Binary probit regression with I-priors

Haziq Jamil
Supervisors: Dr. Wicher Bergsma & Prof. Irini Moustaki

Social Statistics (Year 3)
London School of Economics & Political Science

8-9 May 2017

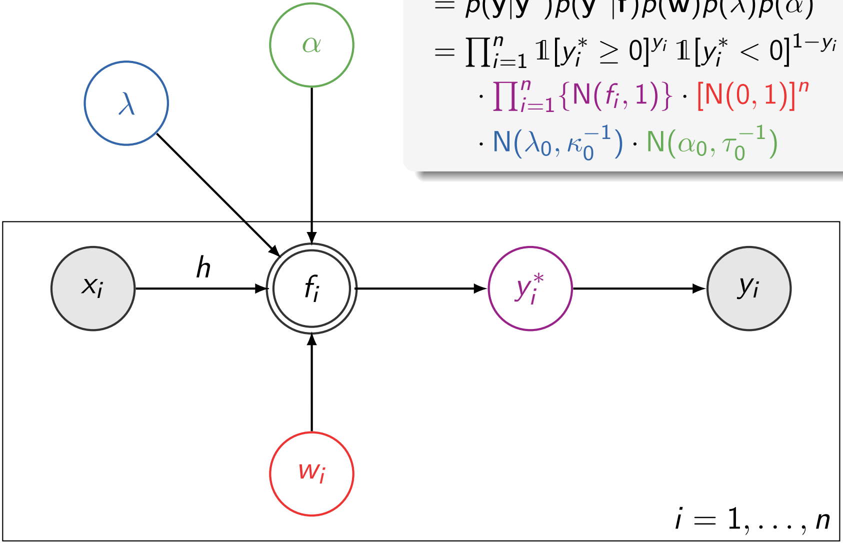
PhD Presentation Event

http://phd3.haziqj.ml

Outline

Variational I-prior probit

$$
\begin{aligned}
&p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda) \\
&= p(\mathbf{y}|\mathbf{y}^*)p(\mathbf{y}^*|\mathbf{f})p(\mathbf{w})p(\lambda)p(\alpha) \\
&= \textstyle\prod_{i=1}^n \mathbb{1}[y_i^* \geq 0]^{y_i}\, \mathbb{1}[y_i^* < 0]^{1-y_i} \\
&\quad \cdot \textstyle\prod_{i=1}^n \{\mathrm{N}(f_i, 1)\} \cdot [\mathrm{N}(0,1)]^n \\
&\quad \cdot \mathrm{N}(\lambda_0, \kappa_0^{-1}) \cdot \mathrm{N}(\alpha_0, \tau_0^{-1})
\end{aligned}
$$



Posterior distribution

- Approximate the posterior by a mean-field variational density

$$p(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda|\mathbf{y}) \approx \prod_{i=1}^n q(y_i^*)q(\mathbf{w})q(\alpha)q(\lambda)$$

where

$$q(y_i^*) \equiv \begin{cases} \mathbb{1}[y_i^* \geq 0]\,\mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 1 \\ \mathbb{1}[y_i^* < 0]\,\mathrm{N}(\tilde{f}_i, 1) & \text{if } y_i = 0 \end{cases} \qquad q(\mathbf{w}) \equiv \mathrm{N}(\tilde{\mathbf{w}}, \tilde{\mathbf{V}}_w)$$

$$q(\lambda) \equiv \mathrm{N}(\tilde{\lambda}, \tilde{v}_w) \qquad q(\alpha) \equiv \mathrm{N}(\tilde{\alpha}, 1/n)$$

$$\tilde{f}_i = \tilde{\alpha} + \textstyle\sum_{k=1}^n h_{\tilde{\lambda}}(x_i, x_k)\tilde{w}_k \qquad \tilde{\alpha} = \frac{1}{n}\textstyle\sum_{k=1}^n \left(\mathrm{E}[\mathbf{y}_i^*] - h_{\tilde{\lambda}}(x_i, x_k)\tilde{w}_k\right)$$

$$\tilde{\mathbf{w}} = \tilde{\mathbf{V}}_w \mathbf{H}_{\tilde{\lambda}}(\mathrm{E}[\mathbf{y}^*] - \tilde{\alpha}\mathbf{1}_n) \qquad \tilde{\mathbf{V}}_w^{-1} = \mathbf{H}_{\tilde{\lambda}}^2 + \mathbf{I}_n$$

$$\tilde{\lambda} = (\mathrm{E}[\mathbf{y}^*] - \tilde{\alpha}\mathbf{1}_n)\mathbf{H}\tilde{\mathbf{w}}/\tilde{v}_\lambda \qquad \tilde{v}_\lambda = \mathrm{tr}(\mathbf{H}^2(\tilde{\mathbf{V}}_w + \tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top))$$

Posterior predictive distribution

- Given new data points x_{new}, interested in

$$
\begin{aligned}
p(y_{\text{new}}|\mathbf{y}) &= \int p(y_{\text{new}}|y^*_{\text{new}}, \mathbf{y}) p(y^*_{\text{new}}|\mathbf{y}) \, \mathrm{d}y^*_{\text{new}} \\
&\approx \int p(y_{\text{new}}|y^*_{\text{new}}) q(y^*_{\text{new}}) \, \mathrm{d}y^*_{\text{new}} \\
&= \begin{cases} \Phi(\tilde{f}_{\text{new}}) & \text{if } y_{\text{new}} = 1 \\ 1 - \Phi(\tilde{f}_{\text{new}}) & \text{if } y_{\text{new}} = 0 \end{cases}
\end{aligned}
$$

where $\tilde{f}_{\text{new}} = \tilde{\alpha} + \sum_{k=1}^n h_{\tilde{\lambda}}(x_{\text{new}}, x_k)\tilde{w}_k$.

- f_{new} represents the estimate of the latent propensity for y_{new}, and its uncertainty is described by $q(y^*_{\text{new}})$.

Variational lower bound

- Since the solutions are coupled, we implement an iterative scheme (as per Algorithm ??)

- Assess convergence by monitoring the lower bound

$$\begin{aligned}\mathcal{L} &= \mathrm{E}_q[\log p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] - \mathrm{E}_q[\log q(\mathbf{y}^*, \mathbf{w}, \alpha, \lambda)] \\ &= \text{const.} + \sum_{i=1}^n \Big(y_i \log \Phi(\tilde{f}_i) + (1 - y_i) \log \big(1 - \Phi(\tilde{f}_i)\big) \Big) \\ &\quad - \frac{1}{2} \Big(\operatorname{tr} \tilde{\mathbf{V}}_w + \operatorname{tr}(\tilde{\mathbf{w}}\tilde{\mathbf{w}}^\top) - \log |\tilde{\mathbf{V}}_w| + \log \tilde{v}_\lambda \Big)\end{aligned}$$

- ISSUE: Different initialisation leads to different converged lower bound values indicating presence of many local optima.

R/iprobit

HJ (2017). *iprobit: Binary Probit Regression with I-priors*. R Package version 0.1.0: GitHub

Fisher's Iris data set

1. Intro. Combine some groups so binary classification problem. For illustration just use sepal length and width (to get nice plots). 2. Fit model. Syntax. Summary. 3. Multiple starting values leads to different L. 4. Plot LB. Plot decision boundary.

Cardiac arrhythmia data set

1. Intro. Number of covariates. 2. Subsample, fit and get SE for out-of-sample test error rates. 3. Compare with other classifiers.

Multilevel example

Not sure what yet. Something that latent propensities might be worth measuring? Maybe fitted probabilities too.

End

Thank you!