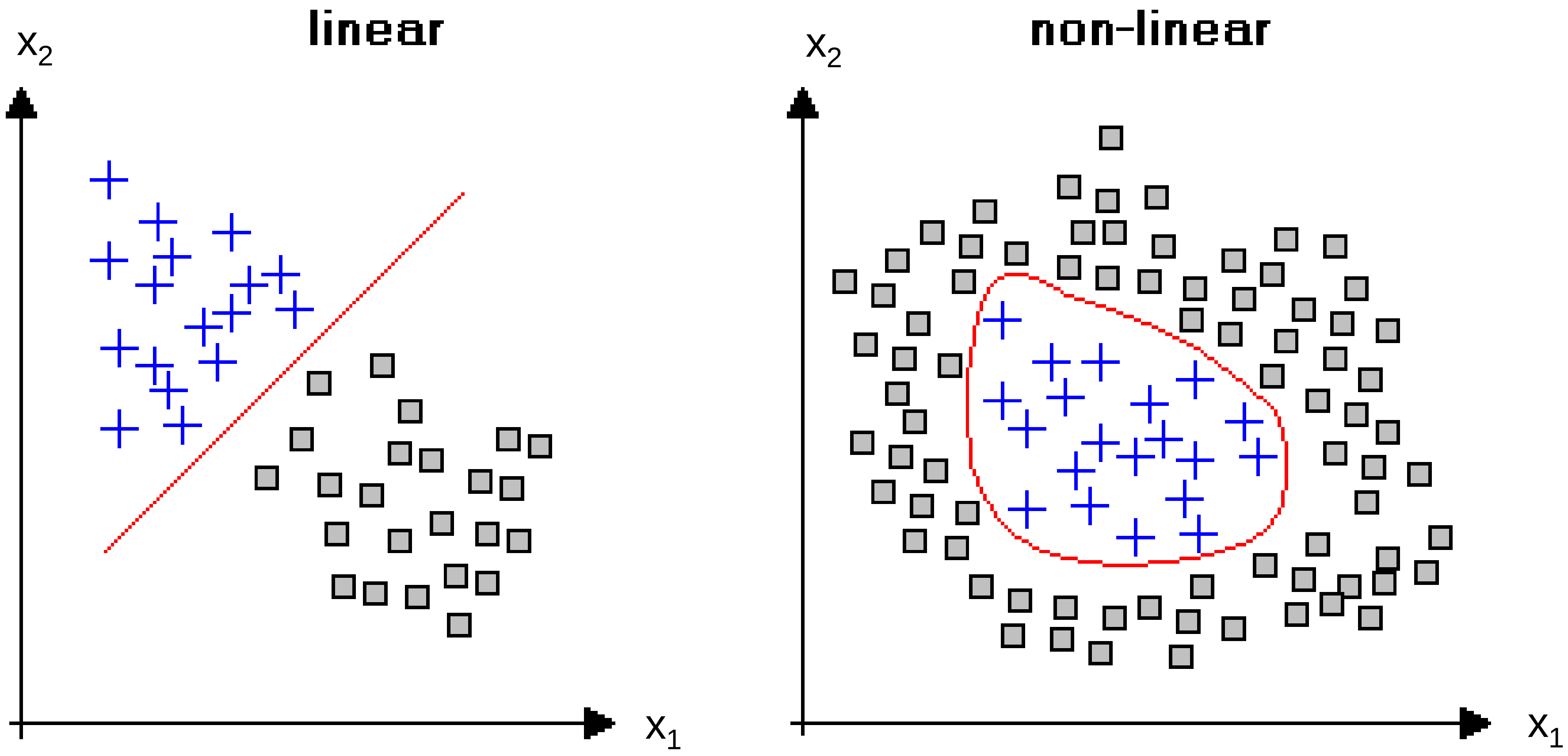
An example of EEG data: how to convert to a standard data frame?



