Notes de Cours de Machine Learning M1 MINT

Slide 20 régression logistique et softmax.

Logistique $\mathbb{P}(Y=1 / X=x) = \sigma(\langle x, w\rangle + b)$

Softmax $\mathbb{P}(Y=k / X=x) = \dfrac{\exp(\langle x, w_k\rangle + b_k)}{\sum_{k=1}^{K} \exp(\langle x, w_k\rangle + b_k)}$

$\sum_{k=1}^{K} \mathbb{P}(Y=k / X=x) = 1$

Cas où $K=2$ classification binaire.

$$\mathbb{P}(Y=1 / X=x) = \frac{\exp(\langle x, w_1\rangle + b_1)}{\exp(\langle x, w_{-1}\rangle + b_{-1}) + \exp(\langle x, w_1\rangle + b_1)}$$

$$= \frac{\exp(\langle x, w_1\rangle + b_1 - \langle x, w_{-1}\rangle - b_{-1})}{1 + \exp(\langle x, w_1\rangle + b_1 - \langle x, w_{-1}\rangle - b_1)}$$

$$= \frac{\exp(\langle x, w_1 - w_{-1}\rangle + b_1 - b_{-1})}{1 + \exp(\langle x, w_1 - w_{-1}\rangle + b_1 - b_{-1})}$$

Si on pose $w = w_1 - w_{-1}$ et $b = b_1 - b_{-1}$

$$= \frac{\exp(\langle x, w\rangle + b)}{1 + \exp(\langle x, w\rangle + b)} = \sigma(\langle x, w\rangle + b).$$

Conclusion : la softmax généralise la sigmoïde.

Slide 22

$$\frac{\mathbb{P}(Y=1 \mid X=x)}{\mathbb{P}(Y=-1 \mid X=x)} = \text{odds}.$$

$\mathbb{P}(Y=1) = \frac{1}{4}$

$\mathbb{P}(Y=-1) = 3/4$

$\text{odds} = \frac{1/4}{3/4} = \frac{1}{3}.$

$\mathbb{P}(Y=1) = 1/2$

$\mathbb{P}(Y=-1) = 1/2$

$\text{odds} = 1$

$\mathbb{P}(Y=1) = 1 \quad \mathbb{P}(Y=-1) = 0$

$\text{odds} = +\infty$

$\mathbb{P}(Y=1) = 0 \quad \mathbb{P}(Y=-1) = 1$

$\text{odds} = 0.$

Dans le cas de la régression logistique

$$\mathbb{P}(Y=1 \mid X=x) = \sigma(\langle x,w\rangle + b). \qquad \sigma : x \to \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

$$\text{Odds} = \frac{\sigma(\langle x,w\rangle + b)}{1 - \sigma(\langle x+w\rangle + b)}$$

$$= \frac{e^{\langle x,w\rangle + b} / 1+e^{\langle x,w\rangle + b}}{1 - \frac{e^{\langle x,w\rangle + b}}{1+e^{\langle x,w\rangle + b}}}$$

$$= \frac{e^{\langle x,w\rangle + b} / \cancel{1+e^{\langle x,w\rangle + b}}}{\frac{1 + \cancel{e^{\langle x,w\rangle + b}} - \cancel{e^{\langle x,w\rangle + b}}}{\cancel{1+e^{\langle x,w\rangle + b}}}}$$

$$= e^{\langle x,w\rangle + b}$$

i_1 et i_2 dont les features sont égales sauf la j

$$x_{i_1}^j = x_{i_2}^j + 1$$

$$x_{i_1}^k = x_{i_2}^k \quad \forall k \in [\![1, d]\!] \setminus \{j\}$$

$$\frac{odds(i_1)}{odds(i_2)} = \frac{\exp(\langle x_{i_1}, w\rangle + b)}{\exp(\langle x_{i_2}, w\rangle + b)}$$

$$= \exp\left(\langle x_{i_1}, w\rangle + \cancel{b} - \langle x_{i_2}, w\rangle - \cancel{b}\right)$$

