Logistic regression

$\{(\bar{x}_n, t_n)\}_{n=1}^{N}, \quad t_n = \begin{cases} 1, & C_1 \\ 0, & C_2 \end{cases}$

$p(C_1|\bar{x}) = \sigma(\bar{w}^T\bar{x})$

$p(C_2|\bar{x}) = 1 - \sigma(\bar{w}^T\bar{x})$

$p(C_k|\bar{x}) = \dfrac{e^{\bar{w}_k^T\bar{x}}}{\sum_\ell e^{\bar{w}_\ell^T \bar{x}}}$

$$p(D|\bar{w}) = \prod_n p(t_n|\bar{w}, \bar{x}_n) = \prod_n \overbrace{\sigma(\bar{w}^T\bar{x})}^{\sigma_n}{}^{t_n} \left(1-\sigma(\bar{w}^T\bar{x})\right)^{1-t_n}$$

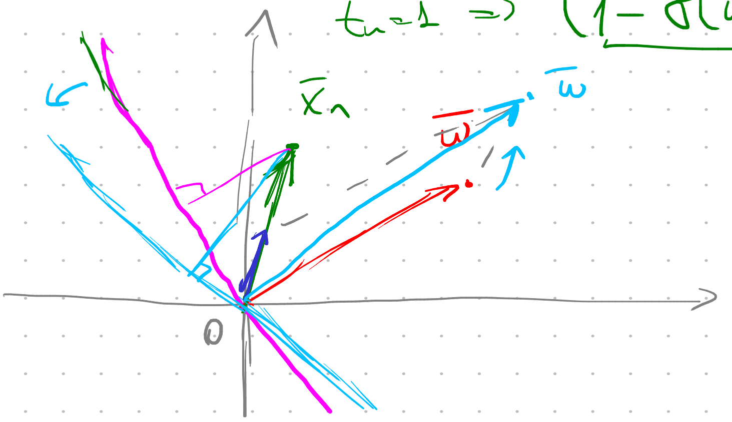
$$\nabla \log p(D|\bar{w}) = \sum_{n=1}^{N} \nabla_{\bar{w}}\left(t_n \log\sigma_n + (1-t_n)\log(1-\sigma_n)\right) =$$

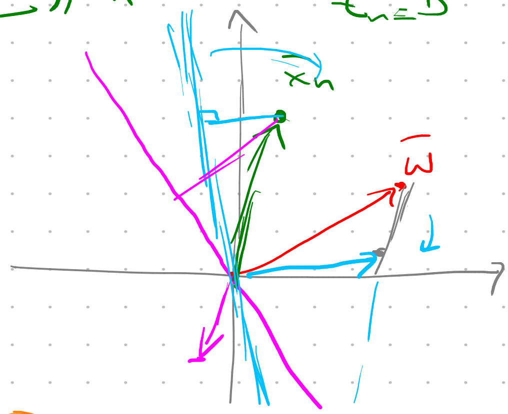
$$= \sum_n \left[t_n \cdot \frac{1}{\sigma_n}\cdot \sigma_n(1-\sigma_n)\cdot\bar{x}_n - (1-t_n)\cdot\frac{1}{\cancel{1-\sigma_n}}\sigma_n\cancel{(1-\sigma_n)}\bar{x}_n\right]$$

$p(C_1|\bar{x}) = \frac{1}{2} \Leftrightarrow \bar{w}^T\bar{x} = 0$

$$= \sum_n \left(t_n - t_n\sigma_n - \sigma_n + t_n\sigma_n\right)\bar{x}_n = \sum_{n=1}^{N}\left(\boxed{t_n} - \sigma(\bar{w}^T\bar{x}_n)\right)\bar{x}_n$$

$t_n = 1 \Rightarrow (1-\sigma(\bar{w}^T\bar{x}_n))\bar{x}_n$ $\qquad$ $t_n = 0 \quad -\sigma(\bar{w}^T\bar{x}_n)\bar{x}_n$





SGD stochastic gradient descent

$$\bar{w} := \bar{w} + \eta \cdot \nabla_{\bar{w}} \log p(t_n|\bar{w}, \bar{x}_n)$$

$2H(\bar{w}-\bar{w}_0)$

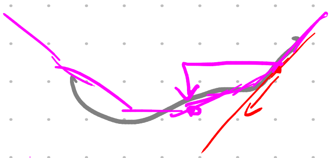
Hessian $\left(\dfrac{\partial^2 f}{\partial w_i \partial w_j}\right)_{ij}$

max

$$f(\bar{w}) \approx f(\bar{w}_0) + \nabla_{\bar{w}} f(\bar{w}_0)^T(\bar{w}-\bar{w}_0) + \frac{1}{2}(\bar{w}-\bar{w}_0)^T H (\bar{w}-\bar{w}_0)$$

$$\nabla_{\bar{w}} f|_2 = \nabla_{\bar{w}} f(\bar{w}_0) + H(\bar{w}_* - \bar{w}_0) = 0$$