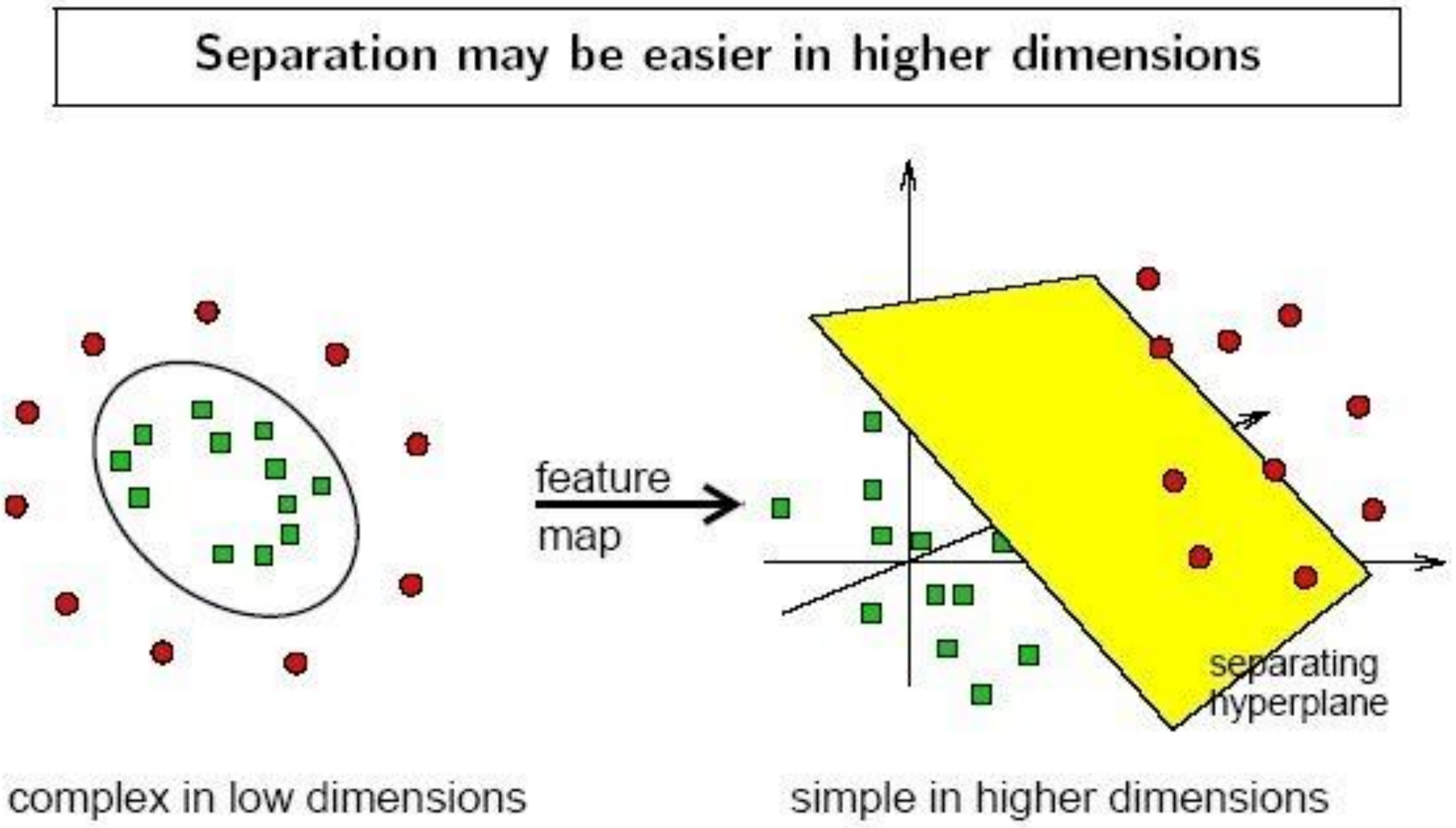
Fitting classification models on the credit card dataset



The decision boundary of a linear model is a hyperplane



Linear models will be a building block for more complicated methods…



How can we apply least-squares to classification problems?

Linear regression model:

$$\hat{y}_i = \sum_{j=0}^{p} w_j x_{ij} = \boldsymbol{w}^T \boldsymbol{x}_i$$

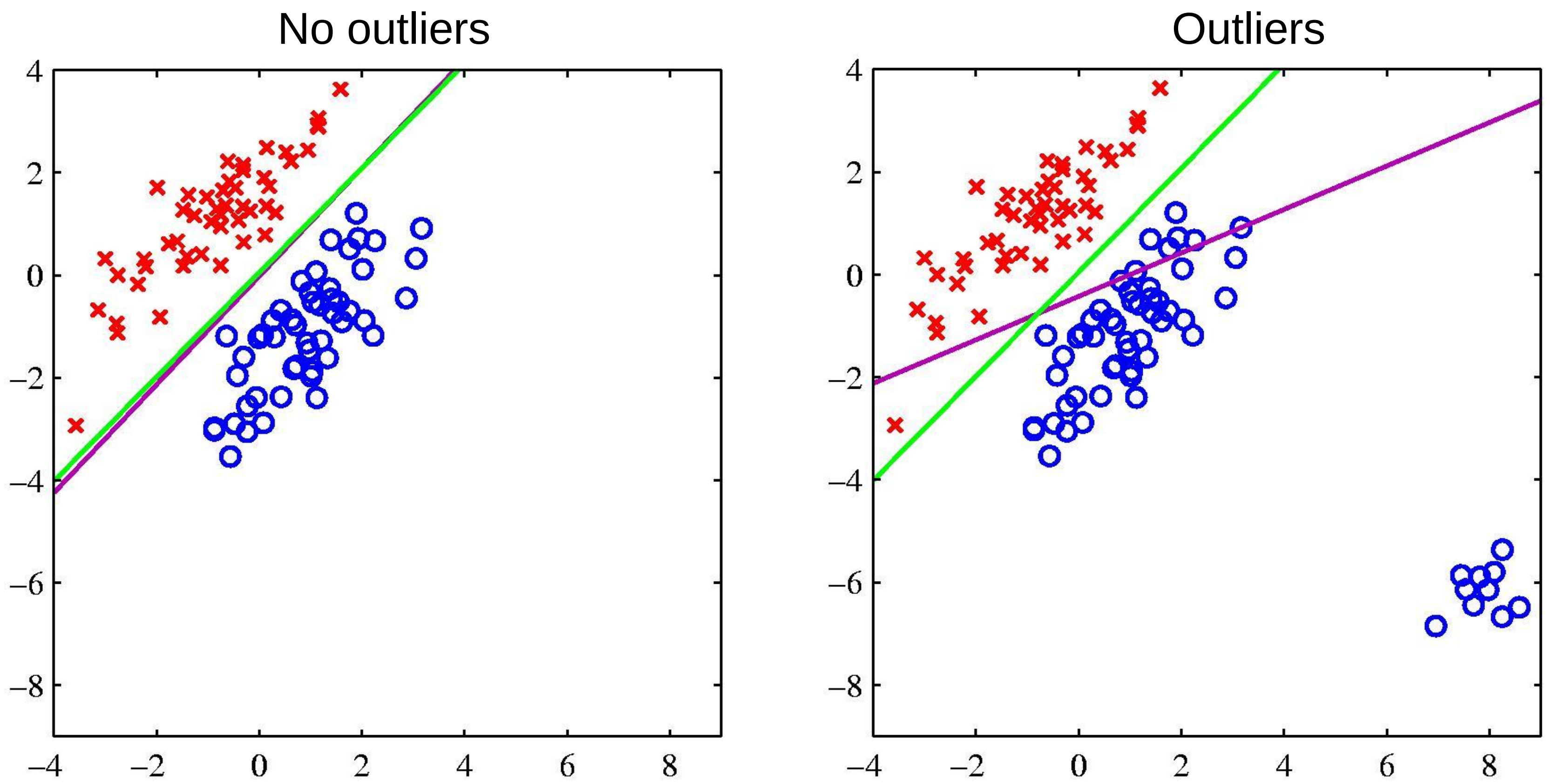
Minimize the residual sum of squares:

$$RSS(\boldsymbol{w}) = \sum_{i=1}^{n} (y_i - \boldsymbol{w}^T \boldsymbol{x}_i)^2$$

Convert to class labels:

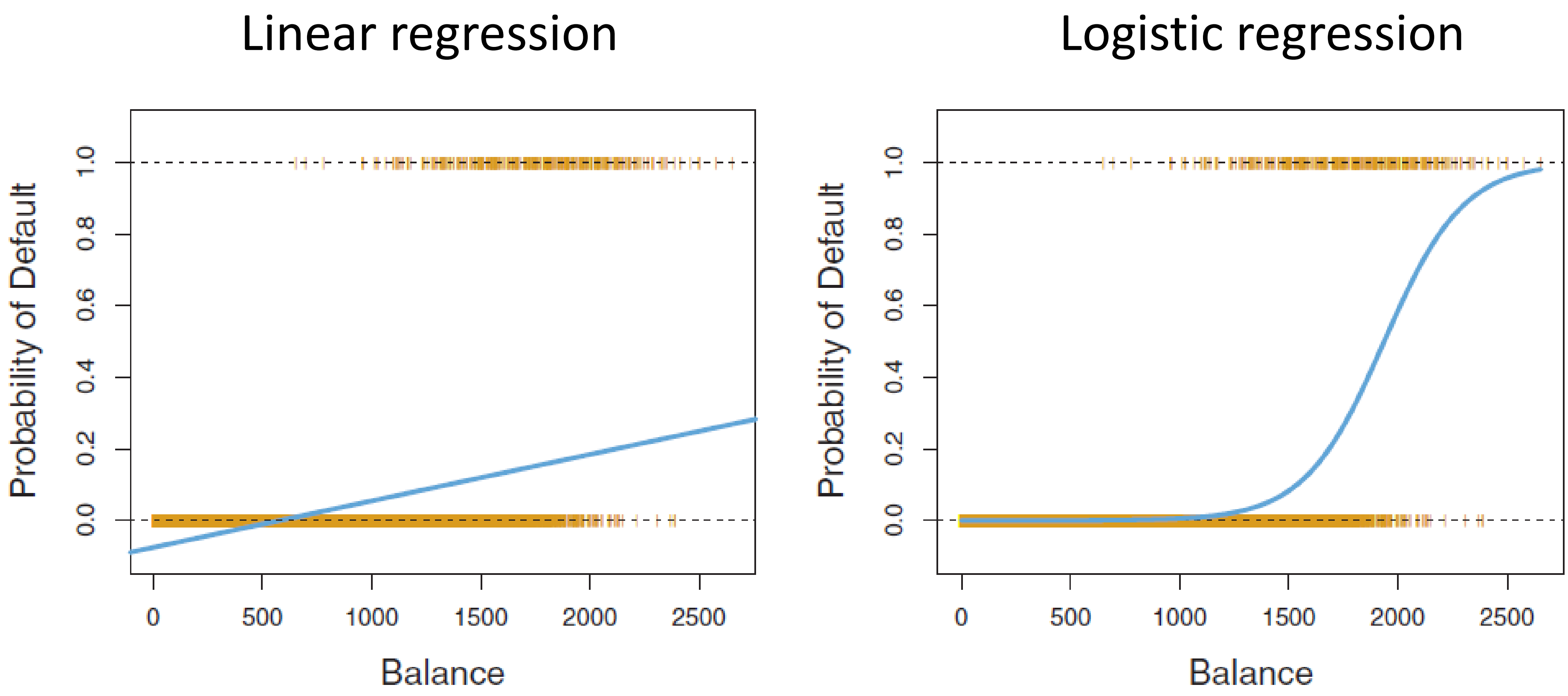
$$\hat{G}_i = \begin{cases} \text{ORANGE}, & \text{if } \hat{y}_i > 0.5 \\ \text{BLUE}, & \text{if } \hat{y}_i \leq 0.5 \,. \end{cases}$$

An advantage of logistic regression over least-squares classification



Purple line = least squares classification **Green line = logistic regression**

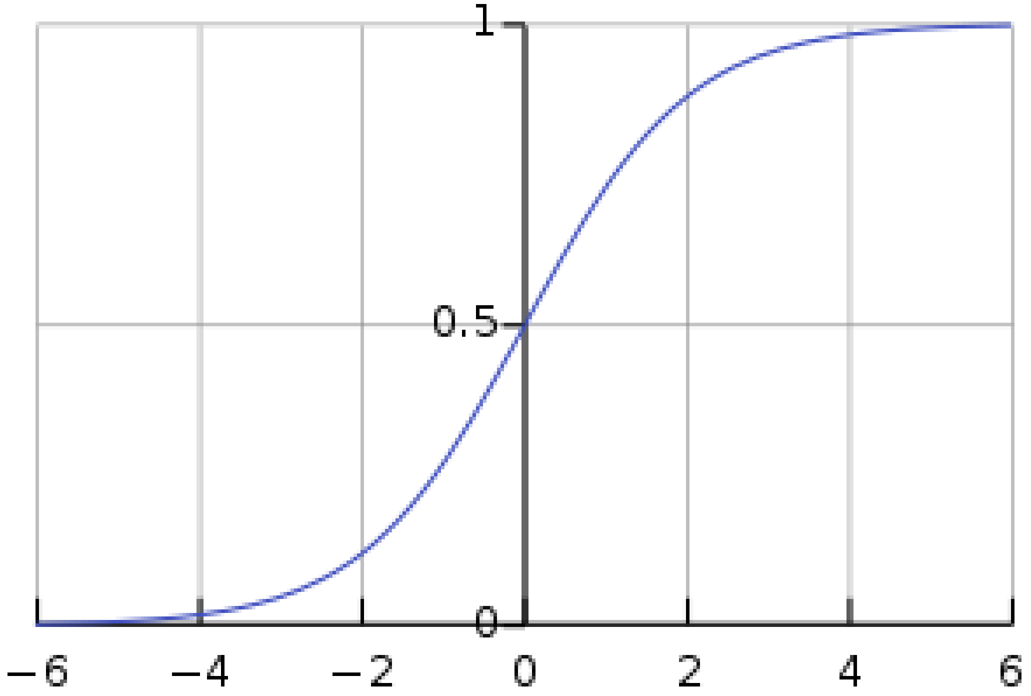
Another advantage of logistic regression over linear regression



$$p(\boldsymbol{x}) = \Pr(y = 1|\boldsymbol{x})$$

Quick math recap: the logistic function (aka sigmoid function)

$$\phi(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{1+e^z}$$



From linear models to posterior probability estimates: the logit-transformation

Linear regression: $$p(x) = w_0 + w_1 x_1$$

Logistic regression: $$p(x) = \frac{e^{w_0 + w_1 x}}{1 + e^{w_0 + w_1 x}}$$

$$\Leftrightarrow \quad \frac{p(x)}{1 - p(x)} = e^{w_0 + w_1 x}$$

Logit transformation:

$$\log\left(\frac{p(x)}{1 - p(x)}\right) = w_0 + w_1 x$$

Logistic regression on the default dataset

	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.6513	0.3612	−29.5	<0.0001
balance	0.0055	0.0002	24.9	<0.0001

