

Bayesian Logistic Regression

Zeel B Patel, Nipun Batra

August 27, 2023

IIT Gandhinagar

MLE

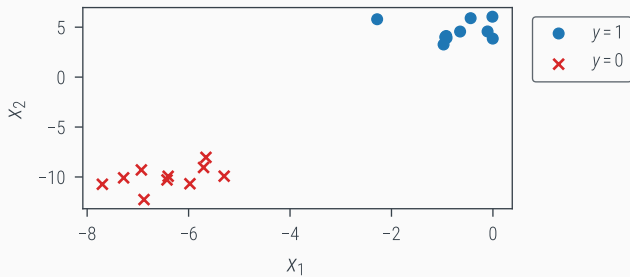
MAP

Fully Bayesian

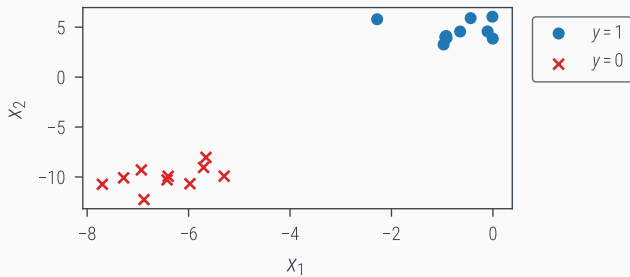
Laplace Approximation

MLE

Data

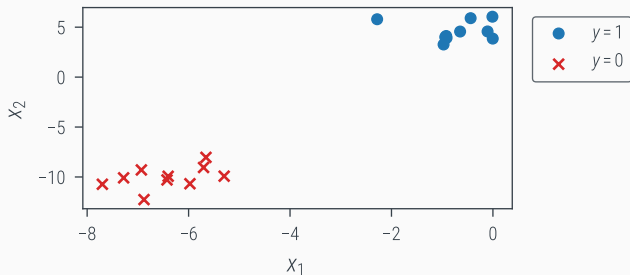


Data



| x1 | x2 | y |
|-------|--------|-----|
| -5.97 | -10.68 | 0 |
| -0.44 | 5.90 | 1 |
| -0.97 | 3.27 | 1 |
| ... | ... | ... |

Data



| x1 | x2 | y |
|-------|--------|-----|
| -5.97 | -10.68 | 0 |
| -0.44 | 5.90 | 1 |
| -0.97 | 3.27 | 1 |
| ... | ... | ... |

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^N$$
$$= \{X, \mathbf{y}\}$$

$$p(\mathcal{D}|\boldsymbol{\theta}) = p(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \boldsymbol{\theta})$$

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= p(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^N \text{Bernoulli} \left(\sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) \right) \quad \left[\sigma(x) = \frac{1}{1 + e^{-x}} \right] \end{aligned}$$

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= p(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^N \text{Bernoulli}\left(\sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)\right) \quad \left[\sigma(x) = \frac{1}{1 + e^{-x}}\right] \\ &= \prod_{i=1}^N \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)^{y_i} \left(1 - \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)\right)^{1-y_i} \end{aligned}$$

$$\begin{aligned} p(\mathcal{D}|\boldsymbol{\theta}) &= p(\mathbf{y}|X, \boldsymbol{\theta}) = \prod_{i=1}^N p(y_i|\mathbf{x}_i, \boldsymbol{\theta}) \\ &= \prod_{i=1}^N \text{Bernoulli}\left(\sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)\right) \quad \left[\sigma(x) = \frac{1}{1 + e^{-x}}\right] \\ &= \prod_{i=1}^N \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)^{y_i} \left(1 - \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)\right)^{1-y_i} \\ \log p(\mathbf{y}|X, \boldsymbol{\theta}) &= \sum_{i=1}^N \left[y_i \log \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right) + (1 - y_i) \log \left(1 - \sigma\left(\boldsymbol{\theta}^T \mathbf{x}_i\right)\right) \right] \end{aligned}$$

$$-\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{i=1}^N \left[y_i \log \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) - (1 - y_i) \log \left(1 - \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) \right) \right]$$

$$-\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{i=1}^N \left[y_i \log \sigma(\boldsymbol{\theta}^T \mathbf{x}_i) - (1 - y_i) \log (1 - \sigma(\boldsymbol{\theta}^T \mathbf{x}_i)) \right]$$

$$\frac{\partial}{\partial \boldsymbol{\theta}} -\log p(\mathcal{D}|\boldsymbol{\theta}) = -\sum_{i=1}^N \left[y_i \frac{\sigma(\boldsymbol{\theta}^T \mathbf{x}_i) (1 - \sigma(\boldsymbol{\theta}^T \mathbf{x}_i))}{\sigma(\boldsymbol{\theta}^T \mathbf{x}_i)} \mathbf{x}_i \right]$$

$$-\log p(\mathcal{D}|\theta) = -\sum_{i=1}^N \left[y_i \log \sigma(\theta^T \mathbf{x}_i) - (1 - y_i) \log (1 - \sigma(\theta^T \mathbf{x}_i)) \right]$$

$$\frac{\partial}{\partial \theta} -\log p(\mathcal{D}|\theta) = -\sum_{i=1}^N \left[y_i \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{\sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right. \\ \left. - (1 - y_i) \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{1 - \sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right]$$