$L = \frac{1}{2}(\bar{y}-X\bar{w})^T(\bar{y}-X\bar{w})$



$$\boxed{\bar{w}_* = w_0 - H^{-1}\cdot\nabla_{\bar{w}} f(\bar{w}_0)}$$

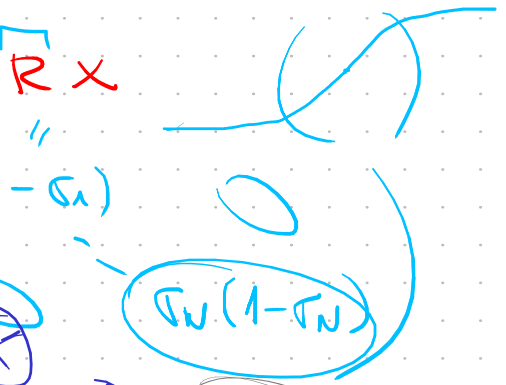
X^TX

$\bar{w}_* = \cancel{\bar{w}_0} + (X^TX)^{-1}(\cancel{X^TX\bar{w}_0} + X^T\bar{y})$

$$\frac{\partial^2 \log p}{\partial w_i \partial w_j} = \frac{\partial}{\partial w_i}\left(\sum_n \left(t_n - \sigma(\bar{w}^T\bar{x}_n)\right)x_{nj}\right) =$$

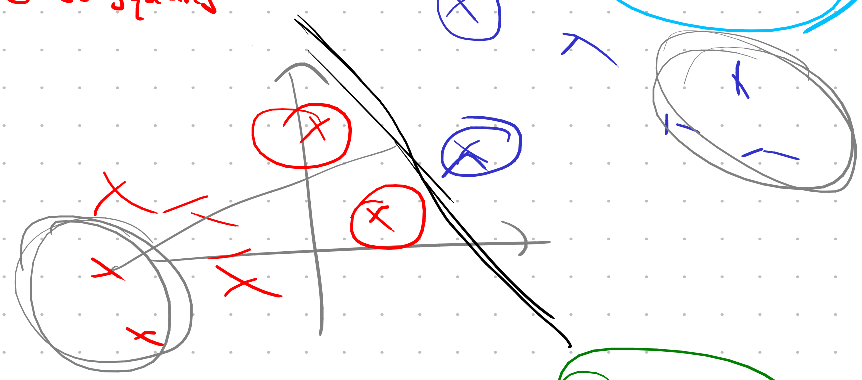
$$= \sum_n \left(-\sigma' \cdot x_{nj}\cdot x_{ni}\right) = -\sum_n \sigma(\bar{w}^T\bar{x}_n)(1-\sigma(\bar{w}^T\bar{x}_n))\, x_{ni} x_{nj}$$

$$H = -\sum_n \sigma_n\cdot(1-\sigma_n)\bar{x}_n\bar{x}_n^T = -X^T R X$$



$$\bar\omega^{(k+1)} := \bar\omega^{(k)} + (X^T R X)^{-1}\cdot X^T\left(\bar t - \bar\sigma\left(\bar\omega^{(k)T}\bar x\right)\right)$$

IRLS – iterative reweighted least squares



$$p(\bar\omega|D) \propto p(\bar\omega)\,p(D|\bar\omega)$$

$$\text{Const} - \frac{1}{2\sigma_0^2}\bar\omega^T\bar\omega$$

$$\log p(\bar\omega|D) = \text{Const} + \sum_n (t_n - \sigma_n)\bar x_n + \log p(\bar\omega)$$

$$\nabla \log p(\bar\omega|D) = \nabla \log p(D|\bar\omega) - \frac{1}{\sigma_0^2}\cdot\bar\omega$$

$$H' = H - \frac{1}{\sigma_0^2}\cdot I$$

one-hot

$$t_n = (0 - 1 - 0)$$

$$p(D|\bar\omega) = \prod_n\prod_k p(C_k|\bar\omega_k, \bar x_n)^{t_{nk}}$$

$$\log p(D|\bar\omega) = \sum_n\sum_k t_{nk}\log y_{nk}$$

$$y_{nk} = \frac{e^{\bar\omega_k^T\bar x_n}}{\sum_s e^{\bar\omega_s^T\bar x_n}}$$

$$y_k = \frac{e^{a_k}}{\sum e^{a_s}}$$

$$\frac{\partial y_k}{\partial a_j} = \begin{cases} \dfrac{e^{a_k}(\sum e^{a_s} - e^{a_k}\cdot e^{a_k})}{(\sum e^{a_s})^2} = y_k(1-y_k), & k=j \\ -\dfrac{e^{a_k}e^{a_j}}{(\sum e^{a_s})^2} = -y_k y_j, & k\neq j \end{cases}$$

$$\nabla_{\bar\omega_j}\log p(D|\bar\omega) = \sum_{n=1}^{N}\sum_{k=1}^{K} t_{nk}\frac{\cancel{y_{nk}}([k=j]-y_{nj})}{\cancel{y_{nk}}}\bar x_n$$

$$= \sum_n\sum_k t_{nk}([k=j]-y_{nj})\bar x_n = \sum_n\left[t_{nj}(1-y_{nj}) - \sum_{k\neq j} t_{nk}y_{nj}\right]\bar x_n$$