$$= \exp\left(\langle x_{i_1} - x_{i_2}, w\rangle\right) = \exp\left(\sum_{k=1}^{d} (x_{i_1}^k - x_{i_2}^k) w_k\right)$$

$$= \exp\left((x_{i_1}^j - x_{i_2}^j) w_j\right) = \exp(w_j).$$

Slide 24

on choisit $\hat{Y}_+ = 1$

si $\mathbb{P}(Y=1 / X=x) \geq \mathbb{P}(Y=-1 / X=x)$

$$\iff \frac{e^{\langle x, w\rangle + b}}{1 + e^{\langle x, w\rangle + b}} \geq \frac{1}{1 + e^{\langle x, w\rangle + b}}$$

$$\iff e^{\langle x, w\rangle + b} \geq 1$$

$$\iff \langle x, w\rangle + b \geq 0$$

$$\iff \langle x, w\rangle \geq -b$$

$\longrightarrow$ règle de classification linéaire

l'espace des features est partagé par un hyperplan d'équation $\langle x, w\rangle + b = 0$.

Slide 27 et suivantes

On veut calculer $\frac{1}{n} \log \mathcal{L}$ ← vraisemblance.

Vraisemblance $= \mathcal{L}(y_1, \dots, y_n, x_1, \dots, x_n ; w, b) = \mathcal{L}$

$$= \prod_{i=1}^{n} \mathbb{P}(Y = y_i / X = x_i)$$

$$\mathbb{P}(Y = 1 / X = x_i) = \sigma(\langle x_i, w \rangle + b) = \frac{e^{\langle x_i, w \rangle + b}}{1 + e^{\langle x_i, w \rangle + b}}$$

$$\mathbb{P}(Y = -1 / X = x_i) = 1 - \sigma(\langle x_i, w \rangle + b)$$

Probit

$\phi(\langle x_i, w \rangle + b)$

$$= \frac{1}{1 + e^{\langle x_i, w \rangle + b}}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{\langle x_i, w \rangle + b}}{1 + e^{\langle x_i, w \rangle + b}} \right)^{\mathbb{1}(y_i = 1)} \left(\frac{1}{1 + e^{\langle x_i, w \rangle + b}} \right)^{\mathbb{1}(y_i = -1)}$$

$$= \prod_{i=1}^{n} \left(\frac{1}{e^{-(\langle x_i, w \rangle + b)} + 1} \right)^{\mathbb{1}(y_i = 1)} \left(\frac{1}{1 + e^{\langle x_i, w \rangle + b}} \right)^{\mathbb{1}(y_i = -1)}$$

$$= \prod_{i=1}^{n} \frac{1}{1 + e^{-y_i (\langle x_i, w \rangle + b)}}$$

$$\log \mathcal{L} = \sum_{i=1}^{n} \log \frac{1}{1 + e^{-y_i (\langle x_i, w \rangle + b)}}$$

$$= - \sum_{i=1}^{n} \log \left(1 + e^{-y_i \langle x_i, w \rangle + b} \right)$$

Empirical loss $= -\frac{1}{n} \log \mathcal{L} = \frac{1}{n} \sum_{i=1}^{n} \log \left(1 + e^{-y_i \langle x_i, w \rangle + b} \right)$

Définir les "bons" $\hat{w}$ et $\hat{b}$ au max. de vraisemblance comme

$$(\hat{w}, \hat{b}) \in \underset{w \in \mathbb{R}^d, b \in \mathbb{R}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + e^{-y_i(\langle x_i, w\rangle + b)}\right).$$

problème d'optimisation strict convexe et différentiable.

Slide 44 et suivantes.

$$\nabla f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell'(y_i, \langle x_i, w\rangle)\, x_i$$

$$\nabla^2 f(w) = \frac{1}{n} \sum_{i=1}^{n} \ell''(y_i, \langle x_i, w\rangle)\, x_i x_i^T$$

f est convexe $\Leftrightarrow$ ℓ l'est pour tout y $y' \mapsto \ell(y, y')$

L - régulière

$$\|\nabla f(w) - \nabla f(w')\|_2 \leq L \|w - w'\|_2.$$

si f est 2 fois diff.

