DS 5220 Homework - 03
KAVANA VENKATESH

1. Given $y^i \sim N(\exp(\omega x^i), 1)$

WKT, Normal distribution $= \frac{1}{\sqrt{2\pi\sigma^2}} \cdot e^{-\frac{(y^i-\mu)^2}{2\sigma^2}}$ —— ①

Consider $\sum_{i=1}^{N} x^i \exp(\omega x^i)$,

$\therefore$ WKT, $\mu = e^{\omega x^i}$ and $\sigma^2 = 1$.

Assuming random variables are IIDs, we've

$$\mathcal{L}(\mu, \sigma^2 | x) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y^i - e^{\omega x^i})^2}{2}}$$

Taking ln on both sides,

$$\ln(\mathcal{L}(\mu, \sigma^2 | x)) = \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}} \cdot e^{-\frac{(y^i - e^{\omega x^i})^2}{2}}\right)$$

$$= \ln\left(\prod_{i=1}^{N} \frac{1}{\sqrt{2\pi}}\right) + \ln \prod_{i=1}^{N} e^{-\frac{(y^i - e^{\omega x^i})^2}{2}}$$

$$= N \cdot \ln \frac{1}{\sqrt{2\pi}} + \sum_{i=1}^{N} \frac{1}{2} \cdot -(y^i - e^{\omega x^i})^2$$

$$= -\frac{N}{2} \ln 2\pi - \frac{1}{2} \sum_{i=1}^{N} (y^i - e^{\omega x^i})^2$$

Differentiating the above eq. and setting it to 0.

Scanned with CamScanner

$$\frac{\partial}{\partial \omega} \ln\left(\mathcal{L}\left(e^{\omega x^i}, 1/\tau\right)\right) = \frac{\partial}{\partial \omega}\left[-\frac{N}{2}\ln 2\pi - \frac{1}{2}\sum_{i=1}^{N}\left(y^i - e^{\omega x^i}\right)^2\right]$$

$$\Rightarrow \frac{\partial}{\partial \omega}\left(-\frac{N}{2}\ln(2\pi)\right) - \frac{1}{2}\frac{\partial}{\partial \omega}\sum_{i=1}^{N}\left(y^i - e^{\omega x^i}\right)^2$$

$$\Rightarrow 0 - \frac{1}{2}\cdot 2\cdot \sum_{i=1}^{N}\left(y^i - e^{\omega x^i}\right)\left(0 - e^{\omega x^i}\cdot x^i\right)$$

$$\Rightarrow 0 + \sum_{i=1}^{N}\left(y^i - e^{\omega x^i}\right)\left(e^{\omega x^i}\cdot x^i\right) = 0.$$

$$\Rightarrow \sum_{i=1}^{N} y^i x^i \cdot e^{\omega x^i} - \sum_{i=1}^{N} e^{\omega x^i}\cdot e^{\omega x^i}\cdot x^i = 0.$$

$$\Rightarrow \sum_{i=1}^{N} x^i\cdot e^{2\omega x^i} = \sum_{i=1}^{N} y^i x^i e^{\omega x^i}.$$

$$\Rightarrow \boxed{\sum x^i \exp(2\omega x^i) = \sum x^i y^i \exp(\omega x^i)}$$

$\therefore$ E is the solution.

2. <u>MAP</u> <u>Estimate</u>:-

Given total no. of trials, $N = N_0 + N_1$.

Bernoulli random variable x, $p(x=1) = \theta$.

Given dataset, $\mathcal{D} = \{x_1, \ldots x_N\}$.

$N_0 \rightarrow$ No. of trials when $x_i = 0$.
$N_1 \rightarrow$ No. of trials when $x_i = 1$.

$$P(\theta) = \begin{cases} 0.2 & \text{if } \theta = 0.6 \\ 0.8 & \text{if } \theta = 0.8 \\ 0 & \text{otherwise} \end{cases}$$

2.1) Write down the log likelihood function $P(D/\theta)$. What's the max. likelihood solution for θ?

