Bayesian Logistic Regression

Zeel B Patel, Nipun Batra

August 27, 2023

IIT Gandhinagar

Outline

MLE

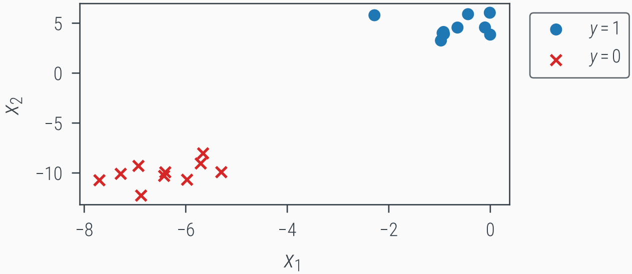
MAP

Fully Bayesian

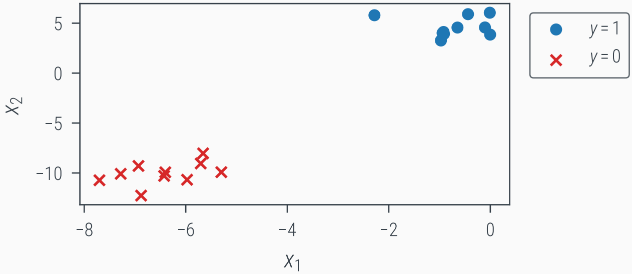
Laplace Approximation

MLE

Data

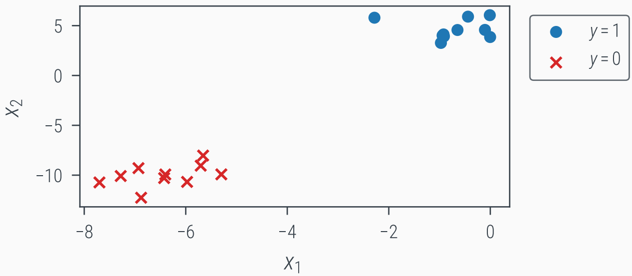


Data



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
...	...	...

$$\begin{aligned} \mathcal{D} &= \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{N} \\ &= \{X, \boldsymbol{y}\} \end{aligned}$$

Likelihood

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

Likelihood

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

Likelihood

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

Likelihood

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

MLE

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

MLE

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

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

MLE

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

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

MLE

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

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

MLE

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

MLE

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

MLE

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

MLE

$$
\begin{aligned}
X^T(\boldsymbol{\sigma}(X\boldsymbol{\theta}) - \boldsymbol{y}) &= 0 \\
X^T\boldsymbol{\sigma}(X\boldsymbol{\theta}) &= X^T\boldsymbol{y} \\
\frac{1}{1 + e^{-X\boldsymbol{\theta}}} &= \boldsymbol{y}
\end{aligned}
$$

MLE

$$
\begin{aligned}
X^T(\boldsymbol{\sigma}(X\boldsymbol{\theta}) - \boldsymbol{y}) &= 0 \\
X^T\boldsymbol{\sigma}(X\boldsymbol{\theta}) &= X^T\boldsymbol{y} \\
\frac{1}{1+e^{-X\boldsymbol{\theta}}} &= \boldsymbol{y} \\
X\boldsymbol{\theta} &= \log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right)
\end{aligned}
$$

MLE

$$
\begin{aligned}
X^T(\boldsymbol{\sigma}(X\boldsymbol{\theta}) - \boldsymbol{y}) &= 0 \\
X^T\boldsymbol{\sigma}(X\boldsymbol{\theta}) &= X^T\boldsymbol{y} \\
\frac{1}{1+e^{-X\boldsymbol{\theta}}} &= \boldsymbol{y} \\
X\boldsymbol{\theta} &= \log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right) \\
\boldsymbol{\theta}_{MLE} &= \left(X^TX\right)^{-1} X^T \log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right)
\end{aligned}
$$

MLE

$$
\begin{aligned}
X^T(\boldsymbol{\sigma}(X\boldsymbol{\theta}) - \boldsymbol{y}) &= 0 \\
X^T\boldsymbol{\sigma}(X\boldsymbol{\theta}) &= X^T\boldsymbol{y} \\
\frac{1}{1 + e^{-X\boldsymbol{\theta}}} &= \boldsymbol{y} \\
X\boldsymbol{\theta} &= \log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right) \\
\boldsymbol{\theta}_{MLE} &= \left(X^T X\right)^{-1} X^T \log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right)
\end{aligned}
$$