]

$$\begin{aligned} -\log p(\mathcal{D}|\theta) &= -\sum_{i=1}^N \left[y_i \log \sigma(\theta^T \mathbf{x}_i) - (1 - y_i) \log (1 - \sigma(\theta^T \mathbf{x}_i)) \right] \\ \frac{\partial}{\partial \theta} -\log p(\mathcal{D}|\theta) &= -\sum_{i=1}^N \left[y_i \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{\sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right. \\ &\quad \left. - (1 - y_i) \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{1 - \sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right] \\ &= -\sum_{i=1}^N \left[y_i (1 - \sigma(\theta^T \mathbf{x}_i)) \mathbf{x}_i - (1 - y_i) \sigma(\theta^T \mathbf{x}_i) \mathbf{x}_i \right] \end{aligned}$$

$$\begin{aligned} -\log p(\mathcal{D}|\theta) &= -\sum_{i=1}^N \left[y_i \log \sigma(\theta^T \mathbf{x}_i) - (1 - y_i) \log (1 - \sigma(\theta^T \mathbf{x}_i)) \right] \\ \frac{\partial}{\partial \theta} -\log p(\mathcal{D}|\theta) &= -\sum_{i=1}^N \left[y_i \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{\sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right. \\ &\quad \left. - (1 - y_i) \frac{\sigma(\theta^T \mathbf{x}_i) (1 - \sigma(\theta^T \mathbf{x}_i))}{1 - \sigma(\theta^T \mathbf{x}_i)} \mathbf{x}_i \right] \\ &= -\sum_{i=1}^N \left[y_i (1 - \sigma(\theta^T \mathbf{x}_i)) \mathbf{x}_i - (1 - y_i) \sigma(\theta^T \mathbf{x}_i) \mathbf{x}_i \right] \\ &= -\sum_{i=1}^N (y_i - \sigma(\theta^T \mathbf{x}_i)) \mathbf{x}_i \end{aligned}$$

$$X^T(\sigma(X\theta) - \mathbf{y}) = 0$$

$$\begin{aligned}X^T(\sigma(X\theta) - \mathbf{y}) &= 0 \\X^T\sigma(X\theta) &= X^T\mathbf{y}\end{aligned}$$

$$X^T(\sigma(X\theta) - \mathbf{y}) = 0$$

$$X^T \sigma(X\theta) = X^T \mathbf{y}$$

$$\frac{1}{1 + e^{-X\theta}} = \mathbf{y}$$

$$X^T(\sigma(X\theta) - \mathbf{y}) = 0$$

$$X^T \sigma(X\theta) = X^T \mathbf{y}$$

$$\frac{1}{1 + e^{-X\theta}} = \mathbf{y}$$

$$X\theta = \log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$$

$$X^T(\sigma(X\theta) - \mathbf{y}) = 0$$

$$X^T \sigma(X\theta) = X^T \mathbf{y}$$

$$\frac{1}{1 + e^{-X\theta}} = \mathbf{y}$$

$$X\theta = \log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$$

$$\theta_{MLE} = (X^T X)^{-1} X^T \log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$$

$$X^T(\sigma(X\theta) - \mathbf{y}) = 0$$

$$X^T \sigma(X\theta) = X^T \mathbf{y}$$

$$\frac{1}{1 + e^{-X\theta}} = \mathbf{y}$$

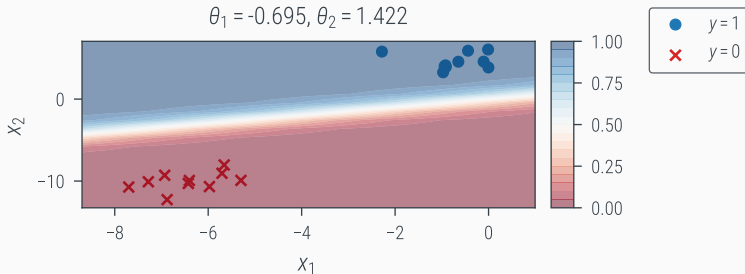
$$X\theta = \log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$$

$$\theta_{MLE} = (X^T X)^{-1} X^T \log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$$

However, $\log\left(\frac{\mathbf{y}}{1 - \mathbf{y}}\right)$ is undefined when $y_i = 0$ or $y_i = 1$, which is always the case.