L- régulière $\Longleftrightarrow$ $\lambda_{max}(\nabla^2 f(w)) \leq L$.

Logistique $\ell(y_i, \langle x_i, w\rangle) = \log\left(1 + e^{-y_i \langle x_i, w\rangle}\right)$

$$\ell'(y_i, \langle x_i, w\rangle) = \frac{-y_i x_i e^{\langle x_i, w\rangle}}{1 + e^{-y_i\langle x_i, w\rangle}}$$

$$= y_i \left(\sigma(y_i\langle x_i, w\rangle) - 1\right) x_i$$

$$\ell''(y_i, \langle x_i, w\rangle) = y_i\left(\sigma'(y_i\langle x_i, w\rangle)\right) \quad x_i$$

$$\sigma(x) = \frac{e^x}{1+e^x}$$

$$\sigma'(x) = \frac{e^x(1+e^x) - e^x e^x}{(1+e^x)^2}$$

$$= \frac{e^x + \cancel{e^{2x}} - \cancel{e^{+2x}}}{(1+e^x)^2} = \frac{e^x}{1+e^x} \cdot \frac{1}{1+e^x}$$

$$= \sigma(x)\left(1-\sigma(x)\right).$$

$$= \underset{\substack{\shortparallel \\ 1}}{y_i^2}\ \underbrace{\sigma(y_i\langle x_i, w\rangle)\left(1-\sigma(y_i\langle x_i, w\rangle)\right) x_i x_i^T}$$

$$\nabla^2 f(w) = \frac{1}{n}\sum_{i=1}^{n} \ell''(y_i, \langle x_i, w\rangle)$$

$$= \frac{1}{n}\sum_{i=1}^{n} \underbrace{\sigma(y_i\langle x_i, w\rangle)\left(1-\sigma(y_i\langle x_i; w\rangle)\right)}_{\sigma(y_i\langle x_i, w\rangle) = \mathbb{P}(Y = y_i / X = x_i)} x_i x_i^T$$

$x \mapsto \sigma(x)(1-\sigma(x))$ son maximum vaut $\frac{1}{4}$

$$\lambda_{max}\left(\nabla^2 f(w)\right) \leq \underbrace{\frac{1}{4n}\lambda_{max}\left(\sum_{i=1}^{n} x_i x_i^T\right)}_{L_{logistic}}$$

Si f est L-régulière :

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{L}{2} \|w - w'\|_2^2.$$

A l'itération :

$$f(w) \leq \underbrace{f(w^k) + \langle \nabla f(w^k), w - w^k \rangle + \frac{L}{2} \|w - w^k\|_2^2}_{(1)}.$$

descente de gradient $w^{k+1} = \arg\min \left((1) \right)$

On veut minimiser $f(w) + g(w)$

$$f(w) + g(w) \leq f(w^k) + \langle \nabla f(w^k), w - w^k \rangle + \frac{L}{2} \|w - w^k\|_2^2 + g(w))$$

$$w^{k+1} = \arg\min_w \cancel{f(w^k)} + \langle \nabla f(w^k), w - w^k \rangle + \frac{L}{2} \|w - w^k\|_2^2 + g(w)$$

$$\frac{L}{2} \| w - (w^k - \frac{1}{L} \nabla f(w^k)) \|_2^2$$

$$= \frac{L}{2} \|w - w^k\|_2^2 + \cancel{2} \frac{L}{\cancel{2}} \langle w - w^k, \frac{1}{\cancel{L}} \nabla f(w^k) \rangle + \frac{L}{2} \frac{1}{L^2} \| \nabla f(w)^k \|_2^2$$

$$w^{k+1} = \arg\min_w \underbrace{\frac{L}{2} \| w - (w^k - \frac{1}{L} \nabla f(w^k)) \|_2^2}_{} + g(w).$$

$$= \arg\min_w \frac{1}{2} \| w - \underbrace{(w^k - \frac{1}{L} \nabla f(w^k))}_{} \|_2^2 + \frac{1}{L} g(w)$$

Si $g = 0$.
on retombe sur la descente de gradient

$g : \mathbb{R}^d \to \mathbb{R}$ convexe (pas forcément diff.)
on définit son opérateur proximal.

$$\text{prox}_g(w) = \underset{w' \in \mathbb{R}^d}{\text{argmin}} \left\{ \frac{1}{2} \|w - w'\|_2^2 + g(w') \right\}$$

Slide S1 : prox du LASSO.

