Part 3:

Generalized linear models

Logistic regression

Approximate inference: Laplace method

3.1. Generalized linear models

$\vec{x} \in \mathbb{R}^d$ - input, y - target

$D = \{(\vec{x}_i, y_i)\}_{i=1}^n$ - a sample

Goal: get $p(y \mid \vec{x})$

General model:

$$\lambda = \lambda(\vec{x}; \vec{\theta}) = \vec{x}^T \vec{\theta} \text{ - linear link f-on}$$

$$p(y \mid \vec{x}) = p(y \mid \lambda)$$

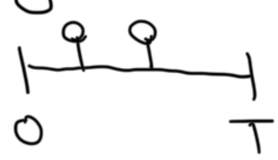
A. Linear regression

$y \in \mathbb{R}$, $\lambda = \vec{\theta}^T \vec{x}$

$y \sim N(\lambda, \sigma^2)$, σ^2 is known / not important

B. Poisson regression

$y \in \mathbb{N} \cup \{0\}$ - number of events



Intensity $\exp(\lambda) > 0$

$y \sim Pois(\exp(\lambda))$

C. Logistic regression

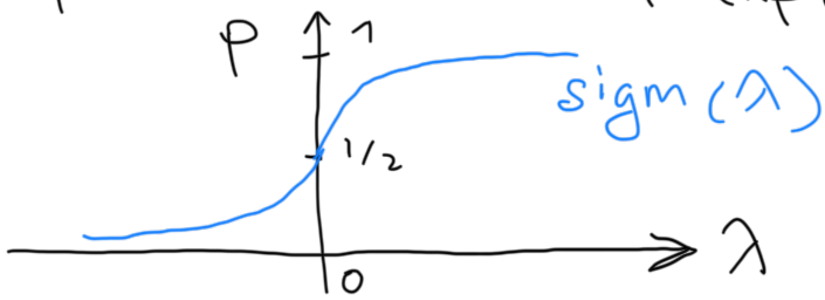
$y \in \{0, 1\}$

$y \sim Bern(p)$ - Bernoulli distrib.

$y = 0$ with probability p

$y = 1$ —||— $1 - p$

$$p = sigm(\lambda) = \frac{1}{1+\exp(-\lambda)} \in [0,1]$$



$sigm(\frac{1}{2}) = \frac{1}{2}$

$sigm(\lambda) \to 1$ if $\lambda \to +\infty$

$sigm(\lambda) \to 0$ if $\lambda \to -\infty$

$$\lambda = \vec{\Theta}^T \vec{x}$$

Data generation process:

$$\boxed{\cdot\vec{x}} \xrightarrow{\vec{\Theta}} \lambda \xrightarrow{sigm(\lambda)} p \to y \in \{0,1\}$$

GLMs are:

- general
- powerful

How to estimate $\vec{\Theta}$ from D for GLMs?

3.2. MLE – maximum likelihood estimate for logistic regression

$$y \in \{0, 1\}$$

$$p = p(y=0 \mid \vec{x}) = \sigma(\vec{x}^T \vec{\Theta}), \quad p_i = \sigma(\vec{x}_i \vec{\Theta})$$

$$p(y \mid \vec{x}, \vec{\Theta}) = p^y (1-p)^{1-y}$$

$$p(\vec{y} \mid X, \vec{\Theta}) = \prod_{i=1}^{n} p_i^{y_i} (1-p_i)^{1-y_i},$$

$$\log p(\vec{y} \mid X, \vec{\Theta}) = \sum_{i=1}^{n} y_i \log p_i + (1-y_i) \log(1-p_i)$$

$$\frac{\partial}{\partial\theta}\log p(y|X,\vec{\theta}) = \sum_{i=1}^{n}(p_i - y_i)\vec{x}_i = 0$$

depends on $\vec{\theta}$ in nonlinear way

3.3. Newton method for optimization

II-order method

$\vec{\theta}^{(0)}$ - initial guess

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} + H^{-1}\nabla \ln p(\vec{y}|X,\vec{\theta}) \quad (1)$$

instead of 'learning rate we use inverse Hessian matrix

$$H = \left\{ \frac{\partial^2 \ln p(\vec{y}|X,\vec{\theta})}{\partial\theta_i \partial\theta_j} \right\}$$

