Binary probit regression with I-priors

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

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

8 May 2017

PhD Presentation Event

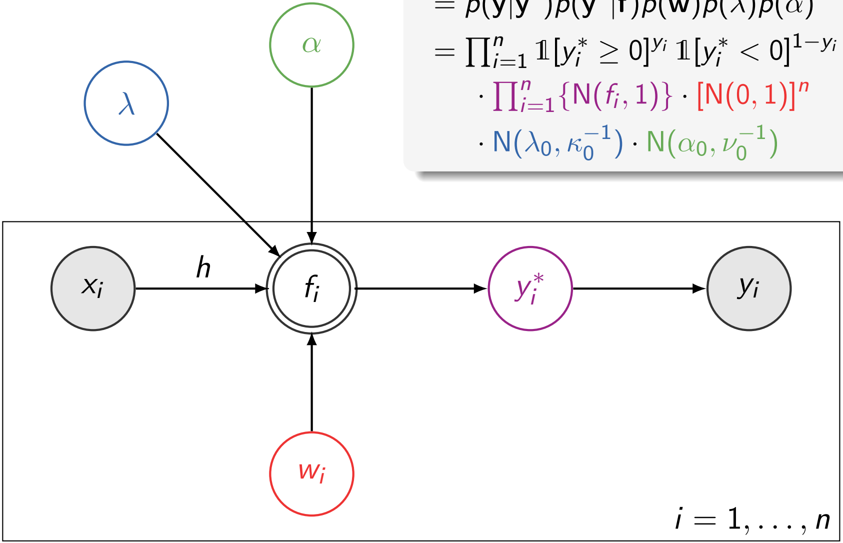
http://phd3.haziqj.ml

Outline

1. Implementation
2. Summary

Variational I-prior probit

$$
\begin{aligned}
& p(\mathbf{y}, \mathbf{y}^*, \mathbf{w}, \alpha, \lambda) \\
& = p(\mathbf{y}|\mathbf{y}^*) p(\mathbf{y}^*|\mathbf{f}) p(\mathbf{w}) p(\lambda) p(\alpha) \\
& = \textstyle\prod_{i=1}^n \mathbb{1}[y_i^* \geq 0]^{y_i} \, \mathbb{1}[y_i^* < 0]^{1-y_i} \\
& \quad \cdot \textstyle\prod_{i=1}^n \{\mathrm{N}(f_i, 1)\} \cdot [\mathrm{N}(0,1)]^n \\
& \quad \cdot \mathrm{N}(\lambda_0, \kappa_0^{-1}) \cdot \mathrm{N}(\alpha_0, \nu_0^{-1})
\end{aligned}
$$



Posterior distribution

- Approximate the posterior by a mean-field variational density

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

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

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

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

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

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

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

$$\tilde f_i = \tilde\alpha + \textstyle\sum_{k=1}^n h_{\tilde\lambda}(x_i, x_k) \tilde w_k \qquad \tilde\alpha = \frac{1}{n} \textstyle\sum_{k=1}^n \left(\mathrm{E}[y_i^*] - h_{\tilde\lambda}(x_i, x_k) \tilde w_k \right)$$

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

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

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

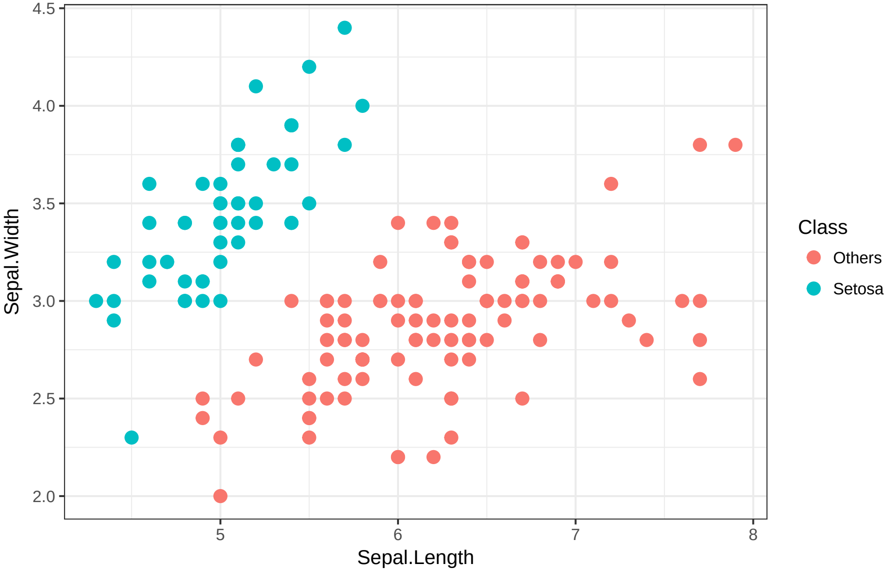
- f_{new} represents the estimate of the latent propensity for y_{new}, and its uncertainty is described by $q(y^*_{\text{new}})$.