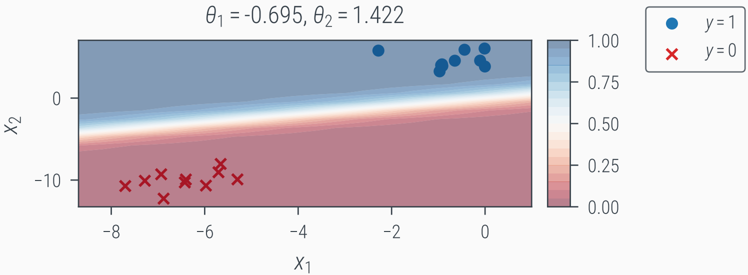
However, $\log\left(\frac{\boldsymbol{y}}{1-\boldsymbol{y}}\right)$ is undefined when $y_i = 0$ or $y_i = 1$, which is always the case.

MLE

- There is no closed form solution for $\boldsymbol{\theta}_{\text{MLE}}$. So, we have to use gradient descent.

MLE

- There is no closed form solution for $\boldsymbol{\theta}_{\text{MLE}}$. So, we have to use gradient descent.



MAP

Prior

- We may use a Gaussian prior on $\boldsymbol{\theta}$.

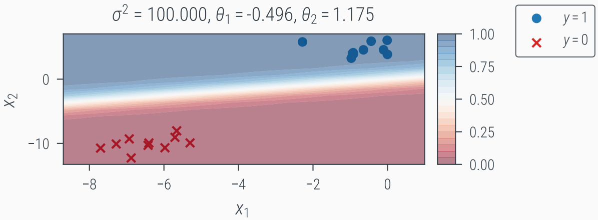
$$p(\boldsymbol{\theta}) = \mathcal{N}(\boldsymbol{\theta}|0, \sigma^2)$$

Log Joint

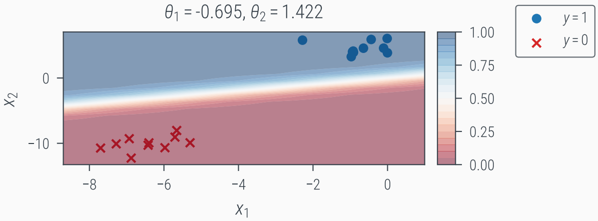
$$
\begin{aligned}
\log p(\boldsymbol{\theta}, \boldsymbol{y}|X) &= \log p(\boldsymbol{y}|X, \boldsymbol{\theta}) + \log p(\boldsymbol{\theta}) \\
&= \sum_{i=1}^{N} y_i \log \boldsymbol{\sigma}\left(\boldsymbol{\theta}^T \boldsymbol{x}_i\right) + (1 - y_i) \log\left(1 - \boldsymbol{\sigma}\left(\boldsymbol{\theta}^T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

MAP with a weak prior

MAP

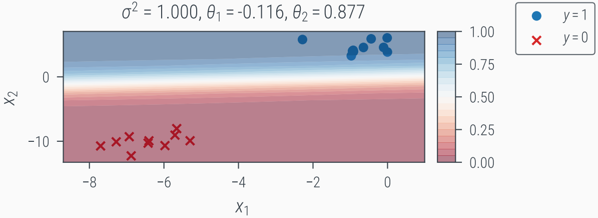


MLE

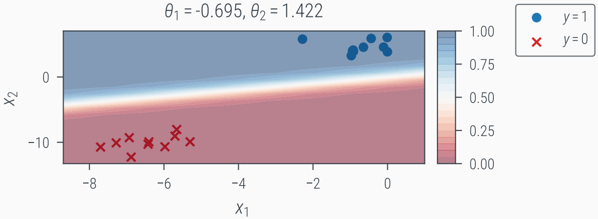


MAP with a medium prior

MAP

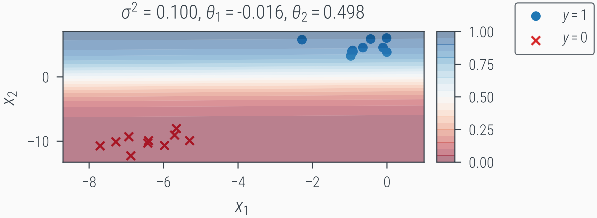


MLE

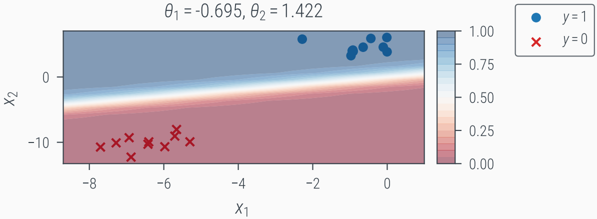


MAP with a strong prior

MAP



MLE



Fully Bayesian

Posterior

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

- 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.