- There is no closed form solution for θ_{MLE} . So, we have to use gradient descent.

- There is no closed form solution for θ_{MLE} . So, we have to use gradient descent.



MAP

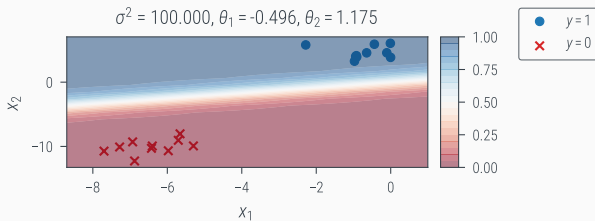
- We may use a Gaussian prior on θ .

$$p(\theta) = \mathcal{N}(\theta|0, \sigma^2)$$

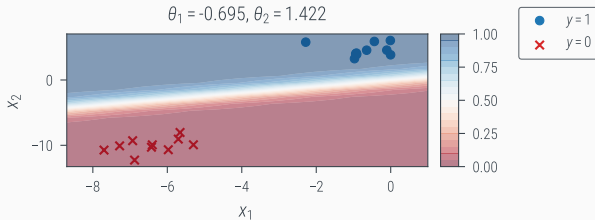
$$\begin{aligned}\log p(\boldsymbol{\theta}, \mathbf{y} | X) &= \log p(\mathbf{y} | X, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \\&= \sum_{i=1}^N \left[y_i \log \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) + (1 - y_i) \log \left(1 - \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) \right) \right] \\&\quad - \frac{1}{2} \frac{\boldsymbol{\theta}^T \boldsymbol{\theta}}{\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \\&= \sum_{i=1}^N \left[y_i \log \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) + (1 - y_i) \log \left(1 - \sigma \left(\boldsymbol{\theta}^T \mathbf{x}_i \right) \right) \right] \\&\quad - \left(c_1 \boldsymbol{\theta}^T \boldsymbol{\theta} + c_2 \right) \quad [c_1 \geq 0]\end{aligned}$$

MAP with a weak prior

MAP

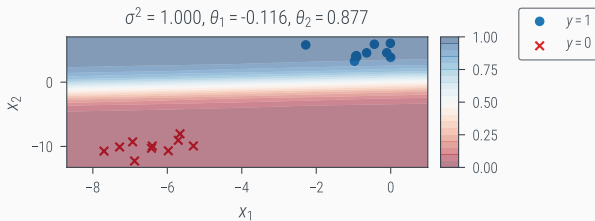


MLE

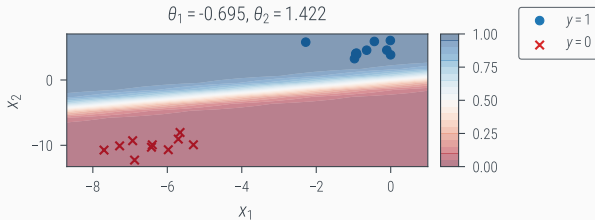


MAP with a medium prior

MAP

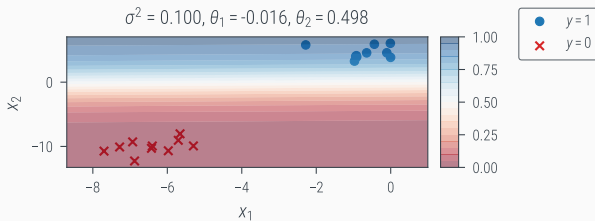


MLE

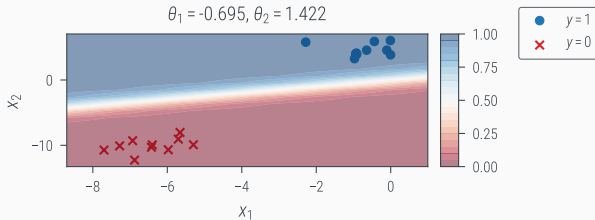


MAP with a strong prior

MAP



MLE



Fully Bayesian

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

- Normal prior and Bernoulli likelihood do not form a conjugate pair. Thus, the denominator is intractable and we cannot find the posterior in closed form.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

- Normal prior and Bernoulli likelihood do not form a conjugate pair. Thus, the denominator is intractable and we cannot find the posterior in closed form.
- We need another method to find the posterior.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int p(\mathcal{D}|\theta)p(\theta)d\theta}$$

- Normal prior and Bernoulli likelihood do not form a conjugate pair. Thus, the denominator is intractable and we cannot find the posterior in closed form.
- We need another method to find the posterior.
- Laplace approximation!

