Workshop 4

COMP90051 Machine Learning

Semester 2, 2020

Learning outcomes

At the end of this workshop you should:

- be able to explain how the optimisation problems for linear regression and logistic regression differ
- be able to implement logistic regression using the iteratively reweighted least-squares (IRLS) algorithm and gradient descent
- be able to explain benefits/drawbacks of IRLS versus gradient descent

Solving logistic regression

Logistic regression optimisation problem:

$$\mathbf{w}^\star \in \arg\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \mu_i)$$

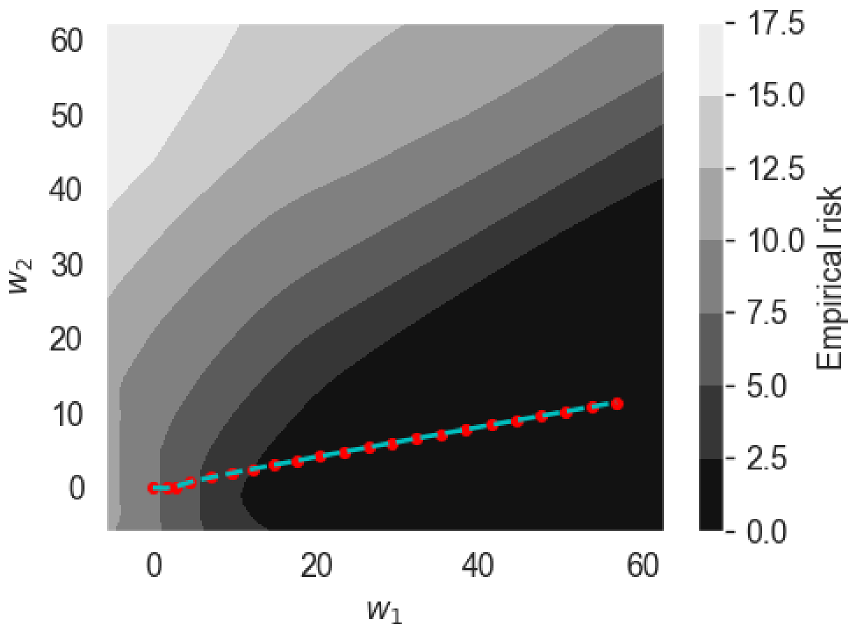
where $\mu_i = \frac{1}{1+\mathrm{e}^{-\mathbf{x}_i^\top \mathbf{w}}}$ and $\ell(y, \mu) = -y \log \mu - (1-y)\log(1-\mu)$

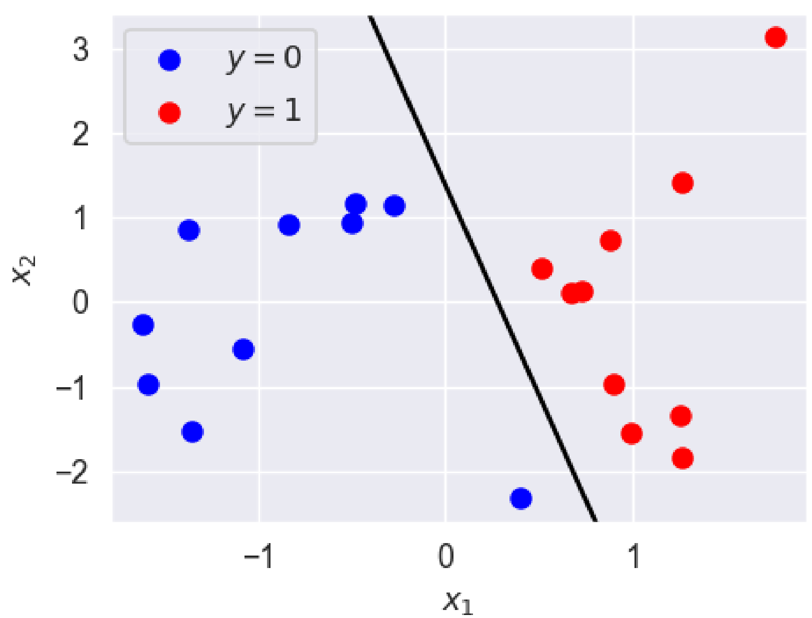
- Can apply gradient descent, but it's slow to converge
- Iteratively reweighted least-squares (IRLS) is faster option:
  - Equivalent to Newton's method (uses the second-order derivative)
  - End up solving a sequence of weighted linear regression problems—interesting connection to last week's workshop!

Worksheet 4

Linearly separable case

- When the data is linearly separable, the optimal weight vector satisfies $\|\mathbf{w}^{\star}\|_{\infty} \to \infty$
- $p(y|\mathbf{x}) = \mathrm{sigmoid}(\mathbf{x}^{\top}\mathbf{w}^{\star})$ transitions abruptly from 0 to 1 at the decision boundary (like a step function)





IRLS in the separable case

A key step in the IRLS algorithm is computing the linearised response:

$$\mathbf{b}_t = \mathbf{X}\mathbf{w}_t + \underbrace{\mathbf{M}_t^{-1}(\mathbf{y} - \boldsymbol{\mu}_t)}_{\mathbf{v}_t}$$

where $\mathbf{M}_t = \mathrm{diag}\big(\boldsymbol{\mu}_t(1 - \boldsymbol{\mu}_t)\big)$ and $\mu_{i,t} = \mathrm{sigmoid}(\mathbf{x}_i^\top \mathbf{w}_t)$

However, when the data is linearly separable

$$\mu_{i,t} \to \begin{cases} 0, & \text{if } y_i = 0 \\ 1, & \text{if } y_i = 1 \end{cases}$$

as $t \to \infty$.

Why is this problematic numerically? We can't compute $\mathbf{M}_t^{-1}$

IRLS in the separable case

To avoid division by zero when computing $\mathbf{M}_t^{-1}$, we compute the second term $\mathbf{v}_t$ as:

$$v_{i,t} = \begin{cases} \dfrac{1-\mu_{i,t}}{\mu_{i,t}}, & y_i = 1 \\ -\dfrac{1}{1-\mu_{i,t}}, & y_i = 0 \end{cases}$$

Optional exercise: verify that this works and implement in worksheet 4