Statistical Learning 3

RG chapter 4, sections 1, 2, 3 and 4

Jakob Söhl

Delft University of Technology

Posterior mode finding for logistic regression

Newton's method for MLE and posterior mode

Consider the logistic regression model:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\log L(\theta) + \log \pi(\theta)\right)$$

with

$$L(\theta) = \prod_{i=1}^{n} p_i^{y_i} (1 - p_i)^{1-y_i} \qquad p_i = \psi(\theta^T x_i)$$

and

$$\psi(z) = \frac{1}{1 + e^{-z}}.$$

For Newton's method we need the partial derivative with respect to θ_j.

Preliminary result

Trivial computation gives that for $y \in \mathbb{R}$:

$$\psi'(z) = \frac{e^{-z}}{(1+e^{-z})^2} = e^{-z}\psi(z)^2 = \psi(z)(1-\psi(z)).$$

Take $z = \theta^T x_i$ and recall $p_i = \psi(\theta^T x_i)$.

$$\psi'(\theta^T x_i) = p_i(1-p_i).$$

REMEMBER: p_i depends on θ, but that is hidden from our notation.

Computing the gradient: likelihood induced term

$$
\begin{aligned}
\frac{\partial \log L(\theta)}{\partial \theta_j} &= \sum_{i=1}^{n} \frac{y_i}{p_i} \frac{\partial p_i}{\partial \theta_j} + \frac{1-y_i}{1-p_i} \frac{\partial (1-p_i)}{\partial \theta_j} \\
&= \sum_{i=1}^{n} \left(\frac{y_i}{p_i} - \frac{1-y_i}{1-p_i} \right) x_{ij} \underbrace{\psi'(\theta^T x_i)}_{p_i(1-p_i)} \\
&= \sum_{i=1}^{n} (y_i - p_i) x_{ij}
\end{aligned}
$$

In matrix-vector notation:

$$
\nabla \log L(\theta) = X^T (y - p).
$$

Computing the Hessian: likelihood induced term

Hessian matrix elements:

$$\begin{aligned}
\frac{\partial}{\partial \theta_k}\frac{\partial \log L(\theta)}{\partial \theta_j} &= \frac{\partial}{\partial \theta_k}\sum_{i=1}^{n}(y_i - p_i)x_{ij} \\
&= -\sum_{i=1}^{n} x_{ij}\frac{\partial p_i}{\partial \theta_k} \\
&= -\sum_{i=1}^{n} x_{ij}x_{ik}\psi'(\theta^T x_i) = -\sum_{i=1}^{n} x_{ij}x_{ik}p_i(1-p_i)
\end{aligned}$$

In matrix-vector notation:

$$H(\theta) = -X^T \mathsf{diag}\left(p_1(1-p_1)\cdots p_n(1-p_n)\right) X.$$

Gradient and Hessian: prior induced terms

Assume $\theta \sim N_p(0, \Sigma_0)$. Then

$$\log \pi(\theta) = -\frac{p}{2}\log(2\pi) - \frac{1}{2}\log|\det \Sigma_0| - \frac{1}{2}\theta^T \Sigma_0^{-1}\theta.$$

This gives

$$\nabla \log \pi(\theta) = -\Sigma_0^{-1}\theta$$

and

$$H(\theta) = -\Sigma_0^{-1}.$$

We aim to compute

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \left(\log L(\theta) + \log \pi(\theta) \right)$$

We found

$$\nabla \left(\log L(\theta) + \log \pi(\theta) \right) = X^T (y - p) - \Sigma_0^{-1} \theta$$

and that the Hessian matrix equals

$$-X^T \mathrm{diag} \underbrace{\left(p_1(1-p_1) \cdots p_n(1-p_n) \right)}_{\Lambda} X - \Sigma_0^{-1}.$$

One step of Newton's method:

$$\theta^{j+1} = \theta^j - (H(\theta^j))^{-1} \nabla F(\theta^j).$$

This becomes (p^j is the vector p with θ^j)

$$\left(X^T \Lambda X + \Sigma_0^{-1} \right) \left(\theta^{j+1} - \theta^j \right) = X^T (y - p^j) - \Sigma_0^{-1} \theta^j.$$

Using logistic regression for classification

Suppose estimate $\hat{\theta}$ has been obtained with Newton's algorithm.

