Logistic regression

Topics we'll cover

1. The logistic regression model
2. Loss function: properties
3. Solution by gradient descent

Logistic regression for binary labels

- Data $x \in \mathbb{R}^d$ and binary labels $y \in \{-1, 1\}$
- Model parametrized by $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$:

$$\Pr_{w,b}(y|x) = \frac{1}{1 + e^{-y(w \cdot x + b)}}$$

The learning problem

Maximum-likelihood principle: given data $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, pick $w \in \mathbb{R}^d$ and $b \in \mathbb{R}$ that maximize

$$\prod_{i=1}^{n} \Pr_{w,b}(y^{(i)} \mid x^{(i)})$$

Take log to get **loss function**

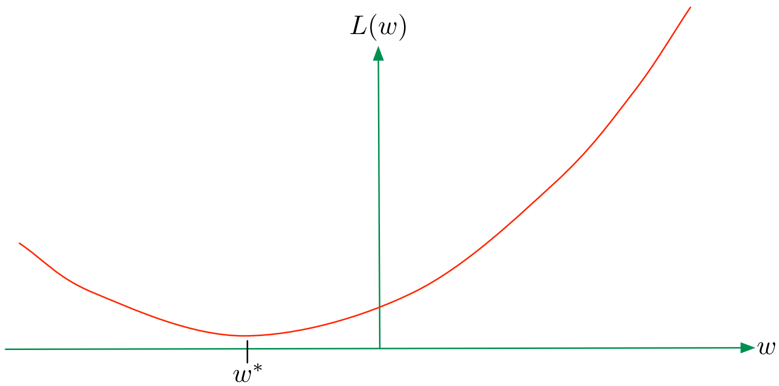
$$L(w, b) = -\sum_{i=1}^{n} \ln \Pr_{w,b}(y^{(i)} \mid x^{(i)}) = \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)} + b)})$$

Goal: minimize $L(w, b)$.

As with linear regression, can absorb b into w.
Yields simplified loss function $L(w)$.

Convexity

- Bad news: no closed-form solution for w
- Good news: $L(w)$ is **convex** in w



How to find the minimum of a convex function? By **local search.**

Gradient descent procedure for logistic regression

Given $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, 1\}$, find

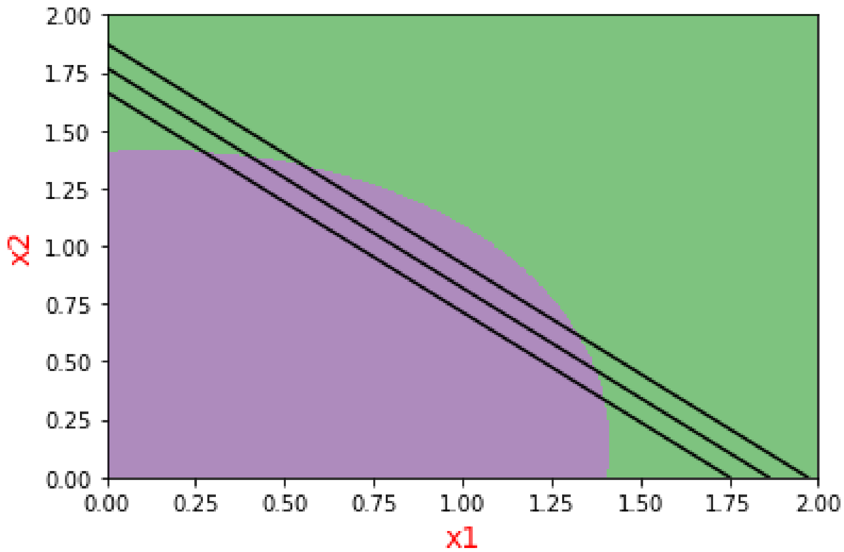
$$\arg\min_{w \in \mathbb{R}^d} L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} \;=\; w_t \;+\; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \underbrace{\Pr_{w_t}(-y^{(i)}|x^{(i)})}_{\text{doubt}_t(x^{(i)}, y^{(i)})},$$

where η_t is a "step size"

Toy example



2.00
1.75
1.50
1.25
1.00
0.75
0.50
0.25
0.00
x2
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
x1