MML Logistic Regression with Translation and Rotation Invariant Priors

Enes Makalic Daniel F. Schmidt

Centre for MEGA Epidemiology
The University of Melbourne

25th Australasian Joint Conference on Artificial Intelligence 2012

Outline

Outline

Problem Description (1)

- We have a binary classification problem
  - Data $\mathbf{y} = (y_1, \ldots, y_n)'$, $y_i = \{-1, +1\}$
  - Matrix of p covariate vectors $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, $\mathbf{x}_j \in \mathbb{R}^n$

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

- Use a logistic regression model (n samples, p predictors)

Problem Description (2)

- Logistic regression model for explaining data $\mathbf{y}$

$$p(\mathbf{y}|\mathbf{X}, \alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \right)$$

  - $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})' \in \mathbb{R}^{p+1}$ is the parameter vector
  - $\alpha \in \mathbb{R}$ is the intercept term
  - $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of logistic regression coefficients

- Log-likelihood for n data points

$$l(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))\right)$$

Problem Description (2)

- Logistic regression model for explaining data $\mathbf{y}$

$$p(\mathbf{y}|\mathbf{X}, \alpha, \boldsymbol{\beta}) = \prod_{i=1}^{n} \left(\frac{1}{1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \right)$$

  - $\boldsymbol{\theta} = (\alpha, \boldsymbol{\beta})' \in \mathbb{R}^{p+1}$ is the parameter vector
  - $\alpha \in \mathbb{R}$ is the intercept term
  - $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of logistic regression coefficients

- Log-likelihood for n data points

$$l(\boldsymbol{\theta}) = -\sum_{i=1}^{n} \log\left(1 + \exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))\right)$$

Motivation

- Task:
  - Estimate parameters
  - Select significant regressors
- Problems with maximum likelihood

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}}(\mathbf{y}) = \arg\max_{\alpha,\beta} \left\{ \prod_{i=1}^{n} \frac{1}{1+\exp(-y_i(\alpha + \mathbf{x}_i'\boldsymbol{\beta}))} \right\}$$

- Minimum Message Length (MML) approach

Outline

Wallace–Freeman Approximation (1)

- The Wallace–Freeman 1987 (WF87) approximation

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{-\log \pi(\boldsymbol{\theta}) + \frac{1}{2} \log |\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| + \frac{k}{2} \log \kappa_k}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})}$$

- As samples goes to infinity ($n \to \infty$)
  - Estimates converge to maximum likelihood
  - Equivalent to the Bayesian Information Criterion (BIC)

Wallace–Freeman Approximation (2)

- WF87 derived under several assumptions
  - Some do not hold in logistic regression
- We use an alternative approximation

$$I_{87}(\mathbf{y}, \boldsymbol{\theta}) = \underbrace{\frac{1}{2} \log \left(1 + \frac{|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) + \mathbf{I}_k| \kappa_k^k}{\pi(\boldsymbol{\theta})^2} \right)}_{I_{87}(\boldsymbol{\theta})} + \underbrace{\frac{k}{2} - \log p(\mathbf{y}|\boldsymbol{\theta})}_{I_{87}(\mathbf{y}|\boldsymbol{\theta})}$$

Outline

Prior Density

- Decision boundary

$$\alpha + \mathbf{x}'\boldsymbol{\beta} = \tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}}) = 0$$

- Re-parameterise

$$\begin{aligned} \tilde{\alpha} &= F_0(\boldsymbol{\theta}) = \alpha \\ \tilde{\beta}_j &= F_j(\boldsymbol{\theta}) = \beta_j/\alpha \quad (j = 1, 2, \ldots, p), \end{aligned}$$

- A prior distribution over $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$

$$\begin{aligned} \pi_{\tilde{\alpha}}(\tilde{\alpha}) &= \frac{a}{2\tilde{\alpha}^2}, \quad \tilde{\alpha} \in [a, \infty) \\ \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}) &= \frac{\Gamma(p/2) r_0}{2\pi^{p/2}} (\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-(p+1)/2}, \quad (||\tilde{\boldsymbol{\beta}}||_2 \geq r_0 > 0) \end{aligned}$$