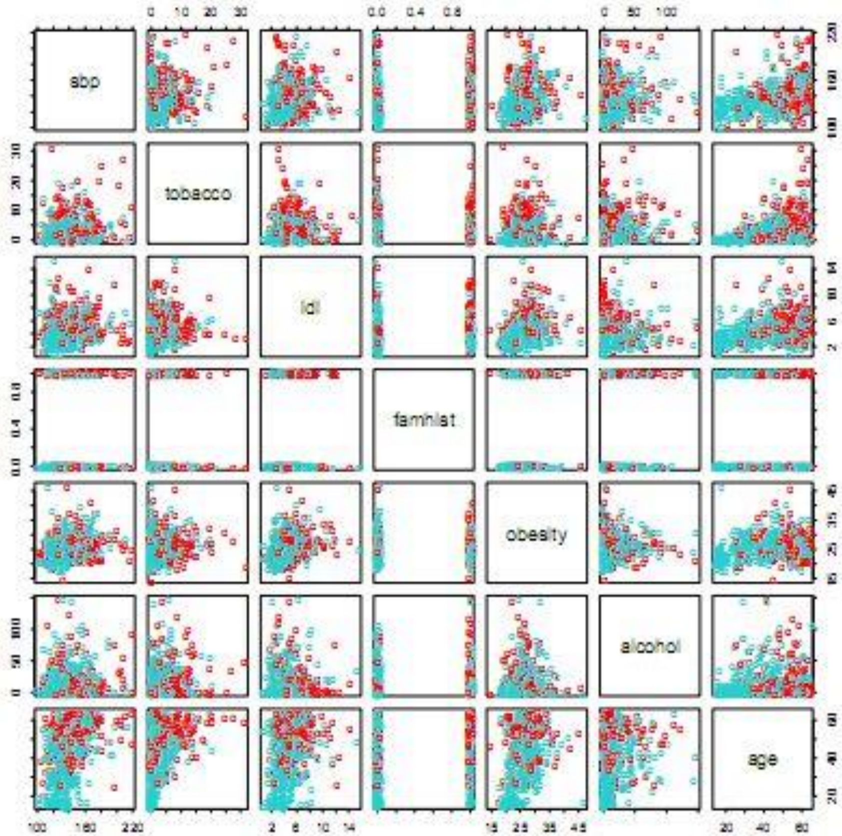
TABLE 4.1. *For the* Default *data, estimated coefficients of the logistic regression model that predicts the probability of* default *using* balance*. A one-unit increase in* balance *is associated with an increase in the log odds of* default *by* 0.0055 *units.*

Making predictions for a new observation with balance 1000 dollar:

$$p(1000) = \frac{e^{w_0+w_1x}}{1+e^{w_0+w_1x}} = \frac{e^{-10,6513+0,0055\times1000}}{1+e^{-10,6513+0,0055\times1000}} = 0,00576$$