Likelihood function for $P(D/\theta)$.

$$\log(P(D/\theta)) = \log \prod_{i=1}^{N} \theta^{x_i} (1-\theta)^{1-x_i}$$

$$= \sum_{i=1}^{N} \left[\log \theta^{x_i} + \log (1-\theta)^{1-x_i}\right]$$

$$= \sum_{i=1}^{N} \left[x_i \log \theta + (1-x_i) \log(1-\theta)\right]$$

Maximum likelihood solution wrt θ

$$\frac{d}{d\theta} \log P(D/\theta) = \frac{d}{d\theta}\left(\sum_{i=1}^{N} \left[x_i \log \theta + (1-x_i) \log (1-\theta)\right]\right) = 0.$$

$$\Rightarrow \sum_{i=1}^{N} \left[x_i \frac{d}{d\theta} \log \theta + (1-x_i) \frac{d}{d\theta} \log(1-\theta)\right] = 0.$$

$$\Rightarrow \sum_{i=1}^{N} \left[\frac{x_i}{\theta} + \frac{(1-x_i)(-1)}{1-\theta}\right] = 0.$$

Scanned with CamScanner

$$= \sum_{i=1}^{N} \frac{x_i}{\theta} = \sum_{i=1}^{N} \frac{(1-x_i)}{1-\theta}$$

$$\Rightarrow \frac{1}{\theta} \sum_{i=1}^{N} x_i = \frac{1}{1-\theta} \cdot \sum_{i=1}^{N} (1-x_i)$$

$$\Rightarrow (1-\theta) \sum_{i=1}^{N} x_i = \theta \cdot \sum_{i=1}^{N} (1-x_i)$$

$$\Rightarrow \sum_{i=1}^{N} x_i - \theta \cdot \sum_{i=1}^{N} x_i = \theta \cdot \sum_{i=1}^{N} (1-x_i)$$

$$\Rightarrow \sum_{i=1}^{N} x_i - \theta \sum_{i=1}^{N} x_i = \theta \cdot \sum_{i=1}^{N} 1 - \theta \cdot \sum_{i=1}^{N} x_i$$

$$\Rightarrow \sum_{i=1}^{N} x_i = \theta \cdot N$$

$$\therefore \boxed{\hat{\theta}_{MLE} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i} = \frac{N_1}{N}$$

2.2) Consider maximizing the posterior distribution, $p(D/\theta) \cdot p(\theta)$ that takes advantage of the prior. What is MAP estimation?

posterior $= p(D/\theta) \cdot p(\theta)$ where
$p(D/\theta)$ = likelihood
$p(\theta)$ = prior

$\therefore \hat{\theta}_{map}$ = argmax log likelihood + prior.

Scanned with CamScanner

$$= \arg\max_{\theta} \left[\sum_{i=1}^{N} x_i \log \theta + (1 - x_i) \log(1-\theta) \right] + \log p(\theta).$$

Given $p(\theta) = \begin{cases} 0.2 & ; \theta = 0.6 \\ 0.8 & \theta = 0.8 \\ 0 & \text{otherwise} \end{cases}$

$\therefore \log p(\theta)$ will always be constant.

$\therefore$ $\boxed{\hat{\theta}_{map} = \hat{\theta}_{mle}}$ because derivation of constant is 0.

we will end up with maximum likelihood equation.

3. <u>Constrained optimization.</u>

3.1) Given regression: $\min_{\theta} \frac{1}{2}\|y - X\theta\|_2^2$ s.t. $w^T\theta = b$. Find the closed form solution.