Prior Density

- Decision boundary

$$\alpha + \mathbf{x}'\boldsymbol{\beta} = \tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}}) = 0$$

- Re-parameterise

$$\begin{aligned} \tilde{\alpha} &= F_0(\boldsymbol{\theta}) = \alpha \\ \tilde{\beta}_j &= F_j(\boldsymbol{\theta}) = \beta_j/\alpha \quad (j = 1, 2, \ldots, p), \end{aligned}$$

- A prior distribution over $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$

$$\begin{aligned} \pi_{\tilde{\alpha}}(\tilde{\alpha}) &= \frac{a}{2\tilde{\alpha}^2}, \quad \tilde{\alpha} \in [a, \infty) \\ \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}) &= \frac{\Gamma(p/2) r_0}{2\pi^{p/2}} (\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-(p+1)/2}, \quad (||\tilde{\boldsymbol{\beta}}||_2 \geq r_0 > 0) \end{aligned}$$

Prior Density

- Decision boundary

$$\alpha + \mathbf{x}'\boldsymbol{\beta} = \tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}}) = 0$$

- Re-parameterise

$$\begin{aligned} \tilde{\alpha} &= F_0(\boldsymbol{\theta}) = \alpha \\ \tilde{\beta}_j &= F_j(\boldsymbol{\theta}) = \beta_j/\alpha \quad (j = 1, 2, \ldots, p), \end{aligned}$$

- A prior distribution over $\tilde{\boldsymbol{\theta}} \in \mathbb{R}^{p+1}$

$$\begin{aligned} \pi_{\tilde{\alpha}}(\tilde{\alpha}) &= \frac{a}{2\tilde{\alpha}^2}, \quad \tilde{\alpha} \in [a, \infty) \\ \pi_{\tilde{\boldsymbol{\beta}}}(\tilde{\boldsymbol{\beta}}) &= \frac{\Gamma(p/2) r_0}{2\pi^{p/2}} (\tilde{\boldsymbol{\beta}}'\tilde{\boldsymbol{\beta}})^{-(p+1)/2}, \quad (||\tilde{\boldsymbol{\beta}}||_2 \geq r_0 > 0) \end{aligned}$$

Fisher information (1)

- Fisher information in the original parameter space

$$|\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})| = \left|(\mathbf{1}_n, \mathbf{X})'\mathbf{V}(\boldsymbol{\theta})(\mathbf{1}_n, \mathbf{X})\right|$$

where

$$\begin{aligned}
\mathbf{1}_n &= (1, 1, \ldots, 1)' \\
\mu_i &= 1/(1 + \exp(-\alpha - \mathbf{x}_i'\boldsymbol{\beta})) \\
\mathbf{V}(\boldsymbol{\theta}) &= \mathrm{diag}(\mu_1(1-\mu_1), \mu_2(1-\mu_2), \ldots, \mu_n(1-\mu_n))
\end{aligned}$$

Fisher information (2)

- Fisher information in the new parametrisation

$$|\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}})| = |\mathbf{J_T}'\mathbf{J}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\mathbf{J_T}| = \tilde{\alpha}^{2p}|(\mathbf{1}_n, \mathbf{X})'\mathbf{V}(\tilde{\boldsymbol{\theta}})(\mathbf{1}_n, \mathbf{X})|$$

where $\mathbf{J_T}$ is the $(k \times k)$ Jacobian transformation matrix

$$\mathbf{J_T} = \begin{pmatrix} 1 & \mathbf{0}'_p \\ \tilde{\boldsymbol{\beta}} & \tilde{\alpha}\mathbf{I}_p \end{pmatrix}, \quad |\mathbf{J_T}| = \tilde{\alpha}^p$$