Logistic regression applied to the South Africa heart disease dataset

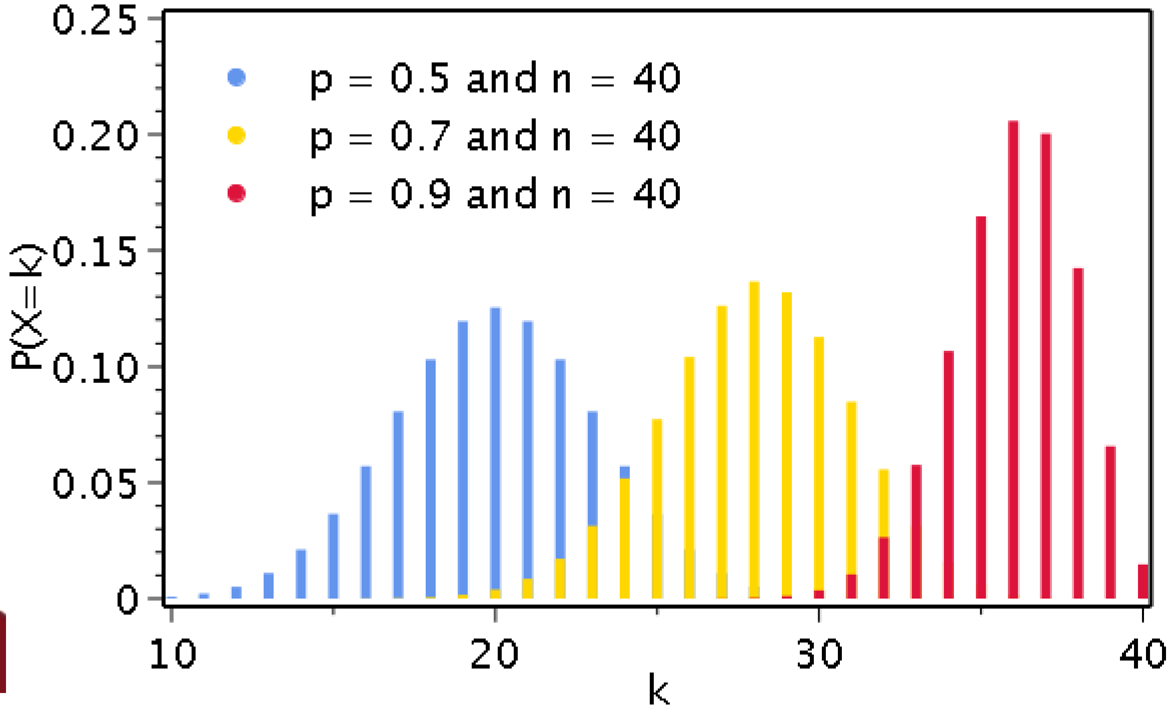
$$\log\left(\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}\right) = w_0 + w_1 x_1 + w_2 x_2 + \cdots = \boldsymbol{w}^T \boldsymbol{x}$$



	Coef.	Std. Error	Z-score
Intercept	-4.130	0.964	-4.285
Sbp	0.006	0.006	1.023
Tobacco	0.080	0.026	3.032
Ldl	0.185	0.057	3.219
Famhist	0.939	0.225	4.178
Obesity	-0.035	0.029	-1.187
Alcohol	0.001	0.004	0.136
age	0.043	0.010	4.184

Quick math recap: the binomial distribution

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$$



Fitting logistic regression models using maximum likelihood estimation

Maximize the likelihood of the training data:

$$l(\boldsymbol{w}) = \prod_{i:y_i=1} p_{\boldsymbol{w}}(\boldsymbol{x}_i) \prod_{i':y_i=0} (1 - p_{\boldsymbol{w}}(\boldsymbol{x}_{i'}))$$

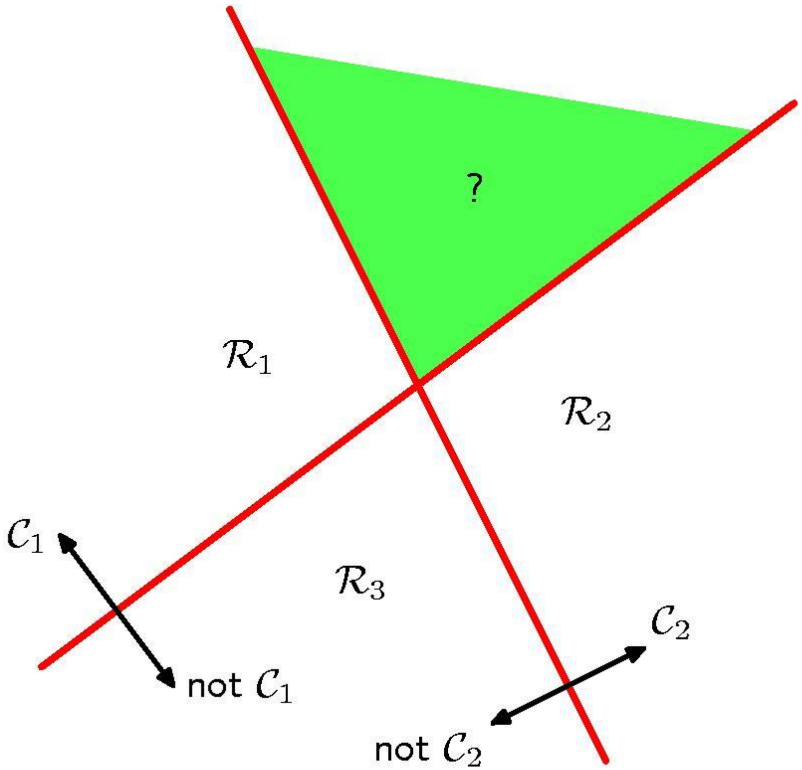
$$= \prod_{i=1}^{n} p_{\boldsymbol{w}}(\boldsymbol{x}_i)^{y_i} (1 - p_{\boldsymbol{w}}(\boldsymbol{x}_i))^{1-y_i}$$