Posterior

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

- 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.

Posterior

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

- 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

A Quick Refresher

$$
\begin{aligned}
\text{Neg. Log Joint } f(\theta) &= -\log p(\mathcal{D}|\theta) - \log p(\theta) \\
&= -\sum_{i=1}^{N} y_i \log \boldsymbol{\sigma}\left(\boldsymbol{\theta}^T \boldsymbol{x}_i\right) + (1-y_i)\log \Big(1 - \boldsymbol{\sigma}\Big(\boldsymbol{\theta}^T \boldsymbol{x}_i \\
&\quad - \left(-\frac{1}{2}\frac{\boldsymbol{\theta}^T\boldsymbol{\theta}}{\sigma^2} - \frac{1}{2}\log(2\pi\sigma^2)\right) \\
\text{Laplace Posterior } q(\boldsymbol{\theta}) &= \mathcal{N}(\theta|\theta_{\text{MAP}}, \nabla^2 f(\theta_{\text{MAP}})^{-1}) \\
&= \mathcal{N}(\theta|\theta_{\text{MAP}}, H^{-1})
\end{aligned}
$$

Predictive distribution

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

Predictive distribution

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

Predictive distribution

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

Predictive distribution

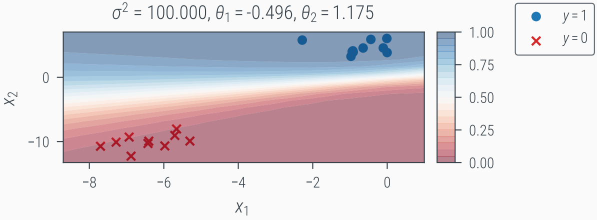
$$
\begin{aligned}
p(y^* = 1|\boldsymbol{x}^*, \mathcal{D}) &= \int p(y^* = 1|\boldsymbol{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathcal{D}) d\boldsymbol{\theta} \\
&\approx \int p(y^* = 1|\boldsymbol{x}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int \boldsymbol{\sigma}(\boldsymbol{\theta}^T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

Predictive distribution

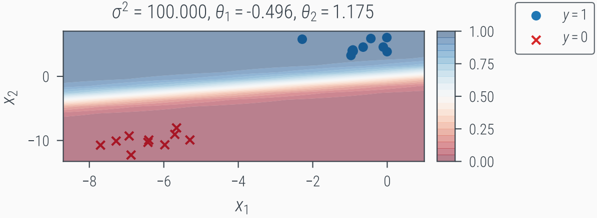
$$
\begin{aligned}
p(y^* = 1 | \boldsymbol{x}^*, \mathcal{D}) &= \int p(y^* = 1 | \boldsymbol{x}^*, \boldsymbol{\theta}) p(\boldsymbol{\theta} | \mathcal{D}) d\boldsymbol{\theta} \\
&\approx \int p(y^* = 1 | \boldsymbol{x}^*, \boldsymbol{\theta}) q(\boldsymbol{\theta}) d\boldsymbol{\theta} \\
&= \int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

Predictive Distribution with a Weak Prior

Laplace

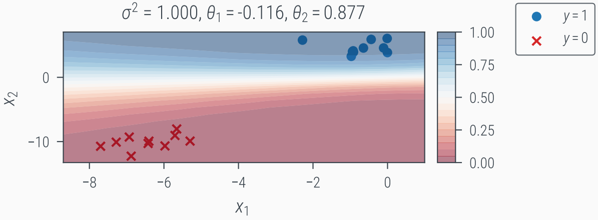


MAP

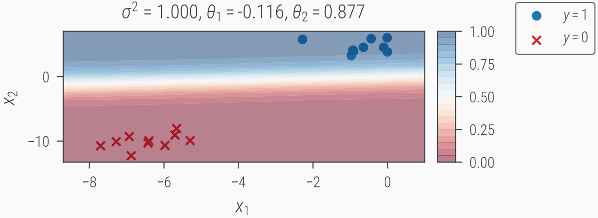


Predictive Distribution with a Medium Prior

Laplace

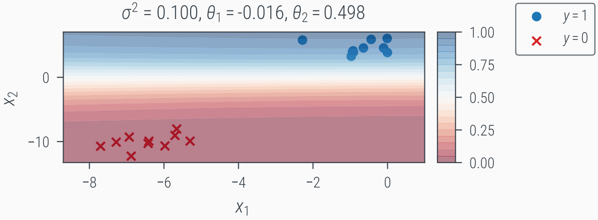


MAP



Predictive Distribution with a Strong Prior

Laplace



MAP