MML Logistic Regression

- The complete MML codelength for logistic regression

$$I_{87}(\mathbf{y}, \tilde{\boldsymbol{\theta}}) = \frac{1}{2}\log\left(1 + \frac{|\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}}) + \mathbf{I}_k|\kappa_k^k}{\pi(\tilde{\boldsymbol{\theta}})^2}\right) + \sum_{i=1}^{n} \log(1 + \exp(-y_i(\tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}})))) + \frac{k}{2}$$

- Model selection requires stating which regressors are in the model

$$\log(p+1) + \log\binom{p}{q}$$

MML Logistic Regression

- The complete MML codelength for logistic regression

$$I_{87}(\mathbf{y}, \tilde{\boldsymbol{\theta}}) = \frac{1}{2}\log\left(1 + \frac{|\mathbf{J}_{\tilde{\boldsymbol{\theta}}}(\tilde{\boldsymbol{\theta}}) + \mathbf{I}_k|\kappa_k^k}{\pi(\tilde{\boldsymbol{\theta}})^2}\right) + \sum_{i=1}^{n}\log(1 + \exp(-y_i(\tilde{\alpha}(1 + \mathbf{x}'\tilde{\boldsymbol{\beta}})))) + \frac{k}{2}$$

- Model selection requires stating which regressors are in the model

$$\log(p+1) + \log\binom{p}{q}$$

Outline

Introduction

- Simulation results
  - Parameter estimation
    - Classification accuracy, AUC and KL divergence
  - Model selection
- Compared against
  - Maximum likelihood
  - David Firth's penalized likelihood estimator
  - AIC, BIC