$$\mathcal{L}(\theta, \alpha) = \frac{1}{2}\|Y - X\theta\|_2^2 + \Sigma\alpha_i (w^T\theta - b).$$

$$= \frac{1}{2}(y - X\theta)^T(Y - X\theta) + \Sigma\alpha_i (w^T\theta - b)$$

Taking derivative wrt θ,

$$\frac{\partial}{\partial\theta}\mathcal{L}(\theta, \alpha) = \frac{\partial}{\partial\theta}\left(\frac{1}{2}\right)\left[Y^TY - Y^TX\theta - X^T\theta^TY + X^T\theta^TX\theta\right] + \left[\Sigma\alpha_i w^T\right]^T$$

$$= \frac{1}{2}\left[-[Y^TX]^T - X^TY + 2X^TX\theta\right] + \alpha^{*T}w \quad \text{where } \Sigma\alpha_i = \alpha^*$$

$$= \frac{1}{2}\left[-2X^TY + 2\theta\right] + \alpha^{*T}w$$

$$(\because X^TX = I).$$

Equating to 0, we have

$$-X^T Y + \theta = -\alpha^{*T} \omega.$$

$$\boxed{\theta^* = X^T Y - \alpha^{*T} \omega} \quad \text{—— ①.}$$

or $$\boxed{\theta^* = X^T Y - \omega^T \alpha.}$$

Scanned with CamScanner

3.2) Now, $\frac{\partial L}{\partial \alpha(\theta^*, \alpha^*)} = \frac{\partial}{\partial \alpha}\left[\frac{1}{2}\left[y^T Y - Y^T X\theta - \theta^T X^T Y - \|\theta\|_2^2\right] + \sum_{i=1}^{N} \alpha_i (\omega^T \theta^* - b)\right]$

$= 0.$

$= \omega^T \theta^* - b = 0.$

$\Rightarrow \omega^T (X^T Y - \sum_{i=1}^{N} \alpha_i^T \omega) - b = 0.$

$\omega^T X^T Y - \omega^* \alpha^* \omega - b = 0$

$\&\ \sum_{i=1}^{N} \alpha_i^T = \alpha^*$

$\Rightarrow \omega^T x^T y - b = \omega^T \alpha^* \omega.$

$$\boxed{\alpha^* = \frac{(\omega^T x^T y - b)}{\|\omega\|_2^2}} \rightarrow ②.$$

Substitute ② in ①,

$$\theta^* = X^T Y - \alpha^* \omega = X^T Y - \left(\frac{\omega^T x^T y}{\|\omega\|_2^2}\right)\omega$$

$$X^T y - \frac{(\omega^T x^T y - b)\omega}{\omega^T \omega} = X^T y - \frac{\omega^T x^T y}{\omega^T} + \frac{b}{\omega^T}$$

$$= X^T y - X^T y + \frac{b}{\omega^T}.$$

$$\Rightarrow \theta^* = \frac{b}{\omega^T}$$

$$\Rightarrow \boxed{\omega^T \theta^* = b}$$ Thus verified.

3.3) without the constraint, we have

$$\mathcal{L}(\theta) = \frac{1}{2}\|Y - X\theta\|_2^2 = \frac{1}{2}\Big[Y^T Y - Y^T X\theta -$$

$$X^T \theta^T Y + X^T \theta^T X \theta\Big].$$

Taking the derivative wrt θ.

$$\frac{\partial \mathcal{L}(\theta)}{\partial \theta} = \frac{1}{2}\left[-2X^T Y + 2\theta\right]$$

Equating to zero, we have

$$X^T Y = \theta \Rightarrow \boxed{\theta = X^T Y}$$

4. Convexity of Logistic regression.

Log likelihood of logistic regression is,

$$l(\omega) = \sum_{i=1}^{N} \left[y^i \phi(x^i)^T \omega - \log(1+e^{\omega^T \phi(x^i)}) \right]$$

→ To prove log likelihood of logistic regression is convex, we prove Hessian is positive semi definite.

$$\nabla^2_{\vec{\omega}} (-l(\omega)) \geq 0 \quad [\because +ve \text{ semi definite}]$$

$$-l(\omega) = -\sum_{i=1}^{N} \left[y^i \phi(x^i)^T . \omega - \log(1+e^{\omega^T \phi(x^i)}) \right]$$

$$\Rightarrow -\left[y \phi^T \omega - \log(1+e^{\phi^T \omega}) \right]$$

→ Gradient, $\nabla_{\vec{\omega}} = -(y\phi^T)^T + \frac{1}{1+e^{\phi^T\omega}} \cdot (0+e^{\phi^T\omega}) . \phi$

