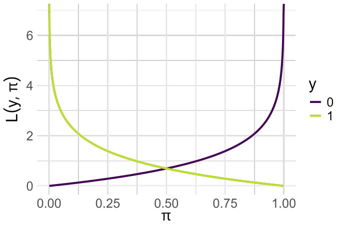
Introduction to Machine Learning

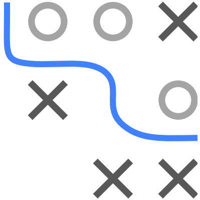
Advanced Risk Minimization
Logistic regression (Deep-Dive)



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

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

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

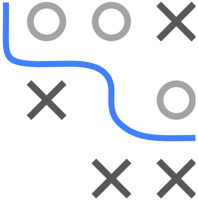
$$\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) \quad = \quad s\left(f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)\right),$$

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 $\frac{\partial f\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right)}{\partial \boldsymbol{\theta}} = \left(\mathbf{x}^{(i)}\right)^\top$.

LOGISTIC REGRESSION: GRADIENT

We find the gradient of logistic regression with the chain rule, s.t.,



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

LOGISTIC REGRESSION: GRADIENT / 2

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left(\pi\left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}\right) - y^{(i)}\right)\left(\mathbf{x}^{(i)}\right)^{\top} \\
&= \left(\pi(\mathbf{X} \mid \boldsymbol{\theta}) - \mathbf{y}\right)^{\top} \mathbf{X}
\end{aligned}
$$

where $\mathbf{X} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)^{\top} \in \mathbb{R}^{n \times d}, \mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)^{\top},$
$\pi(\mathbf{X} \mid \boldsymbol{\theta}) = \left(\pi\left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta}\right), \ldots, \pi\left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta}\right)\right)^{\top} \in \mathbb{R}^{n}.$

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

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

LOGISTIC REGRESSION: HESSIAN

We find the Hessian via differentiation, s.t.,

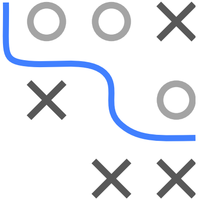
$$
\begin{aligned}
\nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}} = \frac{\partial^2}{\partial \boldsymbol{\theta}^\top \partial \boldsymbol{\theta}} \mathcal{R}_{\text{emp}} &= \frac{\partial}{\partial \boldsymbol{\theta}^\top} \sum_{i=1}^{n} \left(\pi \left(\mathbf{x}^{(i)} \mid \boldsymbol{\theta}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

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

$$
\left(\pi \left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta} \right) (1 - \pi \left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta} \right), \ldots, \pi \left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta} \right) (1 - \pi \left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta} \right) \right).
$$

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

$$\left(\sqrt{\pi\left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta}\right)\left(1-\pi\left(\mathbf{x}^{(1)} \mid \boldsymbol{\theta}\right)\right)}, \ldots, \sqrt{\pi\left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta}\right)\left(1-\pi\left(\mathbf{x}^{(n)} \mid \boldsymbol{\theta}\right)\right)}\right)$$

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

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

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

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

$\Rightarrow \nabla^2_{\boldsymbol{\theta}} \mathcal{R}_{\text{emp}}$ is positive semi-definite $\Rightarrow \mathcal{R}_{\text{emp}}$ is convex.