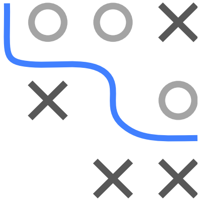
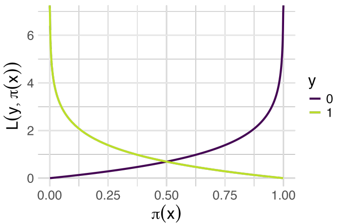
Introduction to Machine Learning



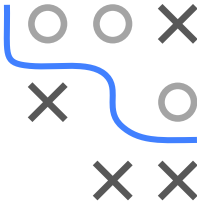
Advanced Risk Minimization
Bernoulli Loss



Learning goals

- Bernoulli (log, logistic, binomial, cross-entropy) loss
- Risk minimizer
- Optimal constant
- Complete separation problem

ON PROBABILITIES



- Likelihood of Bernoulli RV:

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^{n} \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)^{y^{(i)}} (1 - \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right))^{1-y^{(i)}} \qquad y \in \{0, 1\}$$

- Transform into NLL:

$$-\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} -y^{(i)} \log(\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)) - (1 - y^{(i)}) \log(1 - \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right))$$

- Bernoulli loss: loss on single sample

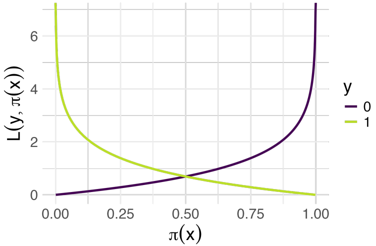
$$L\left(y, \pi(\mathbf{x})\right) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})) \qquad y \in \{0, 1\}$$

ON PROBABILITIES

- Bernoulli loss

$$L(y, \pi(\mathbf{x})) = -y \log(\pi(\mathbf{x})) - (1 - y) \log(1 - \pi(\mathbf{x})) \qquad y \in \{0, 1\}$$

- Confidently wrong predictions are harshly penalized



- A.k.a. Binomial, log, or cross-entropy loss
- Can also write for $y \in \{-1, +1\}$

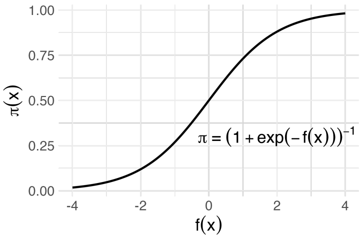
$$L(y, \pi(\mathbf{x})) = -\frac{1 + y}{2} \log(\pi(\mathbf{x})) - \frac{1 - y}{2} \log(1 - \pi(\mathbf{x})) \qquad y \in \{-1, +1\}$$

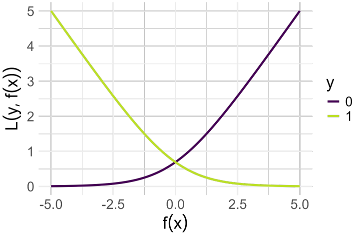
ON DECISION SCORES

- Transform probs into scores (log-odds): $f(\mathbf{x}) = \log(\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})})$
- Then $\pi(\mathbf{x}) = (1 + \exp(-f(\mathbf{x})))^{-1}$
- Yields equivalent loss formulation

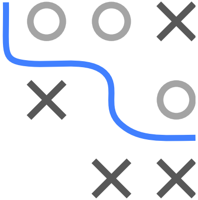
$$L(y, f(\mathbf{x})) = -y \cdot f(\mathbf{x}) + \log(1 + \exp(f(\mathbf{x}))) \quad \text{for } y \in \{0, 1\}$$

- For these and other simple derivations, see deep dive





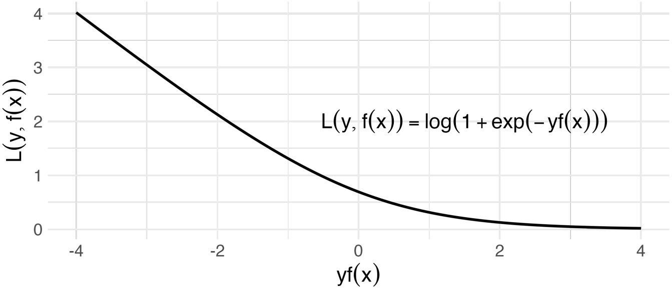
LOSS IN TERMS OF MARGIN



- For $y \in \{-1, +1\}$, loss becomes:

$$L(y, f(\mathbf{x})) = \log(1 + \exp(-y \cdot f(\mathbf{x})))$$

- All loss variants convex, differentiable



RISK MINIMIZER ON PROBS

- For probs and $y \in \{0, 1\}$, the risk minimizer is

$$\pi^*(\tilde{\mathbf{x}}) = \eta(\tilde{\mathbf{x}}) = \mathbb{P}(y = 1 \mid \mathbf{x} = \tilde{\mathbf{x}})$$

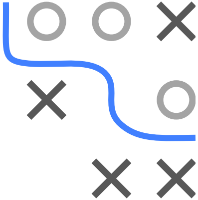
Proof: We have seen before

$$\mathcal{R}(f) = \mathbb{E}_x \left[L(1, \pi(\mathbf{x})) \cdot \eta(\mathbf{x}) + L(0, \pi(\mathbf{x})) \cdot (1 - \eta(\mathbf{x}))\right]$$