Equivalent to minimizing the negative log-likelihood:

$$l_{\log}(\boldsymbol{w}) = -\sum_{i=1}^{n} \left(y_i \log p_{\boldsymbol{w}}(\boldsymbol{x}_i) + (1 - y_i) \log(1 - p_{\boldsymbol{w}}(\boldsymbol{x}_i))\right)$$

From binary classification to multi-class classification using linear models: the one-versus-all approach

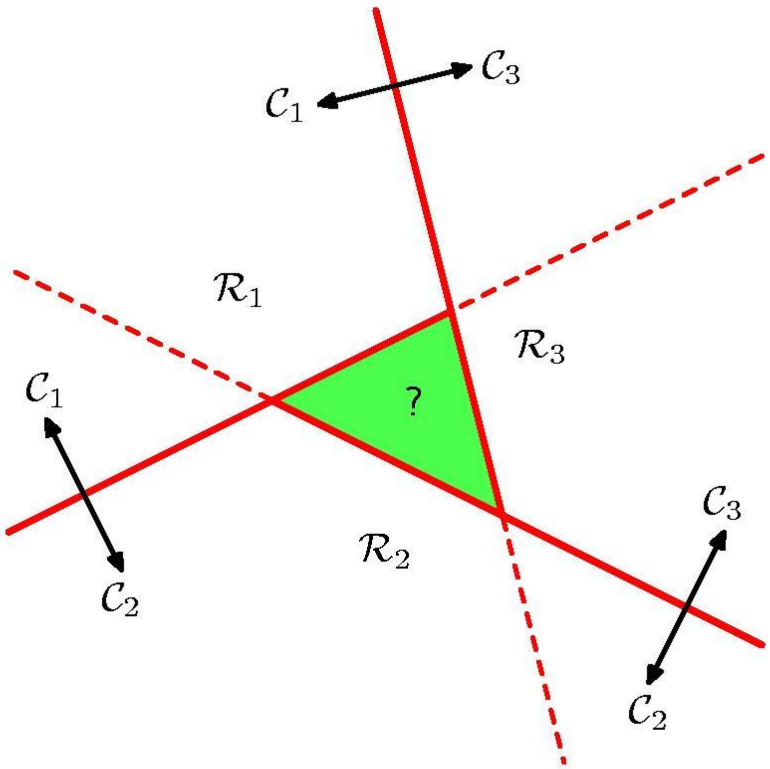
- One-versus-all:
  - **For every class**, solve a binary classification problem where the observations of this class are considered as positive and all the rest as negative
  - For test data, assign observations to the class for which the corresponding **model gives the highest value or highest posterior probability estimate**



$$f(\boldsymbol{x}) = \operatorname{argmax}_{k \in 1,\dots,K} f_k(\boldsymbol{x})$$

From binary classification to multi-class classification using linear models: the one-versus-one approach

- One-versus-one:
  - **For every pair of classes**, build a model where only the observations of these two classes are used
  - For test data, plug the data in every classifier and **apply a voting strategy**
  - For probability estimation, use postprocessing methods



Both approaches can lead to ambiguous decision boundaries!

