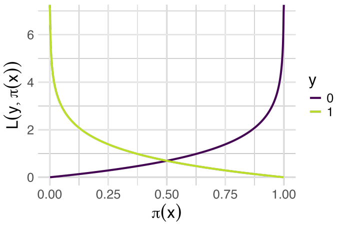
Introduction to Machine Learning

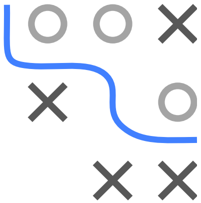
Advanced Risk Minimization
Logistic regression (Deep-Dive)



Learning goals

- Derive the gradient of the logistic regression
- Derive the Hessian of the logistic regression
- Show that the logistic regression is a convex problem

LOGISTIC REGRESSION: RISK PROBLEM



Given n observations $(\mathbf{x}^{(i)}, y^{(i)}) \in \mathcal{X} \times \mathcal{Y}$ with $\mathcal{X} = \mathbb{R}^d, \mathcal{Y} = \{0, 1\}$ we want to minimize the risk

$$\mathcal{R}_{\text{emp}} = -\sum_{i=1}^{n} y^{(i)} \log(\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)) + (1 - y^{(i)}) \log(1 - \pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)))$$

with respect to $\boldsymbol{\theta}$ where the probabilistic classifier

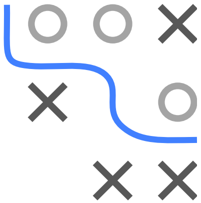
$$\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = s(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))$$

the sigmoid function $s(f) = \frac{1}{1+\exp(-f)}$ and the score $f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) = \boldsymbol{\theta}^\top \mathbf{x}$

NB: Note that $\frac{\partial}{\partial f} s(f) = s(f)(1 - s(f))$ and $\frac{\partial f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = (\mathbf{x}^{(i)})^\top$

From now on we abbreviate $\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)$ as $\pi_{\boldsymbol{\theta}}^{(i)}$

LOGISTIC REGRESSION: GRADIENT



We find the gradient of logistic regression with the chain rule:

$$
\begin{aligned}
\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} &= -\sum_{i=1}^{n} \frac{\partial}{\partial \pi_{\boldsymbol{\theta}}^{(i)}} y^{(i)} \log(\pi_{\boldsymbol{\theta}}^{(i)}) \frac{\partial \pi_{\boldsymbol{\theta}}^{(i)}}{\partial \boldsymbol{\theta}} + \\
&\qquad \frac{\partial}{\partial \pi_{\boldsymbol{\theta}}^{(i)}} (1 - y^{(i)}) \log(1 - \pi_{\boldsymbol{\theta}}^{(i)}) \frac{\partial \pi_{\boldsymbol{\theta}}^{(i)}}{\partial \boldsymbol{\theta}} \\
&= -\sum_{i=1}^{n} \frac{y^{(i)}}{\pi_{\boldsymbol{\theta}}^{(i)}} \frac{\partial \pi_{\boldsymbol{\theta}}^{(i)}}{\partial \boldsymbol{\theta}} - \frac{1 - y^{(i)}}{1 - \pi_{\boldsymbol{\theta}}^{(i)}} \frac{\partial \pi_{\boldsymbol{\theta}}^{(i)}}{\partial \boldsymbol{\theta}} \\
&= -\sum_{i=1}^{n} \left(\frac{y^{(i)}}{\pi_{\boldsymbol{\theta}}^{(i)}} - \frac{1 - y^{(i)}}{1 - \pi_{\boldsymbol{\theta}}^{(i)}} \right) \frac{\partial s(f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}))}{\partial f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta})} \frac{\partial f(\mathbf{x}^{(i)} \mid \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \\
&= -\sum_{i=1}^{n} \left(y^{(i)} (1 - \pi_{\boldsymbol{\theta}}^{(i)}) - (1 - y^{(i)}) \pi_{\boldsymbol{\theta}}^{(i)} \right) (\mathbf{x}^{(i)})^{\top}
\end{aligned}
$$