Equi-probability curves are obtained from considering

$$\mathcal{C}_c := \left\{ x \in \mathbb{R}^p \ : \ \frac{1}{1 + e^{-\hat{\theta}^T x}} = c \right\},$$

for $c \in (0, 1)$.

This gives a linear decision boundary:

$$\mathcal{C}_c := \left\{ x \in \mathbb{R}^p \ : \ \hat{\theta}^T x = \log(c/(1-c)) \right\}$$

Example from section 4.3 in RG

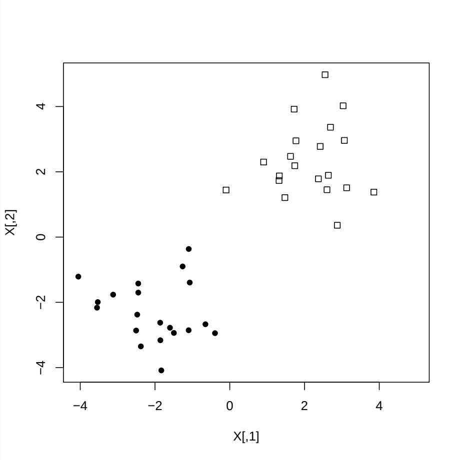


Figure 1: Test data: circle/square distinguishes label. Script *logmap*.

Logistic regression model fitted using Newton's algorithm

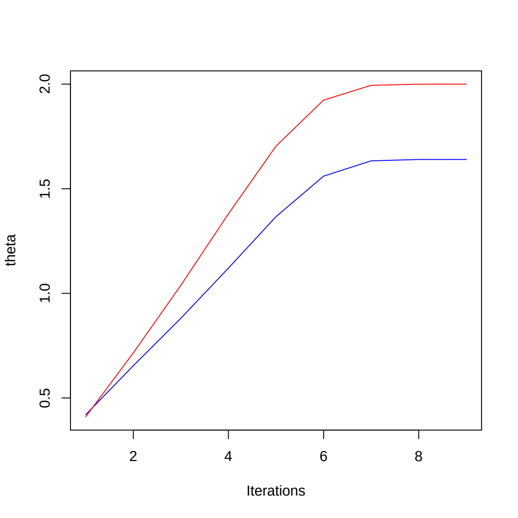


Figure 2: Iterates for θ_1 (blue) and θ_2 (red). Run Newton's algorithm until $\|\theta[it] - \theta[it-1]\|^2 < 10^{-6}$

Visualisation of decision boundary

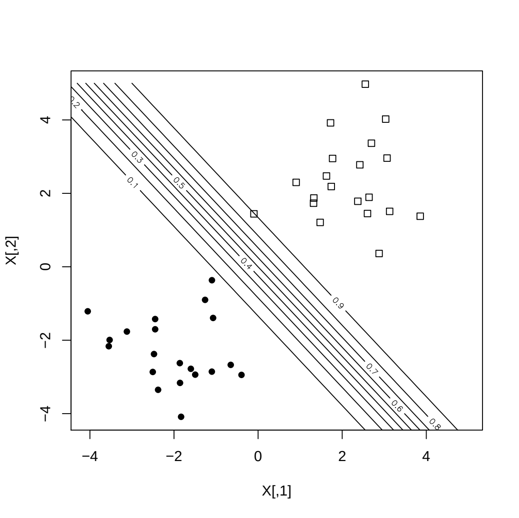


Figure 3: Visualisation of decision boundaries $\mathcal{C}_c$ for various values of c.

Laplace approximation

Laplace approximation

Method for finding a Gaussian approximation to the posterior.

Suppose $\tilde{\Theta}$ is the posterior mode and let $G(\theta)$ and $H(\theta)$ be the gradient and Hessian-matrix of

$$\theta \mapsto \log f_{\Theta|X}(\theta \mid x).$$

$$\log f_{\Theta|X}(\theta \mid x) \approx \log f_{\Theta|X}(\tilde{\Theta} \mid x) + (\theta - \tilde{\Theta})^T G(\tilde{\Theta}) + \frac{1}{2}(\theta - \tilde{\Theta})^T H(\tilde{\Theta})(\theta - \tilde{\Theta})$$

Under smoothness assumptions $G(\tilde{\Theta}) = 0$ and then

$$f_{\Theta|X}(\theta \mid x) \approx\propto \exp\left(\frac{1}{2}(\theta - \tilde{\Theta})^T H(\tilde{\Theta})(\theta - \tilde{\Theta})\right).$$

Thus:

$$f_{\Theta|X}(\theta \mid x) \approx \varphi\left(\theta; \tilde{\Theta}, -H(\tilde{\Theta})^{-1}\right),$$

where $\varphi(x; \mu, \Sigma)$ is the density of the $N(\mu, \Sigma)$-distribution, evaluated at x.