Statement We solve quadratic problem in one iteration via Newton's method

$$L(\vec{\theta}) = (\vec{\theta} - \vec{\mu})^T A (\vec{\theta} - \vec{\mu}) \to \min_{\vec{\theta}},$$

$$A \succeq 0$$

$\vec{\theta}^{(0)}$ - initial guess

$$\vec{\theta}^{(t+1)} = \vec{\theta}^{(t)} - H^{-1} \left.\frac{\partial L(\vec{\theta})}{\partial\vec{\theta}}\right|_{\vec{\theta}=\vec{\theta}^{(0)}}$$

Now we need $\frac{\partial L(\theta)}{\partial\vec{\theta}}$ and H

$$\begin{cases} \frac{\partial L(\vec{\theta})}{\partial\vec{\theta}} = 2A(\vec{\theta} - \vec{\mu}) \\ H = 2A, \quad H^{-1} = \frac{1}{2}A^{-1} \end{cases}$$

$$\vec{\theta}^{(1)} = \vec{\theta}^{(0)} - \frac{1}{2}A^{-1}\, 2A(\vec{\theta}^{(0)} - \vec{\mu}) =$$

$$= \vec{\theta}^{(0)} - \vec{\theta}^{(0)} + \vec{\mu} = \vec{\mu}$$

Let's apply Newton's method for the optimization for parameters of Logistic regression

$$\ln p(\vec{y} \mid \vec{\theta}) = \sum_{i=1}^{n} y_i \log p_i + (1-y_i)\log(1-p_i)$$

$$p_i = \frac{1}{1+\exp(-\vec{x}_i^T\vec{\theta})}; \quad \vec{x}_i^T \vec{\theta} = t_i$$

$$\vec{p} = (p_1, \dots, p_n)^T$$

$$\frac{\partial p_i}{\partial \theta_j} = -\sigma^2(t_i)\exp(-t_i)(-x_{ij})$$

$$= \sigma(t_i)(1-\sigma(t_i))\, x_{ij}$$

$$= p_i(1-p_i)\, x_{ij}$$

$$y_i = 1: \ \frac{\partial}{\partial\theta_j} p(y_i \mid \vec{\theta}) = \frac{1}{p_i}\cdot\frac{\partial p_i}{\partial \theta_j} = $$

$$= (1-p_i)\, x_{ij}$$

$$y_i = 0: \ \frac{\partial}{\partial\theta_j} p(y_i \mid \vec{\theta}) = -\frac{1}{1-p_i}\,\frac{\partial p_i}{\partial\theta_j} = $$

$$= (0-p_i)\, x_{ij}$$

$$\nabla \ln p(\vec{y} \mid \vec{\theta}) = (\vec{p}-\vec{y})^T X,$$

Hessian $\quad H = X^T \underline{R}\, X,$

$$R = \operatorname{diag}(\vec{p}(1-\vec{p})) = \begin{pmatrix} p_1(1-p_1) & & 0 \\ & \ddots & \\ 0 & & p_n(1-p_n) \end{pmatrix}$$