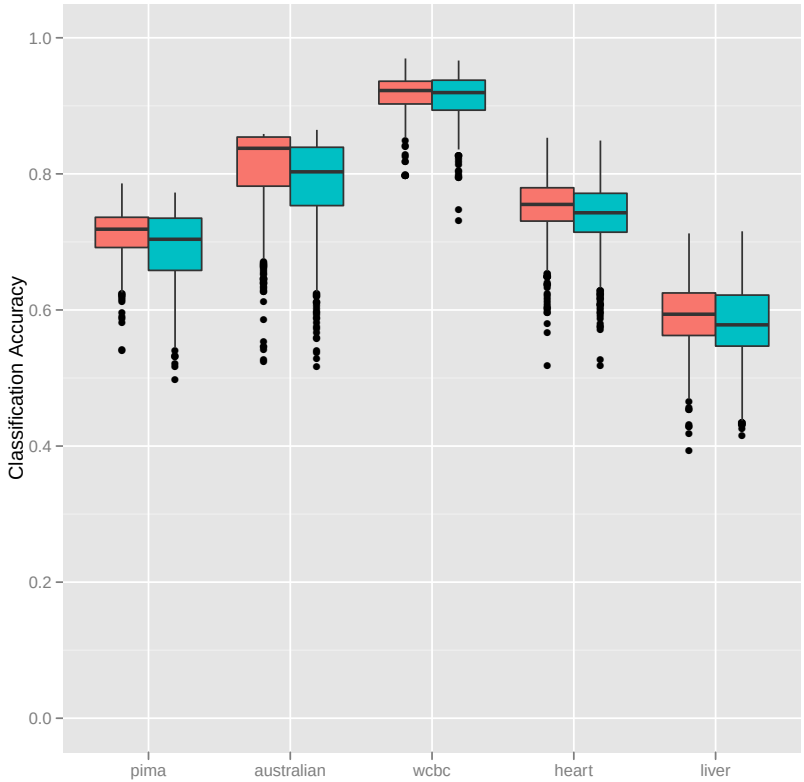
Parameter estimation

n	ρ	$\hat{\boldsymbol{\theta}}_{\mathrm{FR}}(\mathbf{y})$			$\hat{\boldsymbol{\theta}}_{87}(\mathbf{y})$		
		CA	AUC	KL	CA	AUC	KL
25	0.0	72.45	81.11	0.58	72.43	83.10	0.57
	0.2	77.77	87.25	0.47	85.45	94.71	0.40
	0.5	79.92	89.37	0.43	90.44	97.74	0.30
	0.7	80.46	90.10	0.41	91.92	98.45	0.26
	0.9	81.01	90.57	0.41	93.12	98.85	0.24
50	0.0	78.48	87.60	0.49	77.50	87.69	0.52
	0.2	84.06	92.95	0.36	86.89	95.30	0.31
	0.5	86.42	94.88	0.30	91.15	97.83	0.24
	0.7	87.24	95.44	0.29	92.63	98.48	0.21
	0.9	87.84	95.83	0.27	93.66	98.85	0.19
100	0.0	81.47	90.31	0.41	81.50	90.39	0.40
	0.2	87.26	95.23	0.30	87.95	95.69	0.28
	0.5	89.89	96.97	0.24	91.26	97.72	0.21
	0.7	90.69	97.45	0.22	92.69	98.38	0.18
	0.9	91.33	97.79	0.20	93.84	98.85	0.15
250	0.0	82.93	91.49	0.37	82.95	91.50	0.37
	0.2	88.67	96.07	0.26	88.84	96.17	0.26
	0.5	91.39	97.71	0.20	91.76	97.88	0.19
	0.7	92.33	98.17	0.18	92.85	98.40	0.17
	0.9	92.95	98.47	0.17	93.89	98.82	0.15

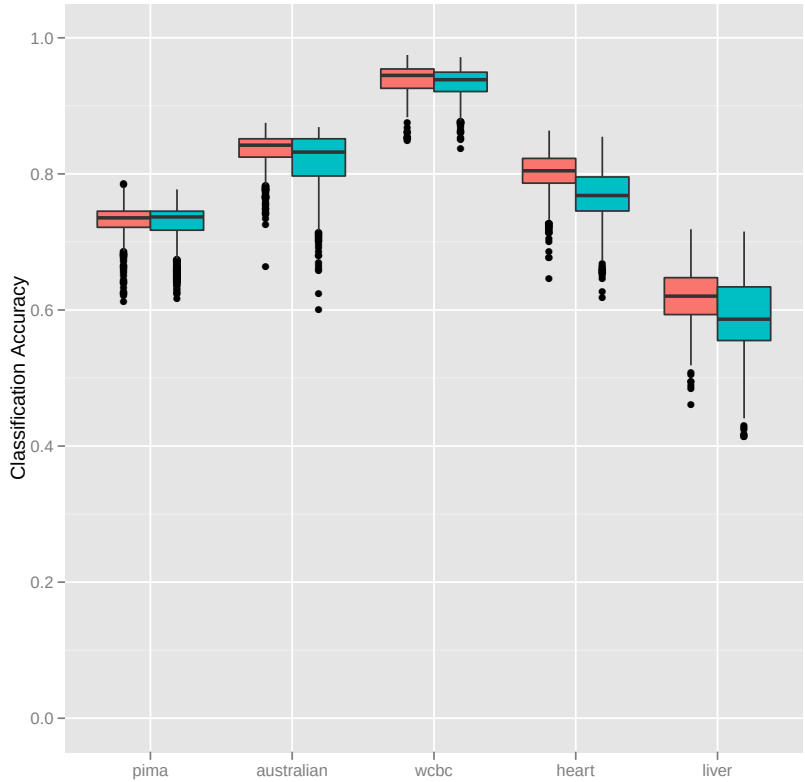
Model selection (1)

- Data sets
  - pima (8 predictors, 768 samples)
  - australian (15 predictors, 690 samples)
  - wcbc (10 predictors, 683 samples)
  - liver (6 predictors, 345 samples)
  - heart (13 predictors, 270 samples)

Model selection (2)