For fixed $\mathbf{x}$, minimize inner part pointwise, use $c \in (0, 1)$ for best value:

$$\begin{aligned}
\frac{d}{dc}(-\log c \cdot \eta(\mathbf{x}) - \log(1 - c) \cdot (1 - \eta(\mathbf{x}))) &= 0 \\
-\frac{\eta(\mathbf{x})}{c} + \frac{1 - \eta(\mathbf{x})}{1 - c} &= 0 \\
\frac{-\eta(\mathbf{x}) + \eta(\mathbf{x})c + c - \eta(\mathbf{x})c}{c(1 - c)} &= 0 \\
c &= \eta(\mathbf{x})
\end{aligned}$$

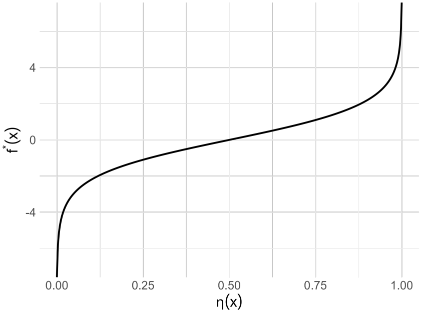
RISK MINIMIZER ON SCORES



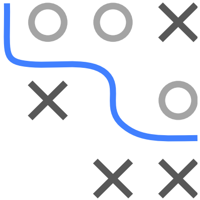
- For $y \in \{-1, 1\}$ and scores $f(\mathbf{x})$: RM is pointwise log-odds

$$f^*(\mathbf{x}) = \log(\frac{\eta(\mathbf{x})}{1 - \eta(\mathbf{x})})$$

- Undefined for $\eta(\mathbf{x}) \in \{0, 1\}$
- Monotonously increasing in $\eta(\mathbf{x})$, with $f^*(\mathbf{x}) = 0$ if $\eta(\mathbf{x}) = 0.5$



EMPIRICAL OPTIMAL CONSTANT MODELS



- Optimal constant probability model for labels $\mathcal{Y} = \{0, 1\}$ is

$$\hat{\theta} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} y^{(i)}$$

- Fraction of class-1 observations in observed data

- Optimal constant score model:

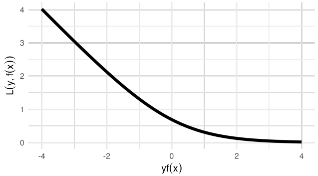
$$\hat{\theta} = \arg\min_{\theta} \mathcal{R}_{\text{emp}}(\boldsymbol{\theta}) = \log \frac{n_+}{n_-} = \log \frac{n_+/n}{n_-/n}$$

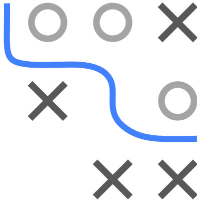
n_- and n_+ are nr. of neg. and pos. observations

- Again shows connection to log-odds

OPTIMIZATION PROPERTIES: CONVERGENCE

- In case of **complete separation**, optimization might fail





- Loss strictly decreasing in margin $yf(\mathbf{x})$:

$$L(yf(\mathbf{x})) = \log(1 + \exp(-yf(\mathbf{x})))$$

- f linear in $\boldsymbol{\theta}$, e.g., **log. regr.** with $f(\mathbf{x} \mid \boldsymbol{\theta}) = \boldsymbol{\theta}^\top \mathbf{x}$

- Assume data separable, so we can find $\boldsymbol{\theta}$:

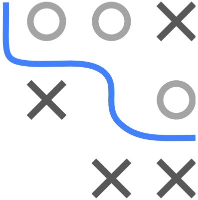
$$y^{(i)} f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = y^{(i)} \boldsymbol{\theta}^T \mathbf{x}^{(i)} > 0 \ \ \forall \mathbf{x}^{(i)}$$

- Can now construct a strictly better $\boldsymbol{\theta}$

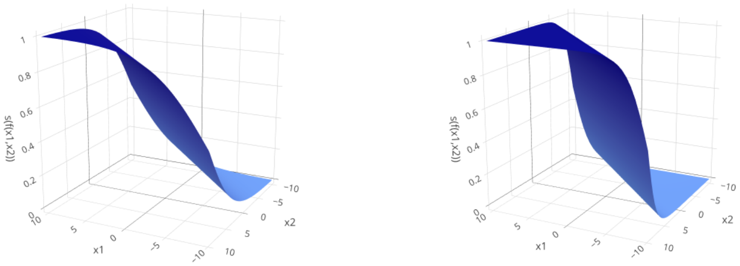
$$\mathcal{R}_{\text{emp}}(2 \cdot \boldsymbol{\theta}) = \sum_{i=1}^{n} L(2y^{(i)} \boldsymbol{\theta}^T \mathbf{x}^{(i)}) < \mathcal{R}_{\text{emp}}(\boldsymbol{\theta})$$

- As $||\boldsymbol{\theta}||$ increases, sum strictly decreases, as argument of L is strictly larger
- Loss is bounded from below, but no global optimium, cannot converge

OPTIMIZATION PROPERTIES: CONVERGENCE



- Geometrically, this translates to an ever steeper slope of the logistic/softmax function, i.e., increasingly sharp discrimination:



- In practice, data are rarely linearly separable and misclassified examples act as counterweights to increasing parameter values

- Can also use **regularization** for robust solutions