Laplace Approximation

$$\begin{aligned}\text{Neg. Log Joint } f(\theta) &= -\log p(\mathcal{D}|\theta) - \log p(\theta) \\ &= -\sum_{i=1}^N \left[y_i \log \sigma(\theta^T \mathbf{x}_i) \right. \\ &\quad \left. + (1 - y_i) \log (1 - \sigma(\theta^T \mathbf{x}_i)) \right] \\ &\quad - \left(-\frac{1}{2} \frac{\theta^T \theta}{\sigma^2} - \frac{1}{2} \log(2\pi\sigma^2) \right)\end{aligned}$$

$$\text{Laplace Posterior } q(\theta) = \mathcal{N}(\theta | \theta_{\text{MAP}}, \nabla^2 f(\theta_{\text{MAP}})^{-1})$$

$$p(y^* = 1 | \mathbf{x}^*, \mathcal{D}) = \int p(y^* = 1 | \mathbf{x}^*, \theta) p(\theta | \mathcal{D}) d\theta$$

$$\begin{aligned} p(y^* = 1 | \mathbf{x}^*, \mathcal{D}) &= \int p(y^* = 1 | \mathbf{x}^*, \theta) p(\theta | \mathcal{D}) d\theta \\ &\approx \int p(y^* = 1 | \mathbf{x}^*, \theta) q(\theta) d\theta \end{aligned}$$

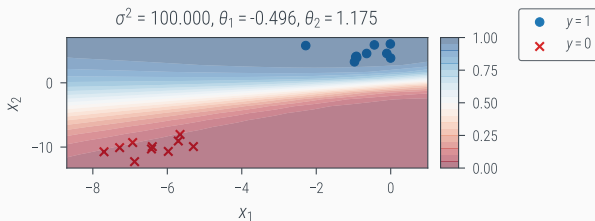
$$\begin{aligned} p(y^* = 1 | \mathbf{x}^*, \mathcal{D}) &= \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \\ &\approx \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \sigma(\boldsymbol{\theta}^T \mathbf{x}^*) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \end{aligned}$$

$$\begin{aligned} p(y^* = 1 | \mathbf{x}^*, \mathcal{D}) &= \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \\ &\approx \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \sigma(\boldsymbol{\theta}^T \mathbf{x}^*) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} \left(\sigma(\boldsymbol{\theta}^T \mathbf{x}^*) \right) \end{aligned}$$

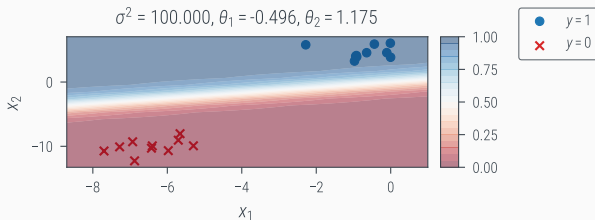
$$\begin{aligned} p(y^* = 1 | \mathbf{x}^*, \mathcal{D}) &= \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \\ &\approx \int p(y^* = 1 | \mathbf{x}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \sigma(\boldsymbol{\theta}^T \mathbf{x}^*) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \mathbb{E}_{q(\boldsymbol{\theta})} \left(\sigma(\boldsymbol{\theta}^T \mathbf{x}^*) \right) \\ &\approx \frac{1}{M} \sum_{i=1}^M \sigma(\boldsymbol{\theta}_i^T \mathbf{x}^*) \end{aligned}$$

Predictive Distribution with a Weak Prior

Laplace

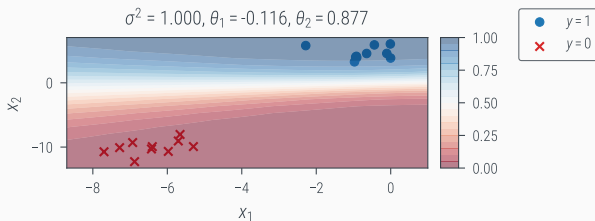


MAP

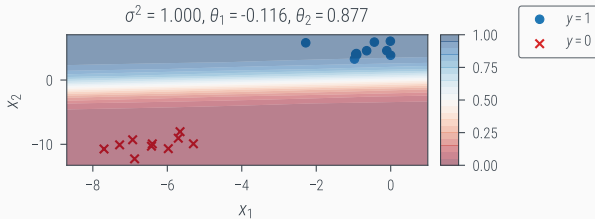


Predictive Distribution with a Medium Prior

Laplace

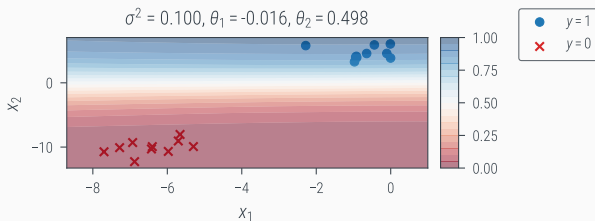


MAP



Predictive Distribution with a Strong Prior

Laplace



MAP

