SPR Day 11

Goals: Review gradient + Hessian
for linear & logistic regr.

Review 1st and 2nd order
gradient descent methods

Show both LinRegr and LogisticRegr are CONVEX

Setup Bayesian Logistic Regression

Joint Model

$$p(t, w|x) = p(w) \prod_{n=1}^{N} \text{Bern}\left(t_n | \sigma(w^T \phi(x_n))\right)$$

$p(w)$: normal prior

$\text{Bern}\left(t_n | \sigma(w^T \phi(x_n))\right)$: Bernoulli likelihood

Peek at Laplace Approximation

1st and 2nd order derivatives of Maximum Lik. objective for both Linear & Logistic Regr.

Φ : N×M feature matrix

t : N×1 output vector

$$\mathcal{L} = \sum_n \log p(t_n | x_n, w)$$

Linear Regr

$$\hat{y} = \Phi w$$

predictions — actual labels

gradient $$\nabla_w \mathcal{L} = \Phi^T (\Phi w - t)$$

Hessian $$\nabla_w \nabla_w \mathcal{L} = \Phi^T \Phi$$

always constant wrt to w

will be positive definite & thus invertible iff $N \geq M$ and full rank

Logistic Regression

apply sigmoid elementwise to vector Φw

gradient $$\nabla_w \mathcal{L} = \Phi^T (\sigma(\Phi w) - t)$$

Hessian $$\nabla_w \nabla_w \mathcal{L} = \Phi^T R(w) \Phi = \sum_n \hat{y}_n(w)(1 - \hat{y}_n(w)) \phi_n \phi_n^T$$

(scalars, vectors)

Let $R(w)$ be N×N diagonal matrix where entry n, n is

$$[R(w)]_{nn} = \hat{y}(w, x_n)(1 - \hat{y}(w, x_n))$$

$$= \sigma(w^T \phi(x_n)) \, \sigma(-w^T \phi(x_n))$$

always > 0

Because R has positive diagonal, Log Reg. Hessian will be invertible if $\Phi^T \Phi$ is invertible and pos. definite

Gradients + Hessians of Linear + Logistic Regr.

$$\alpha(w) = -\sum_{n=1}^{N} \log p(t_n | x_n, w)$$

Linear Regr.

gradient $\nabla_w \alpha = \Phi^T(\Phi w - t)$ — $M \times 1$ vector

Hessian $\nabla_w \nabla_w \alpha = \Phi^T \Phi$ — $M \times M$ matrix

Logistic Regr.

gradient $\nabla_w \alpha = \Phi^T(\sigma(\Phi w) - t)$

$\sigma(\Phi w)$: predicted mean

Hessian $\nabla_w \nabla_w \alpha = \Phi^T R \Phi$ — Bishop 4.98

R: $N \times N$ diagonal matrix

each diagonal entry $= y_n(1-y_n)$

$$= \begin{bmatrix} \sigma(w^T\phi(x_1))\,\sigma(-w^T\phi(x_1)) & & \\ & \ddots & \\ & & \sigma(w^T\phi(x_N))\,\sigma(-w^T\phi(x_N)) \end{bmatrix}$$

Question: Is this objective function convex? Does it have unique global minima? ↔ same!

Punchline: Logistic Regression ML estimation is convex

Function is convex if Hessian is positive definite for all w.

Can show for logistic regression: $u^T H u = (u^T \Phi^T \sqrt{R})(\sqrt{R}\Phi u) = v^T v \geq 0$ thus, is p.d.

From 1st order to 2nd order gradient methods

For any loss $\mathcal{L}(w)$ let $g(w) = \nabla_w \mathcal{L}$ and $H(w) = \nabla_w \nabla_w \mathcal{L}$

First order update:

$$w^{new} \leftarrow w^{old} - \varepsilon\, g(w^{old})$$

step size $\varepsilon > 0$

Second order update (Newton-Raphson)

$$w^{new} \leftarrow w^{old} - H^{-1}(w^{old})\, g(w^{old})$$

When loss is quadratic, Newton update gives optimal w^* in one step

Example: Linear Regression (use defn of g & H from prev. page)

$$\begin{aligned} w^{new} &\leftarrow w^{old} - H^{-1} g \\ &\leftarrow w^{old} - (\Phi^T\Phi)^{-1}(\Phi^T\Phi\, w^{old} - \Phi^T t) \\ &\leftarrow w^{old} - w^{old} + (\Phi^T\Phi)^{-1}\Phi^T t \\ &\leftarrow (\Phi^T\Phi)^{-1}\Phi^T t \end{aligned}$$

"least squares" optimal solution (maximizes likelihood)

For logistic regression, 2nd order methods are gold standard (but still require many iterations)

$$w^{new} \leftarrow w^{old} - (\Phi^T R \Phi)^{-1}\Phi^T(\hat{y} - t)$$

remember $R(w)$ and $\hat{y}(w)$ depend on w

$$\leftarrow (\Phi^T R \Phi)^{-1}(\Phi^T R \Phi\, w^{old} - \Phi^T \hat{y} + \Phi^T t)$$

$$\leftarrow (\Phi^T R \Phi)^{-1}\Phi^T R z \quad \text{where } z = \Phi w^{old} - R^{-1}(\hat{y} - t)$$

looks like least squares with "per-example weights" $R_{nn} = \hat{y}_n(1 - \hat{y}_n)$

Algorithm: 2nd order GD using Hessian (aka Newton method) is known as "Iteratively Reweighted Least Squares" (IRLS)

↳ update r_n given w ↰ until converged!
↳ update w given r_n

Bayesian Logistic Regression

Model:

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

$$p(w) = \mathcal{N}(w \mid \underbrace{m_0}_{\text{mean}}, \underbrace{S_0}_{\text{covar}})$$ just like linear regr.

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

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

iid across examples

Goals are to estimate the posterior and predictive

Posterior: $p(w \mid \{x_n, t_n\}_{n=1}^N)$ no closed form! Not a Gaussian!

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

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

likelihood; posterior

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 $p(w)$?

Idea: Approximate with a Gaussian: $q(w) = N(m, \beta^{-1})$

- pick mean to match the mode of $p(w)$

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

Can use Gradient Methods to solve this numerically

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

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

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

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

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

Derivation of Taylor approx to density $p(w)$ at $m = \operatorname{argmax}_w p(w)$

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

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!

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

($\ell'(m) = 0$)

2nd 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 $m = \operatorname{argmax}_w \ell(w)$

and precision $\beta = -\ell''(m)$

Approximating the Posterior for Logistic Regression

True posterior intractable:

$$p(w \mid \{x_n, t_n\}_{n=1}^N) = \frac{1}{Z} \cdot \prod_{n=1}^N \text{Bern}(t_n \mid \sigma(w^T \phi(x_n))) \cdot \text{Normal}(w \mid m_0, S_0)$$

But we can apply Laplace Approx!

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

where $m_{MAP} = \underset{w \in \mathbb{R}^M}{\text{argmax}} \log p(w \mid \{x_n, t_n\}_{n=1}^N)$

solved w/ Gradient descent

and precision $S^{-1} = -\ell''(m_{MAP})$

where $\ell'' = \nabla\nabla \log p(w \mid x, t)$ = Hessian of log likelihood + Hessian of log prior

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

Hessian of neg. log prior = S_0^{-1}

Hessian of neg. log likelihood = $\Phi^T R(m_{MAP}) \Phi$

$$= \sum_{n=1}^N r(m_{MAP}, x_n) \phi(x_n) \phi(x_n)^T$$

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