$$\vec{\theta}^{(new)} = (X^T R X)^{-1} X^T R \vec{z}$$

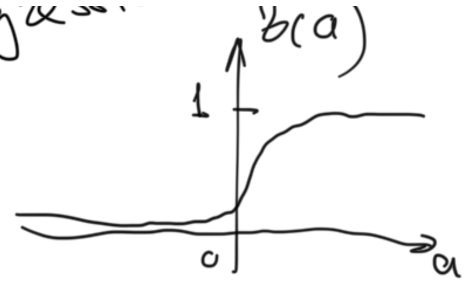
$p_i \in (0,$

$p_i = \sigma(\vec{x}_i^T \vec{\theta})$

$$\vec{z} = X\vec{\theta}^{(old)} - R^{-1}(\vec{p}-\vec{y})$$

$$\vec{\theta}^{(new)} = (X^T X)^{-1} X^T \vec{y}$$ - for linear ...ression

Θ_{MLE} ( ) , regu…



$$\forall \vec{v}: \vec{v}^T H \vec{v} = \underbrace{\vec{v}^T X^T} R X \vec{v} = \underbrace{\vec{v}'^T} \underbrace{R} \vec{v}' = \sum_{i=1}^{d} r_{ii} (v'_i)^2 \geq 0,$$

$\Rightarrow$ H is nonnegative-definite

$\Rightarrow$ the optimization problem is convex

3.4. Probit regression

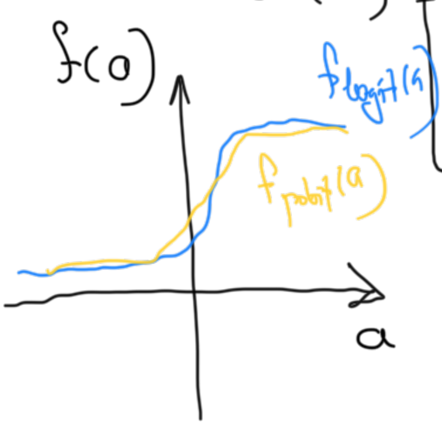
$$p(y=0 \mid a) = f(\underbrace{a}), \quad a = \vec{\Theta}^T \vec{x}$$

$$\boxed{a_i = \vec{\Theta}^T \vec{x}_i} \qquad \boxed{\underbrace{t} \sim p(\cdot)} \;\rightarrow\; \boxed{y_i = I[a_i > t]}$$

$$f(a) = \int_{-\infty}^{a} p(t)\,dt \quad (-\infty, a_i]$$

$$f(a) \begin{cases} \Phi(a) = \int_{-\infty}^{a} \underbrace{N(t \mid 0, 1)}\,dt \end{cases}$$ - probit regression

- no analytical formula for $\Phi(a)$



$$\begin{cases} p_{log}(t) = \dfrac{\exp(-t)}{(1+\exp(-t))^2} \\ f(a) = \sigma(a) = \dfrac{1}{1+\exp(-a)} \end{cases}$$ - logistic regression

---

3.5. Approximate Bayesian Inference: idea

$$\underbrace{p(\Theta \mid D)} \propto \underbrace{p(D \mid \Theta)}\,\underbrace{\pi(\Theta)}$$

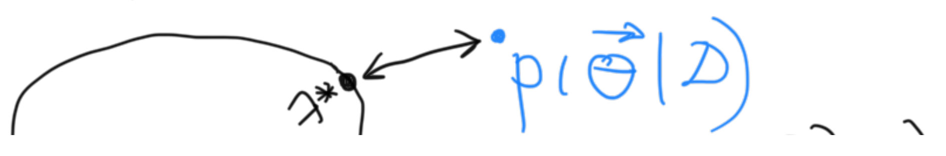
$$\underbrace{p(D)} = ? = \int p(D \mid \Theta)\, \pi(\Theta)\, d\Theta$$

$$p(y \mid D) = \int p(y \mid \vec{x}, \vec{\Theta})\, p(\vec{\Theta} \mid D)\, d\vec{\Theta} = ?$$

$$\underbrace{p(\vec{\Theta} \mid D) \approx q(\vec{\Theta})}$$ - how?

$$p(y \mid D) \approx \int p(y \mid \vec{x}, \vec{\Theta})\, q(\vec{\Theta})\, d\vec{\Theta}$$

$q_{\lambda}(\vec{\Theta})$



$q_{\lambda^*}(\vec{\theta}) \approx p(\vec{\theta}|\mathcal{D})$

3.6. Laplace approximation for approximate Bayesian inference

The problem: we know

$f(z)$ - unnormalized density

but we want to know:

$\boxed{p(z) = \frac{1}{C} f(z)}$, $z \in \mathbb{R}$ (or $\mathbb{R}^d$)

$1 = \int p(z)\, dz = \int \frac{1}{C} f(z)\, dz$

$C = \int f(z)\, dz$ - but we can't evaluate the integral analytically

Ex. Bayesian inference

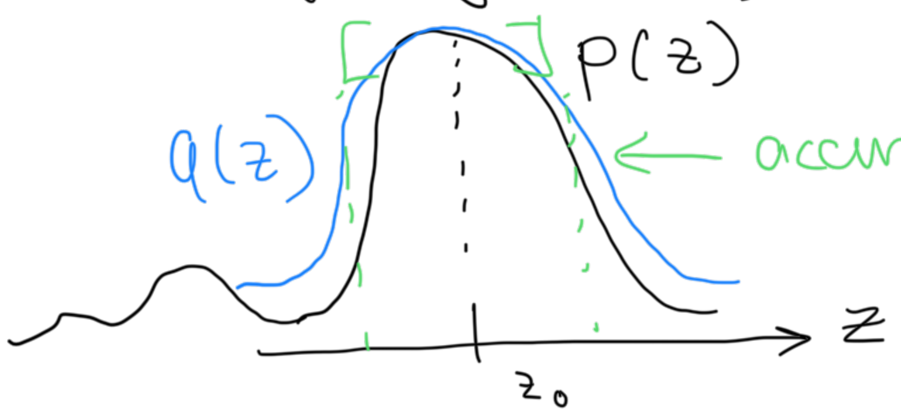
$$p(\theta | \mathcal{D}) = \frac{1}{p(\mathcal{D})} p(\mathcal{D}|\theta)\, p(\theta)$$

$$f(\theta) = p(\mathcal{D}|\theta)\, p(\theta), \quad C = p(\mathcal{D})$$

Solution: find $q(z)$ s.t.

$$q(z) \approx p(z)$$

Idea: $q(z) \approx p(z)$ for the areas of high (max) density of $f(z)$



(*) 1. $\left.\frac{\partial f(z)}{\partial z}\right|_{z=z_0} = 0 \rightsquigarrow z_0$ - the mode of $p(z)$

$$\frac{\partial}{\partial z}\Big|_{z=z_0}$$

2. $$\ln f(z) \approx \ln f(z_0) - \frac{1}{2} A (z - z_0)^2,$$

– Taylor decomposition for $\log f(z)$ around z_0

- No first-order term because of $(*)$
- $A = -\frac{\partial^2}{\partial z^2} \ln f(z)\Big|_{z=z_0}$

