# Multiclass logistic regression — Lecture 9

Now consider $K > 2$:

$$p(C_k \mid \vec{\varphi}) \equiv y_k(\vec{\varphi}) = \frac{e^{a_k}}{\sum_{j=1}^{K} e^{a_j}}.$$

$$a_k = \vec{w}_k^T \cdot \vec{\varphi}$$

The goal is to determine $\{\vec{w}_k\}$ directly.

Use $\vec{t}_n = \{ t_{n,1} \dots \underbrace{t_{n,k}}_{=1 \text{ if } \vec{t}_n \in C_k \text{, all other entries } = 0} \dots t_{n,K} \}$

$$T = \overbrace{\begin{pmatrix} & & \\ & & \\ & & \end{pmatrix}}^{K} \Big\} N \qquad T_{nk} = t_{nk}$$

Then

$$\mathcal{L} = p(T \mid \vec{w}_1 \dots \vec{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} \underbrace{p(C_k \mid \vec{\varphi}_n)}_{= y_{nk}}^{t_{nk}} \quad \vec{\varphi}_n = \vec{\varphi}(\vec{x}_n)$$

$$= \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}.$$

$$E(\vec{w}_1, \dots, \vec{w}_K) = -\log \mathcal{L} = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \log y_{nk}.$$

Now, consider

$$\underbrace{\frac{\partial E}{\partial \vec{w}_j}}_{\text{vector derivative}} = -\sum_n \sum_k t_{nk} \frac{1}{y_{nk}} \underbrace{\frac{\partial y_{nk}}{\partial \vec{w}_j}}_{= \frac{\partial y_{nk}}{\partial a_j} \underbrace{\frac{\partial a_j}{\partial \vec{w}_j}}_{\vec{\varphi}_n}} =$$

note that $\frac{\partial a_{j'}}{\partial \vec{w}_j} = 0, \; j' \neq j$

$$\frac{\partial y_k}{\partial a_j} = y_k \delta_{kj} - \frac{e^{a_k}}{\left(\sum_{j'} e^{a_{j'}}\right)^2} e^{a_j} =$$

$$= y_k \delta_{kj} - y_k y_j .$$

$$\boxed{\equiv} - \sum_{n,k} \frac{t_{nk}}{y_{nk}} \left[ y_{nk} \delta_{kj} - y_{nk} y_{nj} \right] \vec{\varphi}_n =$$

$$= \sum_{n,k} \left[ t_{nk} y_{nj} - t_{nk} \delta_{kj} \right] \vec{\varphi}_n = \sum_n \left[ y_{nj} - \underline{\underline{t_{nj}}} \right] \vec{\varphi}_n .$$

$$\uparrow$$

$$\sum_k t_{nk} = 1 , \quad \forall n .$$

Moreover,

$$H = \overbrace{\frac{\partial E}{\partial \vec{w}_k \partial \vec{w}_j}}^{MK \times MK \text{ matrix}} = \sum_n \left[ \frac{\partial y_{nj}}{\partial \vec{w}_k} \right] \vec{\varphi}_n =$$

$$= \sum_n \left[ y_{nj} \delta_{kj} - y_{nk} y_{nj} \right] \underbrace{\vec{\varphi}_n \vec{\varphi}_n^T}_{M \times M \text{ matrix}}$$

Can show that $\underbrace{\vec{u}^T H \vec{u} > 0}_{\text{positive definite}} \Rightarrow$ unique minimum

Thus can do NR algorithm again.

## Probit regression

If desired, $\sigma(a)$ can be replaced by

$$\varphi(a) = \int_{-\infty}^{a} d\theta \, \mathcal{N}(\theta|0,1) \quad \leftarrow \text{cumulative gaussian, or probit function}$$

Indeed, consider $\sigma(a) = \dfrac{1}{1+e^{-a}}$ vs. $\varphi(\lambda a)$.