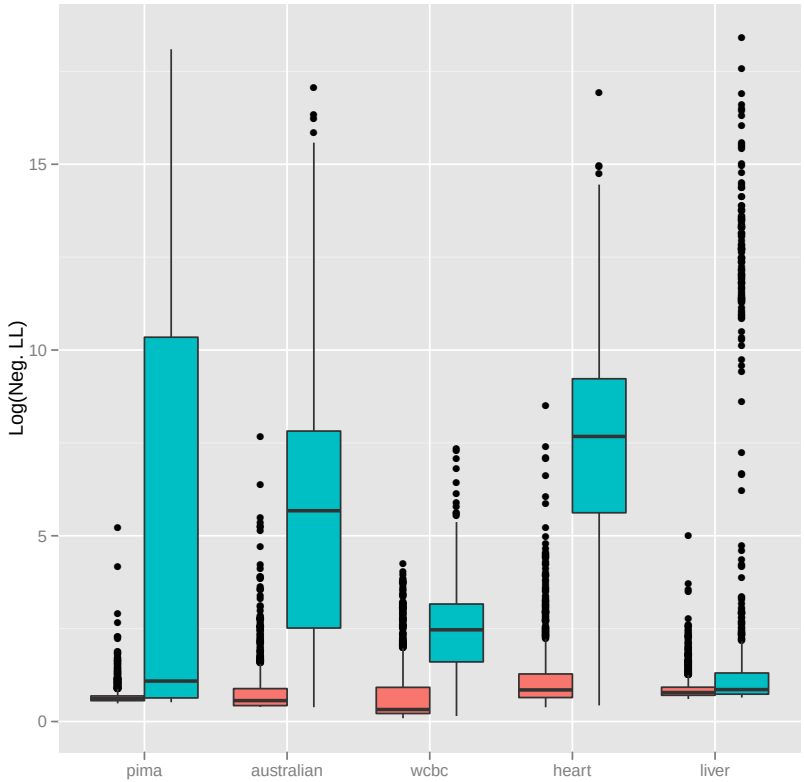


(a) $n = 25$

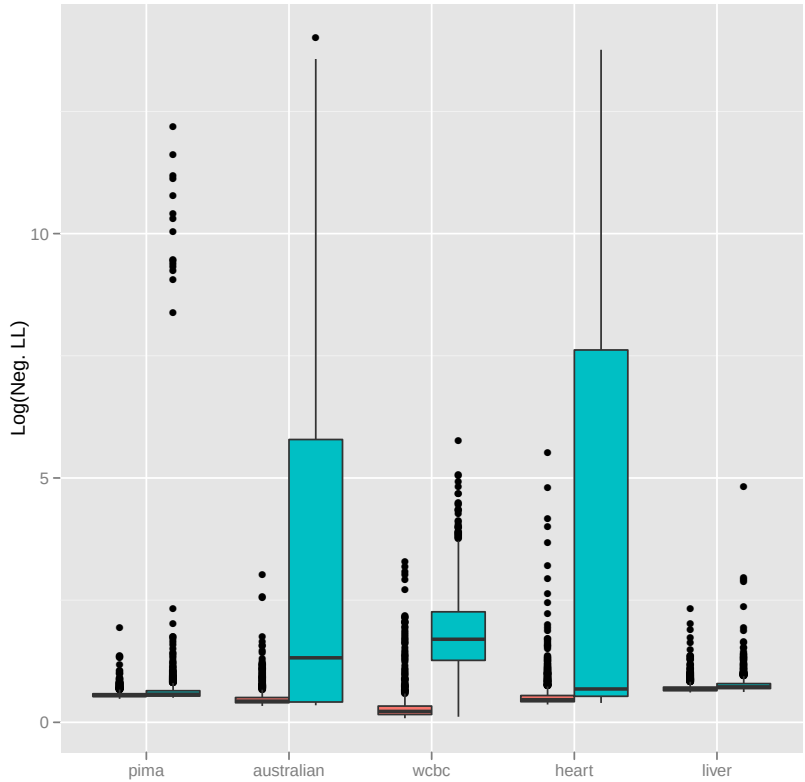


(b) $n = 50$

Model selection (3)



(c) $n = 25$



(d) $n = 50$