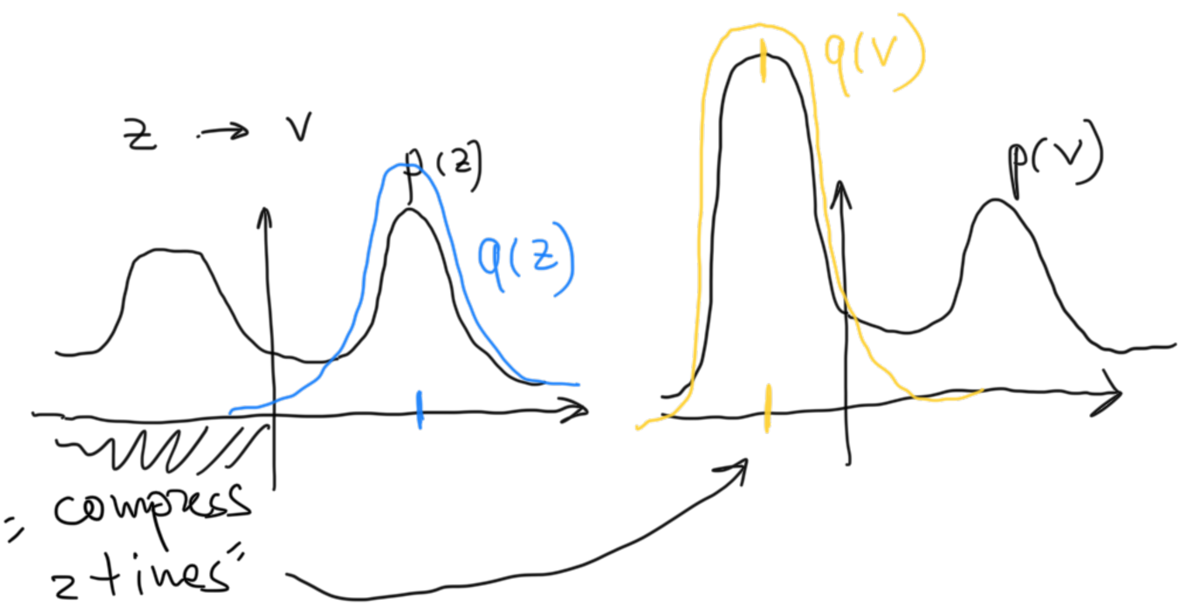
$$f(z) \approx \underbrace{f(z_0)} \exp\big(-\underbrace{\tfrac{1}{2} A (z - z_0)^2}\big) = q(z)$$

$$C = \sqrt{2\pi}\, |A|^{-1/2} / f(z_0)$$

$$q(z) = \mathcal{N}(z \mid z_0, A^{-1})$$

<u>Why:</u> BvM theorem states that for larger samples the posterior is close to Gaus.

<u>Why not:</u>
1. Local approach
2. Unimodal
3. Depends on parametriz.



$$v = \begin{cases} z, & z > 0 \\ 1/2\, z, & z > 0 \end{cases}$$

<u>3.7.</u> <u>Bayesian information criterion</u> (BIC)

$$Z = \int f(z)\,dz \approx \int_{-\infty}^{+\infty} f(z_0) \exp\left(-\frac{1}{2} A (z - z_0)^2\right) dz =$$

$$= f(z_0) \int_{-\infty}^{+\infty} \exp\left(-\frac{1}{2} A (z - z_0)^2\right) dz =$$

$$= \underbrace{f(z_0) \frac{(2\pi)^{p/2}}{|A|^{1/2}}}, \quad \vec{z} \in \mathbb{R}^p$$

$$p(\mathcal{D}) = \int \underbrace{p(\mathcal{D}|\vec{\theta})\, p(\vec{\theta})}\, d\vec{\theta}$$

$$f(\vec{\theta}) = p(\mathcal{D}|\vec{\theta})\, p(\vec{\theta})$$

$$Z = p(\mathcal{D})$$

$$\ln p(\mathcal{D}) \approx \underbrace{\ln p(\mathcal{D}|\vec{\theta}_{MAP})} + \underbrace{\ln p(\vec{\theta}_{MAP})} +$$

$$+ \frac{p}{2} \ln(2\pi) - \frac{1}{2} \ln(\det(A)) \approx$$

$$|\mathcal{D}| = n \to \infty$$

$$\approx \underbrace{\ln p(\mathcal{D}|\vec{\theta}_{MAP})} + 0 + 0 \underbrace{- \frac{1}{2} p \ln(n)}$$

A - diagonal $\mathcal{D}$ - i.i.d. observations

$$A = \begin{pmatrix} \frac{\partial^2 \ln p(\mathcal{D}|\vec{\theta})}{\partial \theta_1^2} & & 0 \\ & \ddots & \\ 0 & & \frac{\partial^2 \ln p(\mathcal{D}|\theta)}{\partial \theta_p^2} \end{pmatrix}$$

$$\frac{\partial^2 \ln p(\mathcal{D}|\vec{\theta})}{\partial \theta_i^2} = \frac{\partial^2 \sum_{j=1}^{n} \ln p(y_j|\vec{\theta})}{\partial \theta_i^2} \approx n\, \mathbb{E} \frac{\partial^2 \ln p(y_j|\vec{\theta})}{\partial \theta_i^2}$$

