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

The latent variable motivation

- Consider binary responses $y_1, \dots, y_n$ together with their corresponding covariates $x_1, \dots, x_n$.
- For $i = 1, \dots, n$, model the responses as

$$y_i \sim \mathrm{Bern}(p_i).$$

- Assume that there exists continuous, underlying latent variables $y_1^*, \dots, y_n^*$, such that

$$y_i = \begin{cases} 1 & \text{if } y_i^* \geq 0 \\ 0 & \text{if } y_i^* < 0. \end{cases}$$

- Model these continuous latent variables according to

$$y_i^* = f(x_i) + \epsilon_i$$

where $(\epsilon_1, \dots, \epsilon_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi}^{-1})$ and $f \in \mathcal{F}$ (some RKHS).

Using I-priors

- Assume an I-prior on f. Then,

$$f(x_i) = \alpha + \sum_{k=1}^{n} h_\lambda(x_i, x_k) w_k$$
$$(w_1, \dots, w_n) \sim \mathrm{N}(\mathbf{0}, \mathbf{\Psi})$$

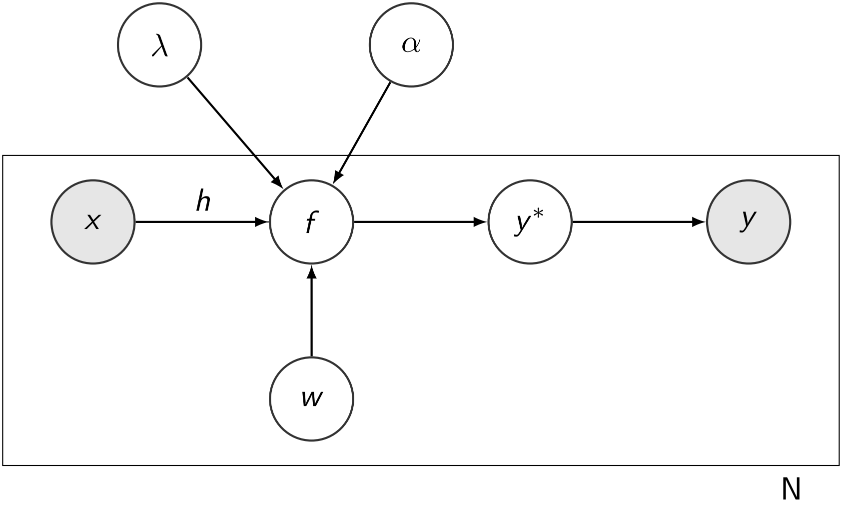
- For now, consider iid errors $\mathbf{\Psi} = \psi \mathbf{I}_n$. In this case,

$$\begin{aligned} p_i = \mathrm{P}[y_i = 1] &= \mathrm{P}[y_i^* \geq 0] \\ &= \mathrm{P}[\epsilon_i \leq f(x_i)] \\ &= \Phi\Big(\psi^{1/2}(\alpha + \textstyle\sum_{k=1}^n h_\lambda(x_i, x_k) w_k)\Big) \end{aligned}$$

  where Φ is the CDF of a standard normal.

- No loss of generality compared with using an arbitrary threshold τ or error precision ψ. Thus, set $\psi = 1$.

The probit I-prior model



Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) \, \mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^{n} \left[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \right] \cdot \mathrm{N}(\alpha \mathbf{1}_n, \mathbf{H}_\lambda^2) \, \mathrm{d}\mathbf{f}
\end{aligned}
$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$\begin{aligned} p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\ &= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \Big] \cdot \mathrm{N}(\alpha \mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f} \end{aligned}$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral

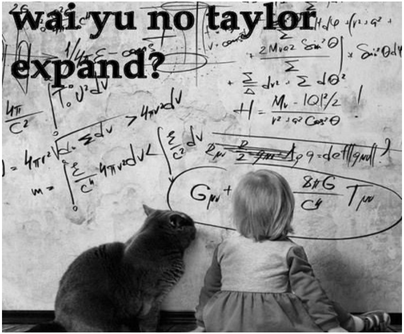
Estimation

- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \Big] \cdot \mathrm{N}(\alpha \mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}
\end{aligned}
$$

  for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.
- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step

Estimation



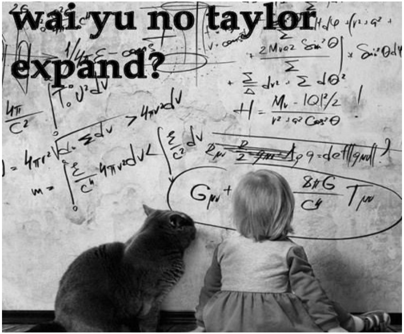
- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$
\begin{aligned}
p(\mathbf{y}) &= \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f} \\
&= \int \prod_{i=1}^n \left[\Phi(f_i)^{y_i} \big(1 - \Phi(f_i)\big)^{1-y_i} \right] \cdot \mathrm{N}(\alpha\mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}
\end{aligned}
$$

  for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step
  - ✓ Laplace approximation

Estimation



- Denote $f_i = f(x_i)$ for short.
- The marginal density

$$p(\mathbf{y}) = \int p(\mathbf{y}|\mathbf{f})p(\mathbf{f})\,\mathrm{d}\mathbf{f}$$
$$= \int \prod_{i=1}^{n} \Big[\Phi(f_i)^{y_i}\big(1-\Phi(f_i)\big)^{1-y_i}\Big] \cdot \mathrm{N}(\alpha\mathbf{1}_n, \mathbf{H}_\lambda^2)\,\mathrm{d}\mathbf{f}$$

for which $p(\mathbf{f}|\mathbf{y})$ depends, cannot be evaluated analytically.