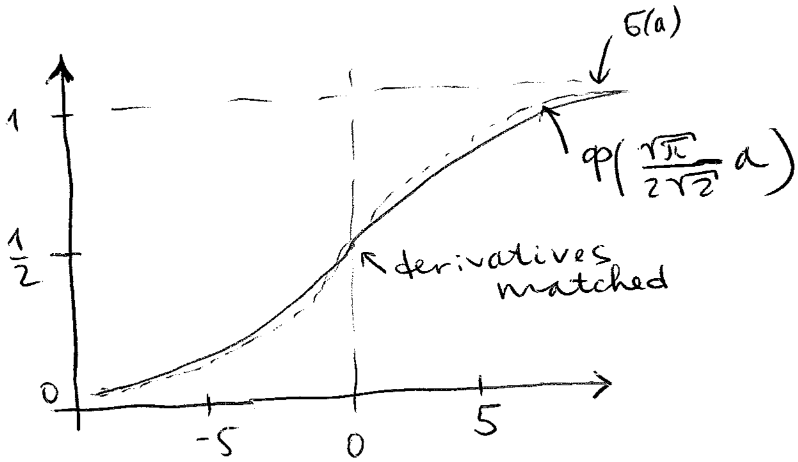
$$\left.\frac{d\sigma(a)}{da}\right|_{a=0} = \sigma(0)(1-\sigma(0)) = \frac{1}{4}.$$

But $$\left.\frac{d\varphi(\lambda a)}{da}\right|_{a=0} = \left.\frac{d\varphi(\lambda a)}{d(\lambda a)}\lambda\right|_{a=0} = \left.\frac{\lambda}{\sqrt{2\pi}} e^{-(\lambda a)^2}\right|_{a=0} = \underline{\underline{\frac{\lambda}{\sqrt{2\pi}}}}.$$

Match derivatives at $a=0$:

$$\frac{\lambda}{\sqrt{2\pi}} = \frac{1}{4} \Rightarrow \lambda^2 = \frac{\pi}{8}.$$

So, $\sigma(a) \approx \varphi\left(\frac{\sqrt{\pi}}{2\sqrt{2}} a\right)$:




Can repeat the ML analysis for the logistic regression with probit regression ⇒ similar results in practice.

# Laplace approximation

(saddle-point appr'n)

Consider

$$p(z) = \frac{f(z)}{Z}$$

$$\int dz\, p(z) = 1 \Rightarrow Z = \int dz\, f(z).$$

$$f(z) \simeq f(z_0)\, e^{-\frac{A}{2}(z-z_0)^2}$$

$$A = -\frac{d^2}{dz^2} \log f(z)\Big|_{z_0} \qquad \underbrace{\frac{df(z)}{dz}\Big|_{z_0} = 0}_{@\ \text{max}}$$

$$\log f(z) \simeq \log f(z_0) - \frac{1}{2}A(z-z_0)^2$$

Then $p(z) \Rightarrow q(z) = \sqrt{\frac{A}{2\pi}}\, e^{-\frac{A}{2}(z-z_0)^2}.$

Likewise, $f(\vec{z}) \simeq f(\vec{z}_0)\, e^{-\frac{1}{2}(\vec{z}-\vec{z}_0)^T A(\vec{z}-\vec{z}_0)}$ ← stationary point

$$p(\vec{z}) \Rightarrow q(\vec{z}) = \frac{|A|^{1/2}}{(2\pi)^{D/2}}\, e^{-\frac{1}{2}(\vec{z}-z_0)^T A(\vec{z}-\vec{z}_0)} =$$

$$= \mathcal{N}(\vec{z}\,|\,\vec{z}_0, A^{-1}).$$

$A = -\vec{\nabla}_{\vec{z}} \vec{\nabla}_{\vec{z}} \log f(\vec{z})\Big|_{\vec{z}_0}$ — $D \times D$ matrix

Note that

$$Z = \int d\vec{z}\, f(\vec{z}) \simeq \underbrace{f(\vec{z}_0) \frac{(2\pi)^{D/2}}{|A|^{1/2}}}_{\text{saddle-point approx'n}}$$

# Model comparison

Consider a set of models $\{M_i\}$ with prms $\{\vec{\theta}_i\}$. Define likelihood

$$p(\mathcal{D}|\vec{\theta}_i, M_i), \text{ then}$$

$$\underbrace{p(\mathcal{D}|M_i)}_{\text{model evidence}} = \int d\vec{\theta}_i \, p(\mathcal{D}|\vec{\theta}_i, M_i)\, p(\vec{\theta}_i|M_i)$$