Variational lower bound

- Since the solutions are coupled, we implement an iterative scheme (as per Algorithm **??**)
- Assess convergence by monitoring the lower bound

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

Fisher's Iris data set



Fisher's Iris data set - Model fitting

- Varitional inference for I-prior probit models implemented in R package iprobit (still lots of work to do!).

```
R> system.time(
+   (mod <- iprobit(y, X))
+ )

##
## |=================================                  |  61%
##  Converged after 6141 iterations.
##  Training error rate: 0 %
##     user  system elapsed
##   67.857   6.396  74.277
```

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

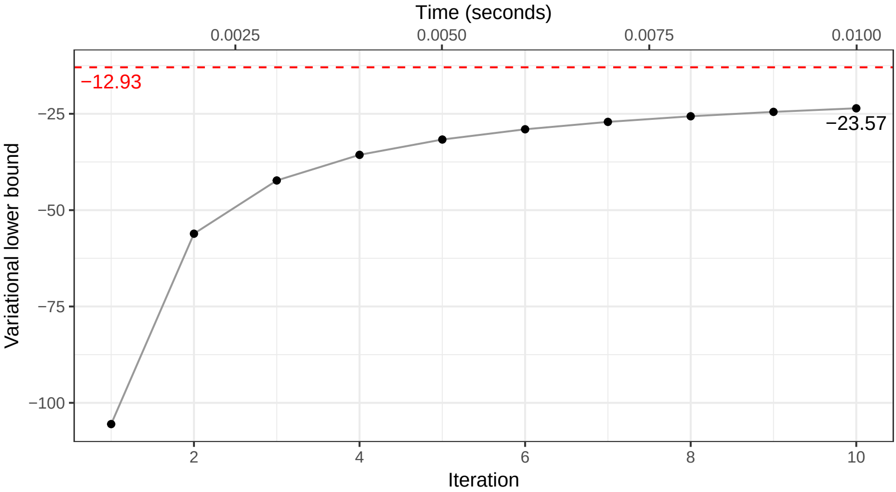
Fisher's Iris data set - Model summary

```
R> summary(mod)

##
## Call:
## iprobit(y = y, X, maxit = 10000)
##
## RKHS used: Canonical
##
##             Mean   S.E.    2.5%   97.5%
## alpha  -4.1730 0.0816 -4.3330 -4.0129
## lambda  1.2896 0.0142  1.2618  1.3175
##
## Converged to within 1e-05 tolerance. No. of iterations: 6141
## Model classification error rate (%): 0
## Variational lower bound: -12.93486
```

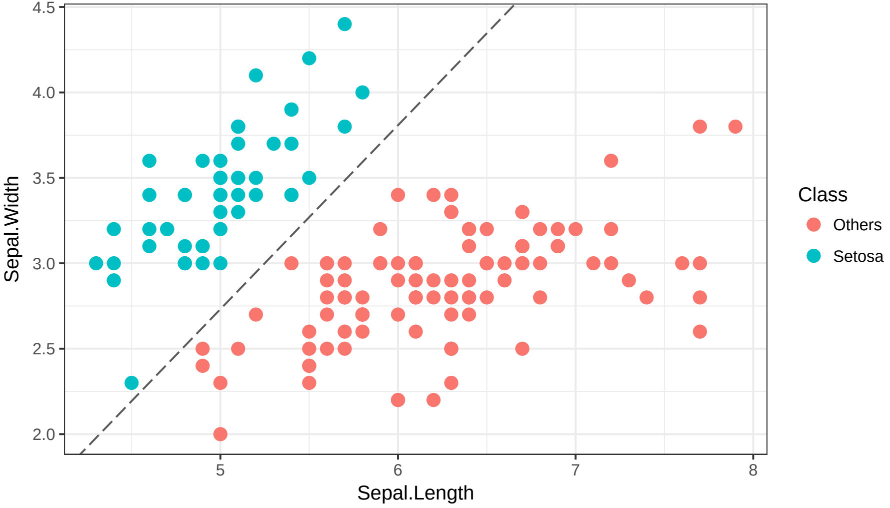
Fisher's Iris data set - Lower bound

```
R> iplot_lb(mod, niter.plot = 10)
```