Multinomial logistic regression

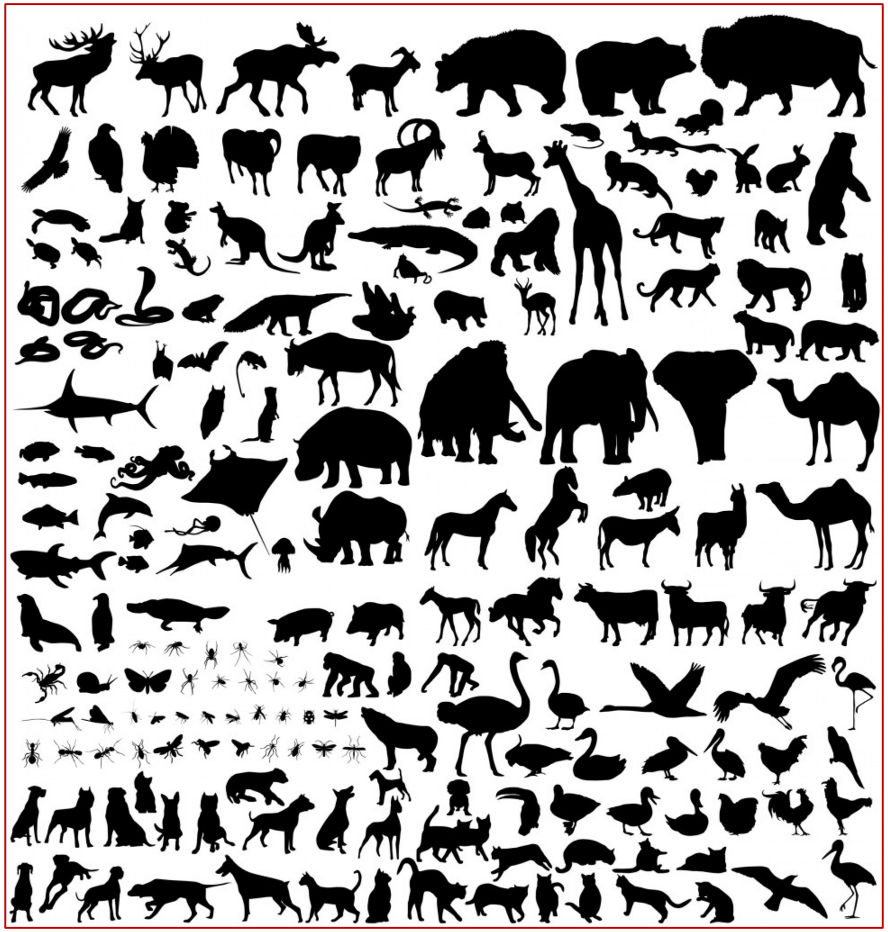
Model the K posterior probabilities via linear functions:

$$
\begin{aligned}
\log\left(\frac{\Pr(y=1|\boldsymbol{x})}{\Pr(y=K|\boldsymbol{x})}\right) &= \boldsymbol{w}_1^T\boldsymbol{x} \\
\log\left(\frac{\Pr(y=2|\boldsymbol{x})}{\Pr(y=K|\boldsymbol{x})}\right) &= \boldsymbol{w}_2^T\boldsymbol{x} \\
&\vdots \\
&= \boldsymbol{w}_{K-1}^T\boldsymbol{x}
\end{aligned}
$$

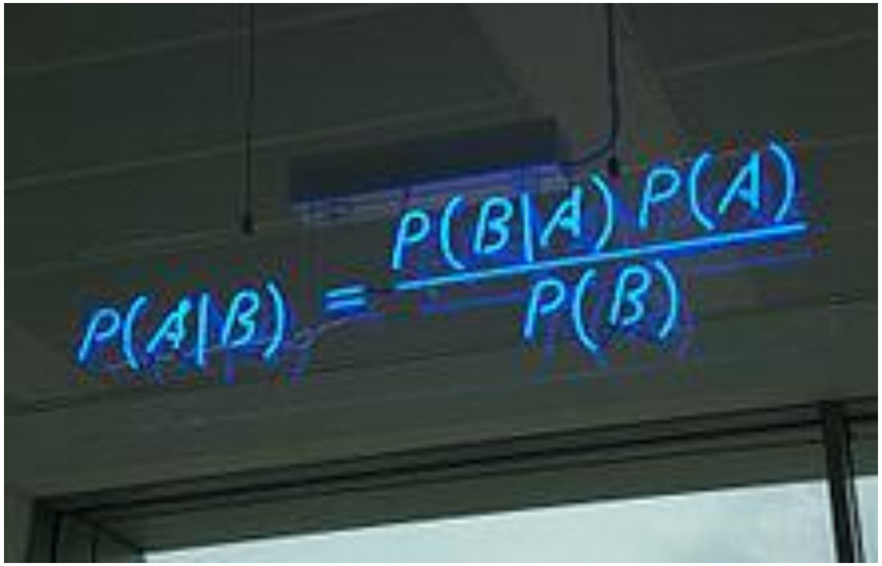
$$
\Pr(y=k|\boldsymbol{x}) = \frac{e^{\boldsymbol{w}_k^T\boldsymbol{x}}}{1+\sum_{l=1}^{K-1} e^{\boldsymbol{w}_l^T\boldsymbol{x}}}
$$

$$
\Pr(y=K|\boldsymbol{x}) = \frac{1}{1+\sum_{l=1}^{K-1} e^{\boldsymbol{w}_l^T\boldsymbol{x}}}
$$

Extreme multi-class classification



Math recap for the next lecture: Bayes' rule



Thomas Bayes

Born: 1702 in London, England

$$P(A) = \sum_n P(A \cap B_n) = \sum_n P(A|B_n)P(B_n)$$

Bayes' rule: an example

- Assume a drug test that produces 99% true positive results for drug users and 99% true negative results for non-drug users.
- Suppose that 0.5% of people are users of the drug.
- What is the **probability** that a randomly selected individual with a positive test is a user?