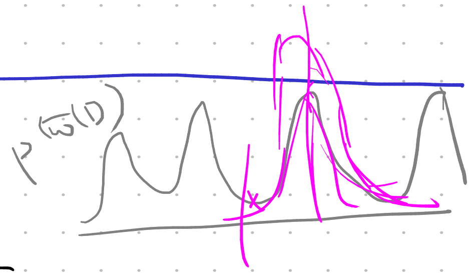
$$= \sum_n\left[t_{nj} - \Big(\underbrace{\sum_k t_{nk}}_{=1}\Big)y_{nj}\right]\bar x_n = \sum_n (t_{nj} - y_{nj})\bar x_n$$

$$\nabla_{\bar w_i}\left(\nabla_{\bar w_j}\dots\right) = \nabla_{\bar w_i}\left(\sum_n (t_{nj} - y_{nj})\bar x_n\right) =$$

$$= \begin{cases} -\sum_n y_{nj}(1-y_{nj})\bar x_n & i=j \\ -\sum_n y_{nj}\, y_{ni}\, \bar x_n & i\neq j \end{cases}$$

---

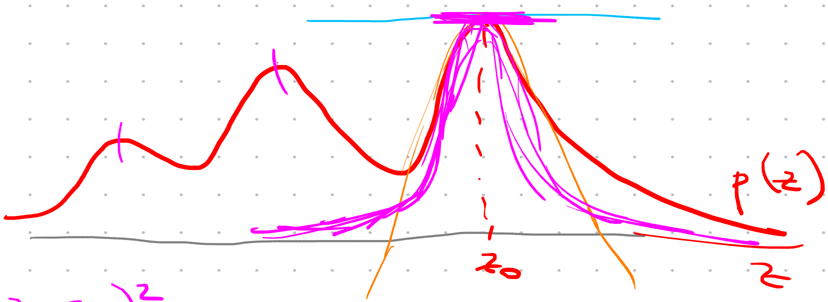
1) $\log p(D|\bar w) \to \max$
2) $\log p(\bar w|D) \to \max$



3) $$p(C_1|\bar x, D) = \int p(C_1|\bar w, \bar x)\cdot p(\bar w|D)\, d\bar w =$$

$$= \int \sigma(\bar w^T \bar x)\cdot p(\bar w|D)\, d\bar w = \mathbb{E}_{p(\bar w|D)}\left[\sigma(\bar w^T\bar x)\right]$$

Laplace approximation



$$\log p(z) \approx \log p(z_0) + \cancel{\frac{\partial \log p}{\partial z}\Big|_{z_0}(z-z_0)} + \frac{1}{2}\frac{\partial^2 \log p}{\partial z^2}\Big|_{z_0}(z-z_0)^2 =$$

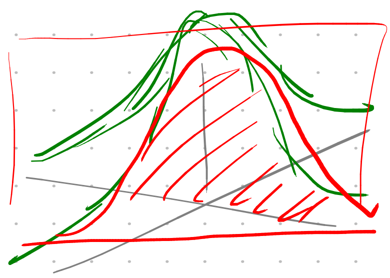
$$= \log p(z_0) - \frac{1}{2}A\cdot(z-z_0)^2, \quad \text{где } A = -\frac{\partial^2 \log p}{\partial z^2}\Big|_{z_0}$$

$$p(z) \approx p(z_0)\cdot e^{-\frac{1}{2}A(z-z_0)^2} \approx \mathcal{N}(z|z_0, 1/A)$$

$$\log p(\bar z) \approx \log p(\bar z_0) - \frac{1}{2}(\bar z - \bar z_0)^T A(\bar z - \bar z_0), \quad \text{где } A = -\nabla\nabla \log p$$

$$p(C_1|D, \bar x) = \int \sigma(\bar w^T \bar x)\, p(\bar w|D)\, d\bar w \approx [\bar w = \bar w_{MAP}] \approx$$

$$\approx \int \sigma(\underbrace{\bar w^T \bar x}_{a})\, \mathcal{N}(\bar w|\bar w_{MAP}, \Sigma_N)\, d\bar w =$$



$$\sigma(\bar w^T \bar x) = \int_{-\infty}^{\infty} \delta(\bar w^T \bar x - a)\cdot \sigma(a)\, da$$