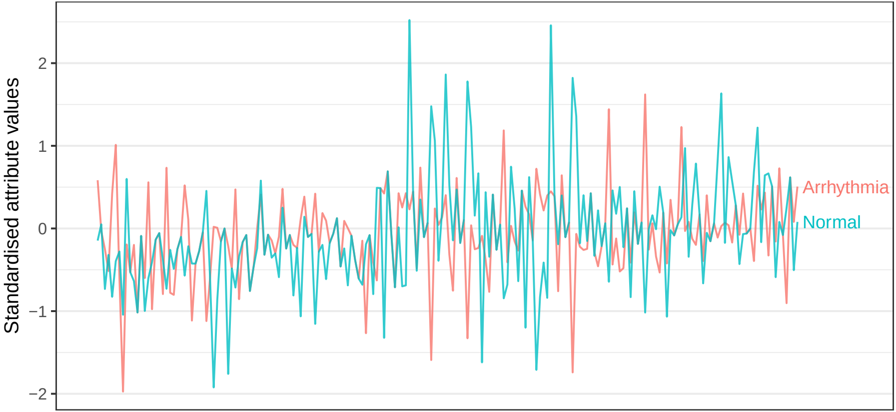
Fisher's Iris data set - Decision boundary

```
R> iplot_decbound(mod)
```



Cardiac arrhythmia data set

- Detect the presence of cardiac arrhythmia based on various ECG data and other attributes such as age and weight ($n = 451, p = 194$).



H. A. Guvenir et al. (1998). *UCI Machine Learning Repository: Arrhythmia Data Set*. URL: `https://archive.ics.uci.edu/ml/datasets/Arrhythmia`

Cardiac arrhythmia data set - Model fit

- Fit an I-prior probit model using Canonical and FBM kernel. The full data set takes about 35 seconds.

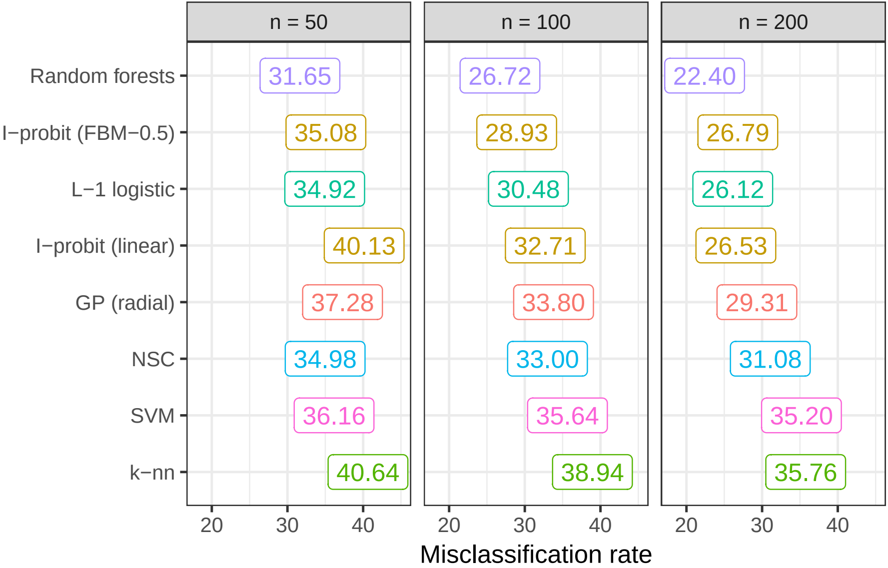
```
R> mod <- iprior(y, X, kernel = "FBM")
```

- Compare against popular classifiers: 1) k-nearest neighbours; 2) support vector machine; 3) Gaussian process classification; 4) random forests; 5) nearest shrunken centroids (Tibshirani et al. 2003); and 6) L-1 penalised logistic regression.

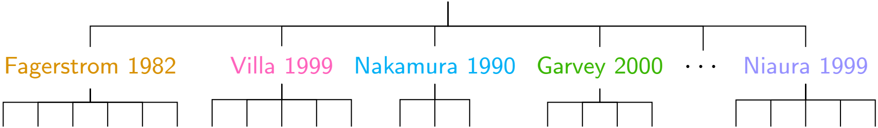
- Experiment set-up:
  - Form training set by sub-sampling $n_{\text{sub}} \in \{50, 100, 200\}$ data points.
  - Use remaining data as test set.
  - Fit model on training set and obtain test error rates.
  - Repeat 100 times.

