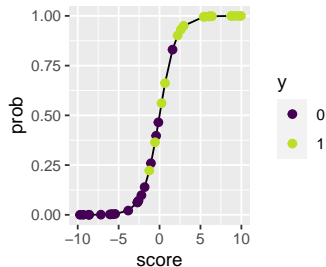


Introduction to Machine Learning

Classification: Logistic Regression



Learning goals

- Understand the definition of the logit model
- Understand how a reasonable loss function for binary classification can be derived
- Know the hypothesis space that belongs to the logit model



MOTIVATION

A **discriminant** approach for directly modeling the posterior probabilities $\pi(\mathbf{x} \mid \boldsymbol{\theta})$ of the labels is **logistic regression**.

For now, let's focus on the binary case $y \in \{0, 1\}$ and use empirical risk minimization.

$$\arg \min_{\boldsymbol{\theta} \in \Theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \arg \min_{\boldsymbol{\theta} \in \Theta} \sum_{i=1}^n L\left(y^{(i)}, \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right).$$



A naive approach would be to model

$$\pi(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}.$$

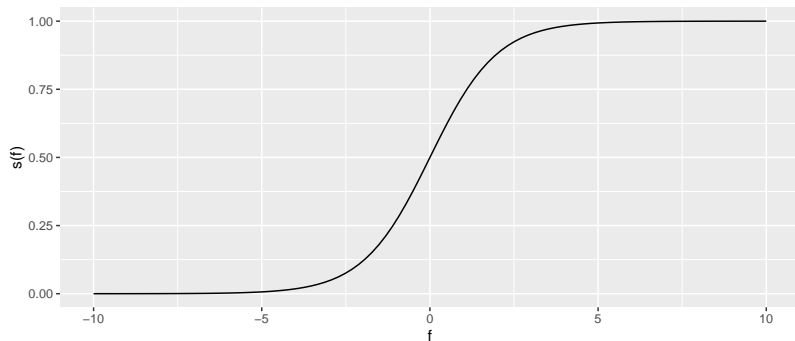
NB: We will often suppress the intercept in notation.

Obviously this could result in predicted probabilities $\pi(\mathbf{x} \mid \boldsymbol{\theta}) \notin [0, 1]$.

LOGISTIC FUNCTION

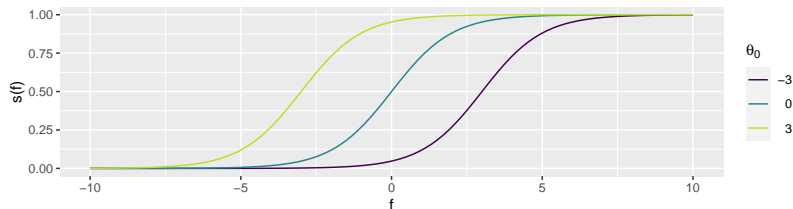
To avoid this, logistic regression “squashes” the estimated linear scores $\theta^\top \mathbf{x}$ to $[0, 1]$ through the **logistic function** s :

$$\pi(\mathbf{x} \mid \theta) = \frac{\exp(\theta^\top \mathbf{x})}{1 + \exp(\theta^\top \mathbf{x})} = \frac{1}{1 + \exp(-\theta^\top \mathbf{x})} = s(\theta^\top \mathbf{x})$$

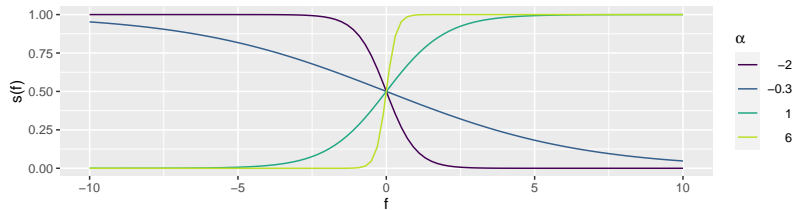


LOGISTIC FUNCTION / 2

The intercept shifts $s(f)$ horizontally $s(\theta_0 + f) = \frac{\exp(\theta_0 + f)}{1 + \exp(\theta_0 + f)}$



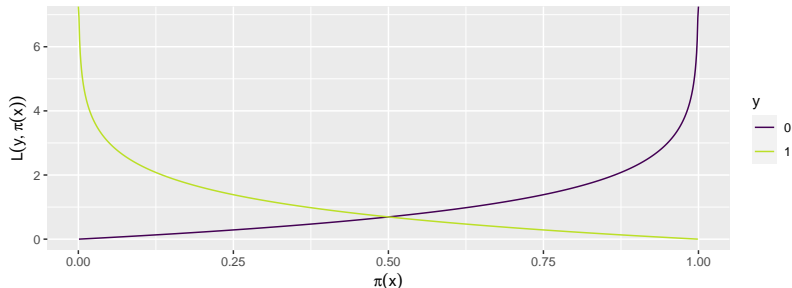
Scaling f like $s(\alpha f) = \frac{\exp(\alpha f)}{1 + \exp(\alpha f)}$ controls the slope and direction.



