SPR day 10

Bayesian Logistic Regression

- Posterior
- Evidence
- Predictive Posterior

Reading: Sec 4.4 of Bishop PRML
Sec 4.5

Outline:
(1) Overview of posterior + predictive
(2) Laplace approximation in 1-dim and M-dim
(3) Laplace approximation for logistic regression
(4) Posterior predictive approximation

Bayesian Logistic Regression

Model:

usually $m_0 = \vec{0}$ all zero
$S_0 = \alpha^{-1} I_M$ with $\alpha > 0$

Prior on weights $w \in \mathbb{R}^M$

$$p(w) = \mathcal{N}(w \mid \overset{\text{mean}}{m_0}, \overset{\text{covar}}{S_0})$$

just like linear regr.

Likelihood of "outputs" $t_n \in \{0,1\}$ (binary)

$$p(t \mid w) = \prod_{n=1}^{N} \mathrm{Bern}(t_n \mid \sigma(w^T \phi(x_n)))$$

treat x as known, fixed

iid across examples

Goals are to estimate the posterior and predictive

Posterior: $p(w \mid t_{1:N})$

no closed form!
Not a Gaussian!

$$p(t_* \mid t_{1:N})$$

if we only denote random variables

Predictive: $= p(t_* \mid x_*, \{x_n, t_n\}_{n=1}^{N})$

if we make known x vals explicit

$$= \int_w \underbrace{p(t_* \mid w, x_*)}_{\text{likelihood}} \underbrace{p(w \mid \{x_n, t_n\}_{n=1}^{N})}_{\text{posterior}} dw$$

Must be a Bernoulli
(r.v. t_* is binary)
But no closed-form
for its parameter

Laplace Approximation in 1D

Given: a random variable $w \in \mathbb{R}$
whose density $p(w)$ is known up to norm. const.

$$p(w) = \frac{1}{Z} f(w) \iff \log p(w) = \log f(w) + \text{const}$$

Here, $f(w) > 0$ is known and evaluatable and differentiable
but computing $Z = \int_w f(w)\, dw$ is hard

How can we estimate the distribution $p(w)$?

mean m precision $\beta > 0$

Idea: Approximate with a Gaussian: $q(w) = \mathcal{N}(m, \beta^{-1})$
- pick mean to match the mode of $p(w)$

$$m = \underset{w \in \mathbb{R}}{\operatorname{argmax}}\ p(w) = \underset{w \in \mathbb{R}}{\operatorname{argmax}}\ f(w)$$

Can use Gradient Methods to solve this numerically

- pick precision to perform best possible 2nd-order Taylor approximation to $p(w)$ at the mode $w = m$

$$\beta = \frac{\partial}{\partial w}\frac{\partial}{\partial w}\left[-\log f(w)\right]\Big|_{w=m}$$

$$= -\ell''(m) \quad \text{where } \ell = \log f(w)$$

Advantages: Gives an approx distribution we can reason about!
Second derivatives are often tractable

Limitations: bad if $p(w)$ multimodal
bad if $p(w)$ ~~not~~ has heavy tails, symmetric about mode

Derivation of Taylor approx to density $p(w)$ at $m = \arg\max_w p(w)$

$$\log p(w) = \log f(w) + \text{const wrt } w$$

by definition of $p(w)$

$$= \ell(w) + \text{const}_1$$

define $\ell(w) = \log f(w)$

note that m is a mode of $\ell(w)$ too!

$$= \ell(m) + \cancelto{0}{\ell'(m)(w-m)} + \frac{\ell''(m)}{2}(w-m)^2 + \text{const}_1$$

2^{nd} order Taylor approx. to func. ℓ at $w = m$

$$= -\frac{1}{2}\left[-\ell''(m)\right](w-m)^2 + \text{const}_2$$

$\ell(m)$ is const wrt w, so group w/ const

$\ell'(m) = 0$ bc. m is a maximer of $f(w)$ & $\ell(w)$ so this term cancels

this is a Gaussian pdf with mean $m = \arg\max_w \ell(w)$ and precision $\beta = -\ell''(m)$

Laplace Approx in M-Dim.

Given: random variable $w \in \mathbb{R}^M$
whose PDF is known up to norm. const.

$$p(w) = \frac{1}{Z} f(w) \longleftrightarrow \log p(w) = \log f(w) + \text{const}$$

We <u>know</u> $f(w) > 0$ and its $1^{st}/2^{nd}$ derivatives
wrt vector w

Idea: Approximation: $q(w) = \text{MVNorm}(\underset{\text{mean}}{m}, \underset{\text{precision matrix}}{\Lambda^{-1}})$: symmetric & positive definite!