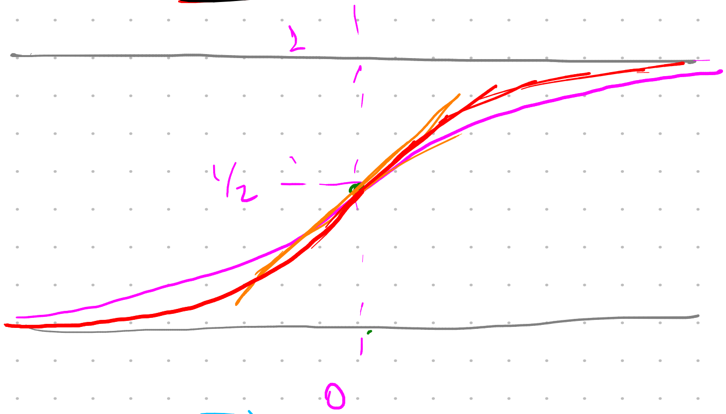
$$= \int_{-\infty}^{\infty}\int \sigma(a)\cdot\delta(\bar w^T\bar x - a)\cdot\mathcal{N}(\_)\,d\bar w\,da =$$

$$= \int_{-\infty}^{\infty}\sigma(a)\cdot\left(\int \delta(\bar w^T\bar x - a)\cdot\mathcal{N}(\bar w|\bar w_{MAP},\Sigma_N)\,d\bar w\right)da =$$

$$= \int_{-\infty}^{\infty} \underbrace{\sigma(a)}_{\simeq\Phi(\lambda a)}\cdot\mathcal{N}(a|\overbrace{\bar w_{MAP}^T\bar x}^{\mu_a}, \overbrace{\bar x^T\Sigma_N\bar x}^{\sigma_a^2})\,da =$$

probit regression

$p(C_1)=\Phi(\bar w^T\bar x)$



probit

$$\Phi(b)=\int_{-\infty}^{b}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx$$

$$\Phi(\lambda a)\approx\sigma(a)$$

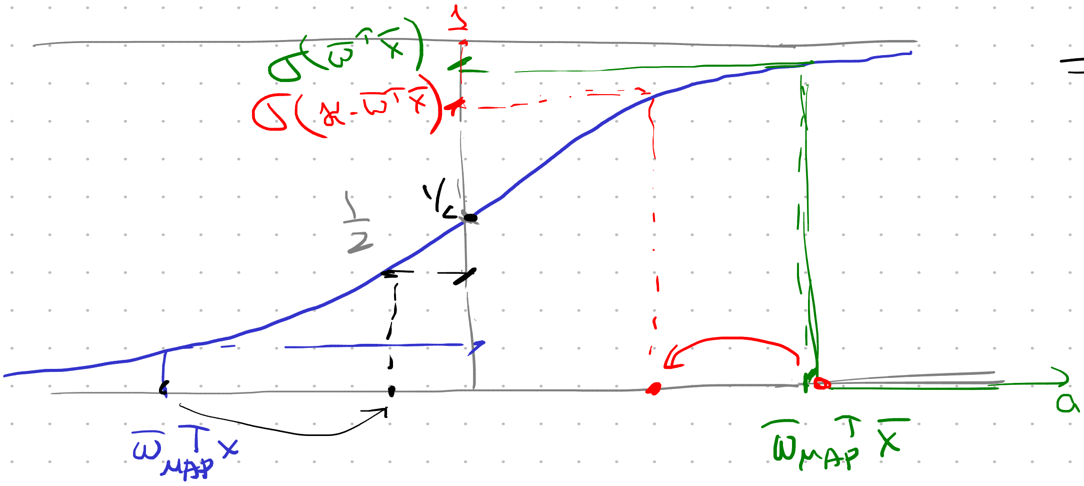
$$\lambda=\sqrt{\pi/8}$$

$\sigma(\mu_a)=\sigma(\bar w_{MAP}^T\bar x)$

$$=\int_{-\infty}^{\infty}\int_{-\infty}^{\lambda a}\mathcal{N}(x|0,1)\cdot\mathcal{N}(a|\mu_a,\sigma_a^2)\,dx\,da = y_{np}.$$

$\ll \bar w_{MAP}^T\bar x$

$$=\Phi\left(\frac{\mu_a}{\sqrt{\sigma_a^2+\frac{1}{\lambda^2}}}\right)=\Phi\left(\frac{\lambda\mu_a}{\sqrt{1+\lambda^2\sigma_a^2}}\right)\approx\sigma\left(\frac{\mu_a}{\sqrt{1+\lambda^2\sigma_a^2}}\right)$$



$$=\sigma\left(\frac{\mu_a}{\sqrt{1+\frac{\pi}{8}\sigma_a^2}}\right)$$