BERNOULLI / LOG LOSS

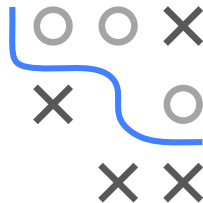
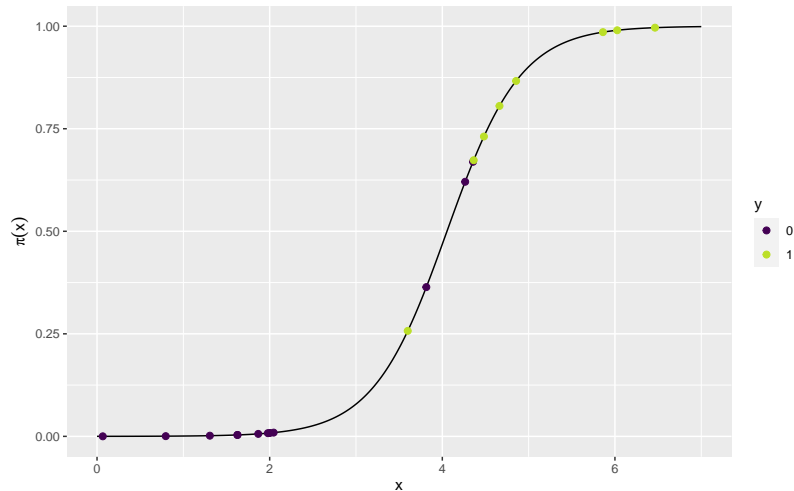
We need to define a loss function for the ERM approach:

- $L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x}))$
- Penalizes confidently wrong predictions heavily
- Called Bernoulli, log or cross-entropy loss
- We can derive it from the negative log-likelihood of Bernoulli / logistic regression model in statistics
- Used for many other classifiers, e.g., in NNs or boosting



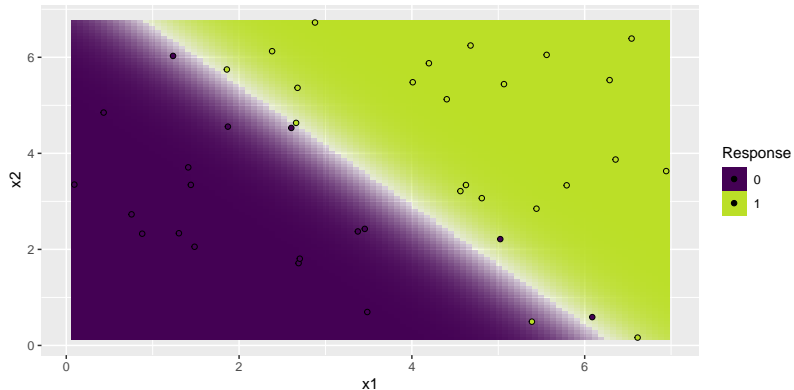
LOGISTIC REGRESSION IN 1D

With one feature $\mathbf{x} \in \mathbb{R}$. The figure shows data and $\mathbf{x} \mapsto \pi(\mathbf{x})$.

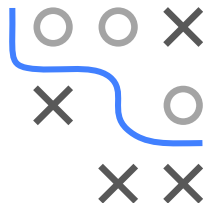
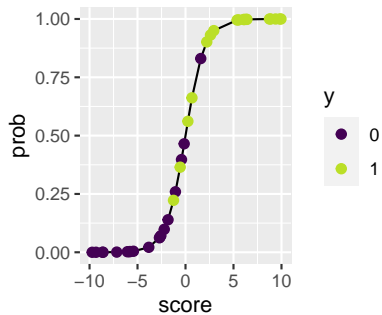
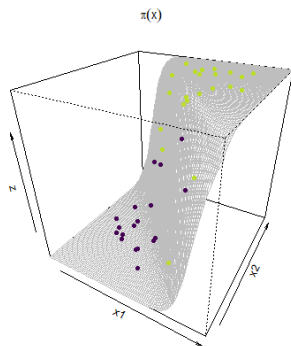


LOGISTIC REGRESSION IN 2D

Obviously, logistic regression is a linear classifier, as $\pi(\mathbf{x} | \boldsymbol{\theta}) = s(\boldsymbol{\theta}^\top \mathbf{x})$ and s is isotonic.



LOGISTIC REGRESSION IN 2D / 2



SUMMARY

Hypothesis Space:

$$\mathcal{H} = \left\{ \pi : \mathcal{X} \rightarrow [0, 1] \mid \pi(\mathbf{x}) = s(\boldsymbol{\theta}^\top \mathbf{x}) \right\}$$

Risk: Logistic/Bernoulli loss function.

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x}))$$

Optimization: Numerical optimization, typically gradient-based methods.