Set m to match mode of f

$$m = \underset{w \in \mathbb{R}^M}{\arg\max}\ \ell(w) \qquad \ell(w) = \log f(w)$$

Set Λ to negative Hessian at mode

$$\Lambda = \nabla_w \nabla_w \left[-\ell(w) \right] \Big|_{w=m}$$

Similar Pro/Con as in 1-dim case above

Approximating the Posterior for Logistic Regression

True posterior intractable, but known to constant via Bayes

$$\log p(w|t_{1:N}) = \log p(w) + \log p(t_{1:N}|w) - \underbrace{\log p(t_{1:N})}_{\substack{\log Z \\ \text{constant} \\ \text{wrt } w}}$$

$$= \underbrace{\log \text{MVNormPDF}(w|m_0, S_0) + \sum_{n=1}^{N} \log \text{BernPMF}(t_n|\sigma(w^T\phi(x_n)))}_{\ell(w)}$$

Thus, we can apply Laplace Approx!

$$p(w|\{x_n, t_n\}_{n=1}^N) \approx \mathcal{N}(w|m_{MAP}, S)$$

with mean vector $m_{MAP} = \underset{w \in \mathbb{R}^M}{\operatorname{argmax}} \ \ell(w)$

solved w/ Gradient ~~descent~~

and precision matrix $S^{-1} = \nabla_w \nabla_w [-\ell(w)] \Big|_{w=m_{MAP}}$

uses plug-in formula

Using formulas for the Hessian of MAP objective, we know:

$$S^{-1} = S_0^{-1} + \Phi^T R(m_{MAP}) \Phi$$

$$= \alpha I_M + \Phi^T R(m_{MAP}) \Phi$$

Recall that $\Phi^T R \Phi = \sum_{n=1}^{N} \underbrace{r(m_{MAP}, x_n)}_{\text{positive scalar}} \underbrace{\phi(x_n)\phi(x_n)^T}_{\substack{M \times M \\ \text{outer product}}}$

$= \sigma(m_{MAP}^T \phi_n)(1 - \sigma(m_{MAP}^T \phi_n))$

Punchline: Laplace parameters $\underset{\text{mean}}{m}, \underset{\text{precision}}{S^{-1}}$ possible to compute

Posterior Predictive for Logistic Regr.

Ideal (intractable) posterior predictive:

$$p(t_* | t_{1:N}) = \int_{w \in \mathbb{R}^M} p(t_* | w)\, p(w | t_{1:N})\, dw$$

Laplace approximation

$$p(w | t_{1:N}) \approx \mathcal{N}(m_{MAP}, S)$$

$$\approx \int_{w \in \mathbb{R}^M} p(t_* | w)\, \mathcal{N}(w | m_{MAP}, S)\, dw$$

Still a tough integral over M dimensions
If M was 1 or 2, could use numerical strategies like trapezoid approx.

Option 1: Monte Carlo

Easy but need many samples L

$$p(t_* | t_{1:N}) = \mathbb{E}_{p(w|t_{1:N})}\left[p(t_* | w) \right]$$

Average of L samples from approx. posterior

$$\approx \frac{1}{L} \sum_{l=1}^{L} p(t_* | w^l)$$

with $w^l \sim \mathcal{N}(m_{MAP}, S)$

Option 2: Probit approx.

Closed-form, hard to port to other models

Likelihood $p(t_* | w) = \begin{cases} \sigma(w^T \phi_*) & \text{if } t_* = 1 \\ \sigma(-w^T \phi_*) & \text{if } t_* = 0 \end{cases}$

See Bishop Fig 4.9

$\sigma(a) \approx \text{NormCDF}(\sqrt{\frac{\pi}{8}} a)$

$$\approx \begin{cases} \text{NormCDF}(\sqrt{\frac{\pi}{8}}\, w^T \phi_*) & \text{if } t_* = 1 \\ 1 - \text{above} & t_* = 0 \end{cases}$$

Makes integral tractable when combined w/ Laplace!

(4.155): $p(t_* = 1 | t_{1:N}) = \sigma\left(m_{MAP}^T \phi_* \cdot \frac{1}{\sqrt{1 + \frac{\pi}{8} \phi_*^T S \phi_*}} \right)$