- Some strategies:
  - ✗ Naive Monte-Carlo integral
  - ✗ EM algorithm with a MCMC E-step
  - ✓ Laplace approximation
  - ✓ MCMC sampling

Laplace's method

- Interested in $p(\mathbf{f}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{f})p(\mathbf{f}) =: e^{Q(\mathbf{f})}$, with normalising constant $p(\mathbf{y}) = \int e^{Q(\mathbf{f})} \, \mathrm{d}\mathbf{f}$. The Taylor expansion of Q about its mode $\tilde{\mathbf{f}}$

$$Q(\mathbf{f}) \approx Q(\tilde{\mathbf{f}}) - \frac{1}{2}(\mathbf{f} - \tilde{\mathbf{f}})^\top \mathbf{A}(\mathbf{f} - \tilde{\mathbf{f}})$$

  is recognised as the logarithm of an unnormalised Gaussian density, with $\mathbf{A} = -\mathrm{D}^2 Q(\mathbf{f})$ being the negative Hessian of Q evaluated at $\tilde{\mathbf{f}}$.

- The posterior $p(\mathbf{f}|\mathbf{y})$ is approximated by $\mathrm{N}(\tilde{\mathbf{f}}, \mathbf{A}^{-1})$, and the marginal by

$$p(\mathbf{y}) \approx (2\pi)^{n/2} |\mathbf{A}|^{-1/2} p(\mathbf{y}|\tilde{\mathbf{f}}) p(\tilde{\mathbf{f}})$$

- Won't scale with large n; difficult to find modes in high dimensions.

R. Kass and A. Raftery (1995). "Bayes Factors". *Journal of the American Statistical Association* 90.430, pp. 773–795, §4.1, pp. 777-778.

Full Bayesian analysis using MCMC

- Assign hyperpriors on parameters of the I-prior, e.g.
  - $\lambda^2 \sim \Gamma^{-1}(a, b)$
  - $\alpha \sim \mathrm{N}(c, d^2)$

  for a hierarchical model to be estimated fully Bayes.
- No closed-form posteriors - need to resort to MCMC sampling.
- Computationally slow, and sampling difficulty results in unreliable posterior samples.

DENSITY PLOTS OF LAMBDA HERE

Variational inference introduction

- Name derived from calculus of variations which deals with maximising or minimising functionals.

$$\begin{array}{lll} \text{Functions} & p : \theta \mapsto \mathbb{R} & \text{(standard calculus)} \\ \text{Functionals} & \mathcal{H} : p \mapsto \mathbb{R} & \text{(variational calculus)} \end{array}$$

- Using standard calculus, we can solve

$$\arg\max_\theta p(\theta) =: \hat\theta$$

e.g. p is a likelihood function, and $\hat\theta$ is the ML estimate.

- Using variational calculus, we can solve

$$\arg\max_p \mathcal{H}(p) =: \tilde p$$

e.g. $\mathcal{H}$ is the entropy $\mathcal{H} = -\int p(x) \log p(x) \, \mathrm{d}x$, and $\tilde p$ is the entropy maximising distribution.

C. M. Bishop (2006). *Pattern Recognition and Machine Learning*. Springer

Variational inference introduction (cont.)

- Consider a statistical model where we have observations $(y_1, \dots, y_n)$ and also some latent variables $(z_1, \dots, z_n)$.
- The z_is could be random effects or some auxiliary latent variables.
- In a Bayesian setting, this could also include the parameters to be estimated.
- **GOAL**: Find approximations for
  - The posterior distribution $p(\mathbf{z}|\mathbf{y})$; and
  - The marginal likelihood (or model evidence) $p(\mathbf{y})$.

Decomposition of the log marginal

- Let $q(\mathbf{z})$ be some density function to approximate $p(\mathbf{z}|\mathbf{y})$. Then the log-marginal density can be decomposed into

$$\begin{aligned}
\log p(\mathbf{y}) &= \log p(\mathbf{y}, \mathbf{z}) - \log p(\mathbf{z}|\mathbf{y}) \\
&= \int \left\{ \log \frac{p(\mathbf{y}, \mathbf{z})}{q(\mathbf{z})} - \log \frac{p(\mathbf{z}|\mathbf{y})}{q(\mathbf{z})} \right\} q(\mathbf{z}) \, d\mathbf{z} \\
&= \mathcal{L}(q) + \mathrm{KL}(q \| p) \\
&\geq \mathcal{L}(q)
\end{aligned}$$

- $\mathcal{L}$ is referred to as the "lower-bound", and it serves as a surrogate function to the marginal.
- Maximising the $\mathcal{L}(q)$ is equivalent to minimising $\mathrm{KL}(q\|p)$.
- Although $\mathrm{KL}(q\|p)$ is minimised at $q(\mathbf{z}) \equiv p(\mathbf{z}|\mathbf{y})$ (c.f. EM algorithm), we are unable to work with $p(\mathbf{z}|\mathbf{y})$.

Comparison

Factorised distributions (Mean field theory)

Variational Bayes EM

Estimation of Gaussian mean and variance

End

Thank you!