Under saddle-point approximation,

$$f(\vec{z}) \Rightarrow p(\mathcal{D}|\vec{\theta}_i, M_i)\, p(\vec{\theta}_i|M_i):$$

$$p(\mathcal{D}|M_i) \simeq p(\mathcal{D}|\vec{\theta}_{i,MAP}, M_i)\, p(\vec{\theta}_{i,MAP}|M_i) \times$$

$$\times \frac{(2\pi)^{M/2}}{|A_i|^{1/2}},$$ where $M$ is the # model prms in model $M_i$ [strictly speaking, $M \to M_i$] and

$$A_i = -\vec{\nabla}_{\vec{\theta}} \vec{\nabla}_{\vec{\theta}} \Big[\log\big(\underbrace{p(\mathcal{D}|\vec{\theta}_i, M_i)\, p(\vec{\theta}_i|M_i)}_{Z\, p(\vec{\theta}_i|\mathcal{D}, M_i),\ \text{s.t.}}\big)\Big]\Big|_{\vec{\theta}_{i,MAP}} =$$

$$\log(\ldots) = \log p(\vec{\theta}_i|\mathcal{D}, M_i) + \underbrace{\log Z}_{\text{const}(\vec{\theta}_i)}$$

$$= -\vec{\nabla}_{\vec{\theta}} \vec{\nabla}_{\vec{\theta}} \Big[\log \underbrace{p(\vec{\theta}_i|\mathcal{D}, M_i)}_{\text{posterior}}\Big]\Big|_{\vec{\theta}_{i,MAP}}.$$

Finally,

$$\boxed{\log p(\mathcal{D}|M_i) \simeq \overbrace{\log p(\mathcal{D}|\vec{\theta}_{i,MAP}, M_i)}^{\log \mathcal{L}} + \underbrace{+\log p(\vec{\theta}_{i,MAP}|M_i) + \frac{M}{2}\log(2\pi) - \frac{1}{2}\log|A_i|}_{\text{"penalty" for model complexity}}}$$

Moreover, assume that priors are given by

$$p(\vec{\theta}_i \mid M_i) = \mathcal{N}(\vec{\theta}_i \mid \vec{0}, \alpha^{-1}\mathbb{1}) \quad \text{as before.}$$

Then

$$A_i = \overbrace{-\vec{\nabla}\vec{\nabla}\left[\log p(\mathcal{D} \mid \vec{\theta}_i, M_i)\right]\Big|_{\vec{\theta}_{i,MAP}}}^{\cong H}$$

$$\underbrace{-\vec{\nabla}\vec{\nabla}\left[\log p(\vec{\theta}_i \mid M_i)\right]}_{\alpha\mathbb{1}}\Big|_{\vec{\theta}_{i,MAP}} = H + \alpha\mathbb{1} \simeq H$$

if $\alpha$ small or we have $\underbrace{\text{lots of data (}N\text{ is large)}}$

note that $\log p(\mathcal{D} \mid \vec{\theta}_i, M_i)$ has $\sum_{n=1}^{N} \dots$ and $\uparrow$ as $N \uparrow$

We have:

$$\log p(\mathcal{D} \mid M_i) \simeq \log p(\mathcal{D} \mid \vec{\theta}_{i,MAP}, M_i) -$$

$$- \frac{\alpha}{2}\, \cancel{\vec{\theta}_{i,MAP}^{T}\, \vec{\theta}_{i,MAP}} - \frac{1}{2}\log|H| \quad \text{in the large-}N\text{ limit}$$

same argument as above ($\alpha$ small and/or $N$ large)

Finally,

$$H = \sum_{n=1}^{N} H_n = N \underbrace{\langle H \rangle}_{\frac{1}{N}\sum_n H_n} \Rightarrow \log|H| = \log|N\langle H\rangle| =$$

$$= \log\left[N^M |\langle H \rangle|\right] =$$

$$= M \log N + \cancel{\log|\langle H \rangle|}$$

We obtain:

$$\left[\log p(\mathcal{D} \mid M_i) \simeq \log p(\mathcal{D} \mid \vec{\theta}_{i,MAP}, M_i) \underbrace{- \frac{M}{2}\log N}_{\text{complexity penalty}}\right]$$

Bayesian information criterion (BIC)