$$\approx \underbrace{n \cdot c}$$

$$\ln \det(A) = \ln\left[(n \cdot c)^p\right] = p \ln n + c'$$

$$\text{BIC} = p \ln n$$

3.8. Bayesian Logistic Regression

$$p(\vec{\theta}) = \mathcal{N}(\vec{\theta} \,|\, \vec{\mu}_0, S_0)$$

$$p(\vec{\theta}|\mathcal{D}) \propto p(\mathcal{D}|\vec{\theta})\, p(\vec{\theta})$$

$$\ln p(\vec{\theta}|\mathcal{D}) \propto + \sum_{i=1}^{n} \left(y_i \ln p_i + (1-y_i) \ln(1-p_i)\right) -$$

$$- \frac{1}{2} (\vec{\theta} - \vec{\mu}_0)^T S_0^{-1} (\vec{\theta} - \vec{\mu}_0), \quad p_i = \sigma(\vec{\theta}^T \vec{x}_i)$$

$$p(\vec{\theta}|D) \approx q(\vec{\theta}) = N(\vec{\theta} \,|\, \underline{\vec{\theta}_{MAP}}, \underline{S_n})$$

$$S_n^{-1} = -\nabla\nabla \ln p(\vec{\theta}|D) = \sum_{i=1}^{n} p_i(1-p_i) \underbrace{\vec{x}_i \vec{x}_i^T}_{p \times p} + S_0^{-1}$$

Predictive distribution

$$p(y=0|\vec{x}, D) = \int p(y=0|\vec{x}, \vec{\theta})\, p(\vec{\theta}|D)\, d\vec{\theta} \approx$$

$$\approx \int \underbrace{p(y=0|\vec{x}, \vec{\theta})}_{\sigma(\vec{x}^T\vec{\theta})} \underbrace{q(\vec{\theta})}_{\text{Gaussian}} d\vec{\theta} =$$

- not analytical ☹

$$\sigma(\vec{x}^T\vec{\theta}) = \int \underbrace{\delta(a - \vec{\theta}^T\vec{x})}_{\text{delta function}} \sigma(a)\, da$$

$$= \underline{\int \sigma(a)\, p(a)\, da}, =$$

$$\underline{p(a)} = \int \delta(a - \vec{\theta}^T\vec{x})\, \underline{q(\vec{\theta})}\, d\vec{\theta}$$

$$p(a) = N(a|\mu_a, \sigma_a^2)$$

$$\mu_a = \mathbb{E}a = \int p(a)\, a\, da = \int q(\vec{\theta})\, \vec{\theta}^T\vec{x}\, d\vec{\theta} = \vec{\theta}_{MAP}^T \vec{x}$$

$$\sigma_a^2 = \mathrm{Var}[a] = \int p(a)\{a^2 - (\mathbb{E}[a])^2\}\, da =$$

$$= \int q(\vec{\theta})\{(\vec{\theta}^T\vec{x})^2 - (\vec{\theta}_{MAP}^T\vec{x})\}\, d\vec{\theta} =$$

$$= \vec{x}^T \underline{S_n^{-1}} \vec{x}$$

$$= \int \underline{\sigma(a)}\, N(a|\mu_a, \sigma_a^2)\, da \approx$$

$$\approx \int \underbrace{\Phi(\lambda a)}_{\text{CDF for standard Gaussian distribution}} N(a|\mu_a, \sigma_a^2)\, da =$$

$$= \Phi\left(\frac{\mu}{(\lambda^{-2} + \sigma^2)^{1/2}}\right) \approx \boxed{\sigma(\kappa(\sigma_a^2)\mu_a)},$$

$$\kappa(\sigma_a^2) = \sqrt{\frac{1}{1 + \pi\sigma_a^2/8}}$$