LOGISTIC REGRESSION: GRADIENT

$$= \sum_{i=1}^{n} (\pi_{\boldsymbol{\theta}}^{(i)} - y^{(i)})(\mathbf{x}^{(i)})^\top$$

$$= (\pi(\mathbf{X} |\ \boldsymbol{\theta}) - \mathbf{y})^\top \mathbf{X}$$

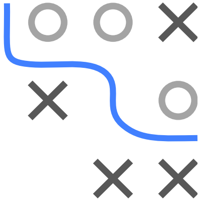
where

- $\mathbf{X} = (\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)})^\top \in \mathbb{R}^{n \times d}$
- $\mathbf{y} = (y^{(1)}, \ldots, y^{(n)})^\top$
- $\pi(\mathbf{X} |\ \boldsymbol{\theta}) = (\pi_{\boldsymbol{\theta}}^{(i)}[1], \ldots, \pi_{\boldsymbol{\theta}}^{(i)}[n])^\top \in \mathbb{R}^n$

$\implies$ The gradient $\nabla_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = (\frac{\partial}{\partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}})^\top = \mathbf{X}^\top (\pi(\mathbf{X} |\ \boldsymbol{\theta}) - \mathbf{y})$

This formula can now be used in gradient descent and its friends

LOGISTIC REGRESSION: HESSIAN



We find the Hessian via differentiation:

$$
\begin{aligned}
\nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \frac{\partial^2}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} &= \frac{\partial}{\partial \boldsymbol{\theta}^\top} \sum_{i=1}^{n} (\pi_{\boldsymbol{\theta}}^{(i)} - y^{(i)})(\mathbf{x}^{(i)})^\top \\
&= \sum_{i=1}^{n} \mathbf{x}^{(i)} (\pi_{\boldsymbol{\theta}}^{(i)} (1 - \pi_{\boldsymbol{\theta}}^{(i)}))(\mathbf{x}^{(i)})^\top \\
&= \mathbf{X}^\top \mathbf{D} \mathbf{X}
\end{aligned}
$$

where $\mathbf{D} \in \mathbb{R}^{n \times n}$ is a diagonal matrix with diagonal

$$
\left(\pi_{\boldsymbol{\theta}}^{(i)}[1](1 - \pi_{\boldsymbol{\theta}}^{(i)}[1], \ldots, \pi_{\boldsymbol{\theta}}^{(i)}[n](1 - \pi_{\boldsymbol{\theta}}^{(i)}[n] \right)
$$

Can now be used in Newton-Raphson and other 2nd order optimizers

LOGISTIC REGRESSION: CONVEXITY

Finally, we check that logistic regression is a convex problem:

We define the diagonal matrix $\bar{\mathbf{D}} \in \mathbb{R}^{n \times n}$ with diagonal

$$\left(\sqrt{\pi_{\boldsymbol{\theta}}^{(i)}[1])(1 - \pi_{\boldsymbol{\theta}}^{(i)}[1]}, \ldots, \sqrt{\pi_{\boldsymbol{\theta}}^{(i)}[n](1 - \pi_{\boldsymbol{\theta}}^{(i)}[n]}\right)$$

which is possible since π maps into (0, 1)

With this, we get for any $\mathbf{w} \in \mathbb{R}^d$ that

$$\mathbf{w}^\top \nabla_{\boldsymbol{\theta}}^2 \mathcal{R}_{\text{emp}} \mathbf{w} = \mathbf{w}^\top \mathbf{X}^\top \bar{\mathbf{D}}^\top \bar{\mathbf{D}} \mathbf{X} \mathbf{w} = (\bar{\mathbf{D}} \mathbf{X} \mathbf{w})^\top \bar{\mathbf{D}} \mathbf{X} \mathbf{w} = \|\bar{\mathbf{D}} \mathbf{X} \mathbf{w}\|_2^2 \geq 0$$

since obviously $\mathbf{D} = \bar{\mathbf{D}}^\top \bar{\mathbf{D}}$

$\implies \nabla_{\boldsymbol{\theta}}^2 \mathcal{R}_{\text{emp}}$ is positive semi-definite $\implies \mathcal{R}_{\text{emp}}$ is convex