T. I. Cannings and R. J. Samworth (2017). "Random-projection ensemble classification". *J. R. Stat. Soc. Ser. B: Stat. Methodol (w. discussion)*, to appear

Cardiac arrhythmia data set - Results



Meta-analysis of smoking cessation



- Data from 27 separate smoking cessation studies, where participants subjected to nicotine gum treatment or placed in control group.
- Some summary statistics:

	Min.	Avg.	Max.	Prop. quit	Odds quit
Control	20	101	617	0.207	0.261
Treated	21	117	600	0.320	0.470

- Raw odds ratio: 1.801.
- Random-effects analysis using a multilevel logistic model estimates this odds ratio as 1.768.

A. Skrondal and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC, §9.5

Meta-analysis of smoking cessation - model

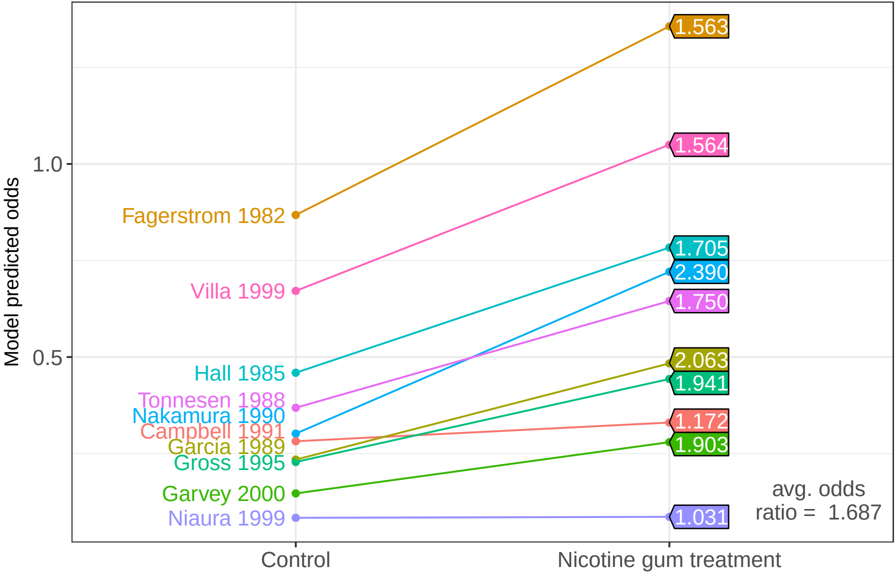
- Let $i = 1, \dots, n_j$ index the patients in study group $j \in 1, \dots, 27$.
- Denote y_{ij} as the binary response variable indicating `Quit` (1) or `Remain` (0), and x_{ij} as patient$_{ij}$'s treatment group indicator.
- Model binary data using I-probit model

$$\begin{aligned}\Phi^{-1}(p_{ij}) &= f(x_{ij}, j) \\ &= f_1(x_{ij}) + f_2(j) + f_{12}(x_{ij}, j)\end{aligned}$$

with $f_1, f_2 \in$ Pearson RKHS, and $f_{12} \in$ ANOVA RKHS.

	Model	Lower bound	Brier score	No. of RKHS param.
1	f_1	-3210.79	0.0311	1
2	$f_1 + f_2$	-3097.24	0.0294	2
3	$f_1 + f_2 + f_{12}$	-3091.21	0.0294	2

Meta-analysis of smoking cessation - results



Summary

- An extension of the I-prior methodology to binary responses.
- Variational inference used to approximate the intractable likelihood.
  - A deterministic approximation of the posterior density by a "close" (in the KL divergence sense), tractable density.
  - It's somewhere between Laplace's method and MCMC sampling.
- Several real-world examples demonstrated the use of I-probit models for classification and inference.
- Further work:
  - R package `iprobit`.
  - Extend to non-iid errors case.
  - Extend to multinomial probit models.
  - Other algorithms (e.g. expectation propagation).

Slides, source code and results are made available at: http://phd3.haziqj.ml

End

Thank you!

References I

Cannings, T. I. and R. J. Samworth (2017). "Random-projection ensemble classification". *Journal of the Royal Statistical Society. Series B: Statistical Methodology (with discussion)*, to appear.

Guvenir, H. A., M. Burak Acar, and H. Muderrisoglu (1998). *UCI Machine Learning Repository: Arrhythmia Data Set*. URL: https://archive.ics.uci.edu/ml/datasets/Arrhythmia.

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

Skrondal, A. and S. Rabe-Hesketh (2004). *Generalized Latent Variable Modeling: Multilevel, Longitudinal, and Structural Equation Models*. Chapman & Hall/CRC.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2003). "Class prediction by nearest shrunken centroids, with applications to DNA microarrays". *Statistical Science*, pp. 104–117.

3 Additional material