Logistic regression example: true posterior and Laplace approximation

- We already derived a Newton algorithm for computing the MAP.
- We have found an expression for the Hessian $H(\theta)$.

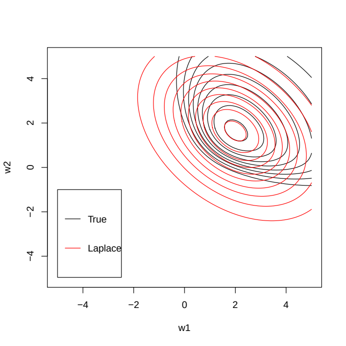
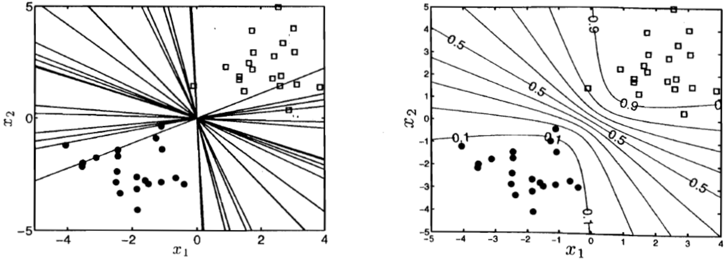


Figure 4: Comparison of contour lines of posterior (with $\theta \sim N(0, 10 \cdot I_2)$ as prior) and its Laplace approximation.

Decision boundaries in logistic regression based on Laplace approximation



(a) Twenty decision boundaries corresponding to instances of $\mathbf{w}$ sampled from the Laplace approximation to the posterior

(b) Contours of $P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \sigma^2)$, computed by using a sample based approximation to $\mathbf{E}_{\mathcal{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})}\ P(T_{\text{new}} = 1|\mathbf{x}_{\text{new}}, \mathbf{w})$

FIGURE 4.7: Decision boundaries sampled from the Laplace approximation and the predictive probability contours.

Explanation on the figure of the previous slide

- The data are X (the design matrix) and y.
- If θ is fixed, the decision boundary is

$$\left\{x_{\text{new}} \in \mathbb{R}^p \;:\; \frac{1}{1+e^{-\theta^T x_{\text{new}}}} = c\right\}.$$

  If all we have is the posterior mode, then we plug this in. This gives linear decision boundaries.

- Now assume we wish to take uncertainty on the estimate of θ into account. Then rather than solving $1+e^{-\theta^T x_{\text{new}}} = 1/c$ for x_{new} we would solve

$$\mathbb{P}(y_{\text{new}} = 1 \mid X, y, x_{\text{new}}) = \mathbb{E}_{\Theta \mid X,y}\left[\frac{1}{1+e^{-\Theta^T x_{\text{new}}}}\right] = c.$$

  Note that the expectation is over the posterior.

- The posterior is not available in close form but can be approximated by the $N(\tilde{\Theta}, -H(\tilde{\Theta})^{-1})$-distribution. Hence, we solve

$$I := \mathbb{E}_{\Theta \sim N(\tilde{\Theta}, -H(\tilde{\Theta})^{-1})}\left[\frac{1}{1+e^{-\Theta^T x_{\text{new}}}}\right] = c.$$

  Note that the expectation is over the Laplace approximation of the posterior.

- The expectation is not known in closed form and hence approximated by Monte Carlo simulation:
  - Sample $\theta_1, \ldots, \theta_M \overset{\text{ind}}{\sim} N(\tilde{\Theta}, -H(\tilde{\Theta})^{-1})$.
  - Approximate I by

$$\frac{1}{M}\sum_{s=1}^{M} \frac{1}{1+e^{-\theta_s^T x_{\text{new}}}}.$$

  - Monte Carlo error can be made arbitrarily small by taking M large; error due to Laplace approximation remains.

- For each realisation of the posterior (whether true posterior or its Laplace approximation), the decision boundary is linear (Figure 4.7(a)).
- However, when averaged over the posterior (reflecting parameter uncertainty), the decision boundary becomes nonlinear (Figure 4.7(b)).