$z \in \mathbb{R}$, $z' \in \mathbb{R}$

$$\underset{z'}{\text{argmin}} \quad \underbrace{\frac{1}{2} (z' - z)^2 + \lambda |z'|}$$

sur $\mathbb{R}_+$ $\quad z' - z + \lambda$. le minimum est atteint pour $z' = z - \lambda$.

sur $\mathbb{R}_-$ $\quad z' - z - \lambda$.

le minimum est atteint $z' = z + \lambda$.

ce minimum est ≤ 0. seulement si $z \leq -\lambda$.

ce minimum est ≥ 0 seulement si $z \geq \lambda$.

Sur $[-\lambda, \lambda]$ le minimum est atteint en 0.

$$z^* = \text{argmin} \ \frac{1}{2} (z' - z)^2 + \lambda |z'| = \begin{cases} z - \lambda & \text{si } z \geq \lambda \\ z + \lambda & \text{si } z \leq -\lambda \\ 0 & \text{si } z \in [-\lambda, \lambda] \end{cases}$$

$$\boxed{z^*(z) = \text{sign}(z) \left(|z| - \lambda \right)_+}$$

si $z \geq \lambda$. $\quad z^*(z) = 1 \cdot (z - \lambda)_+ = z - \lambda$

si $z \leq -\lambda$ $\quad z^*(z) = (-1)(-z - \lambda)_+ = z + \lambda$.

si $z \in [-\lambda, \lambda]$ $\quad z^*(z) = \begin{pmatrix} 1 \\ -1 \end{pmatrix} \underbrace{\left(\underset{\leq \lambda}{|z|} - \lambda \right)_+}_{\leq 0} = 0$.

Exercice 1 de la slide 75

A une itération k $\quad w^k, b^k$

$$f(w,b) + g(w) \leq f(w^k,b^k) + \left\langle \nabla_{w,b} f(w^k,b^k), \begin{pmatrix} w - w^k \\ b - b^k \end{pmatrix} \right\rangle + \frac{L}{2} \left\| \begin{matrix} w - w^k \\ b - b^k \end{matrix} \right\|_2^2 + g(w) \quad (*)$$

on minimise pour obtenir w^{k+1}, b^{k+1}.

$$\nabla_{w,b} f(w^k,b^k) = \begin{pmatrix} \nabla_w f(w^k,b^k) \\ \nabla_b f(w^k,b^k) \end{pmatrix} \in \mathbb{R}^{d+1}$$

$w \in \mathbb{R}^d$, $b \in \mathbb{R}$

$$\begin{pmatrix} w \\ b \end{pmatrix} = \begin{pmatrix} w_1 \\ \vdots \\ w_d \\ b \end{pmatrix} \in \mathbb{R}^{d+1} \qquad \left\| \begin{pmatrix} w \\ b \end{pmatrix} \right\|_2^2 = \sum_{j=1}^{d} w_j^2 + b^2.$$

$$= \|w\|_2^2 + b^2.$$

$$(*) = \langle \nabla_w f(w^k,b^k), w - w^k \rangle + \nabla_b f(w^k,b^k)(b-b^k) + \frac{L}{2}\|w - w^k\|_2^2 + \frac{L}{2}(b-b^k)^2 + g(w).$$

$$w^{k+1} = \underset{w}{\operatorname{argmin}} \langle \nabla_w f(w^k,b^k), w - w^k \rangle + \frac{L}{2}\|w-w^k\|_2^2 + g(w)$$

$$b^k = \underset{b}{\operatorname{argmin}} \nabla_b f(w^k,b^k)(b-b^k) + \frac{L}{2}(b-b^k)^2$$

Conclusion : l'update w^{k+1} est donné par la descente de gradient proximale

l'update b^{k+1} est donné par la descente de gradient.