Hessian, $\nabla_2 \vec{\omega} = 0 + \frac{\partial}{\partial \omega} \cdot \left(\frac{\phi e^{\phi^T . \omega}}{1+e^{\phi^T . \omega}} \right)$

$$\Rightarrow \frac{\partial}{\partial \omega} \cdot \left(\frac{\phi e^{\phi^T \omega}}{1+e^{\phi^T . \omega}} \right)$$

$$\Rightarrow \phi \cdot \left(\frac{\partial}{\partial \omega} \cdot \left(\frac{e^{\phi^T \omega}}{1+e^{\phi^T \cdot \omega}} \right) \right)$$

WKT, $\frac{\partial}{\partial x}\left(\frac{u}{v}\right) = \left(\frac{v \frac{du}{dx} - u \frac{dv}{dx}}{v^2} \right)$

$u = e^{\phi^T \cdot \omega}$, $v = 1 + e^{\phi^T \cdot \omega}$.

$$\Rightarrow \phi \cdot \left(\frac{\left(1+e^{\phi^T \cdot \omega}\right)\left(\frac{d}{d\omega} \cdot e^{\phi^T \omega}\right) - e^{\phi^T \omega}\left(\frac{d}{d\omega}\left(1+e^{\phi^T \omega}\right)\right)}{\left(1+e^{\phi^T \omega}\right)^2} \right)$$

$$= \phi \left(\frac{e^{\phi^T \omega} \cdot \left(\phi^T\right)^T \left(1+e^{\phi^T \omega}\right) - e^{\phi^T \omega}\left(e^{\phi^T \omega} \cdot \phi^T\right)}{\left(1+e^{\phi^T \omega}\right)^2} \right)$$

$$= \phi \left(\frac{\left(\phi^T\right)^T e^{\phi^T \omega} \left(1+e^{\phi^T \omega}\right) - \phi^T e^{2\phi^T \omega}}{\left(1+e^{\phi^T \omega}\right)^2} \right)$$

$$\boxed{\nabla^2_{\omega} = \frac{\phi^2 e^{\phi^T \omega}}{\left(1+e^{\phi^T \omega}\right)^2}} \geq 0. \quad \left(\because \text{+ve Value} \geq 0 \;\; \text{Squared Value} \right)$$

$\therefore$ This is a convex function.

5) Let all the points using $\phi(x)$ be obtained as,

5.1) $$\phi(x) = [1, x_1, x_2, x_1 x_2].$$

class	$x(x_1, x_2)$	$\phi(x)\,[1, x_1, x_2, x_1 x_2]$
+	(1, 1)	(1, 1, 1, 1)
+	(−1, −1)	(1, −1, −1, 1)
−	(−1, 1)	(1, −1, 1, −1)
−	(1, −1)	(1, 1, −1, −1).

we can see that we can ignore the first coordinate as its the same for all the points. From simple intuition we can see that the points can be linearly separated by boundary $c_4 = 0$ and hence, $\omega = [0, 0, 0, 1]$ and this is the hyperplane that linearly separates +ve & −ve points.

<u>Note:-</u> Both classes are not linearly separable in original space but can be linearly separated if we transform the feature space

5.2) $K(X, Z) = \langle \phi(x), \phi(z) \rangle$

$= [1, X_1, X_2, X_1 X_2]^T [1, Z_1, Z_2, Z_1 Z_2]$

$= 1 + X_1 Z_1 + X_2 Z_2 + X_1 X_2 Z_1 Z_2.$

$= 1 + X^T Z + X_1 X_2 Z_1 Z_2.$

6.1) <u>Constructing Kernels</u>:-

Given $K_1(x, z) : \mathbb{R}^d * \mathbb{R}^d \to R$

$K_2(x, z) : \mathbb{R}^d * \mathbb{R}^d \to R$

To prove: $K(x, z) \Rightarrow c_1 K_1(x, z) + c_2 K_2(x, z)$ for $c_1, c_2 \geq 0$

K_1 has its feature map ϕ_1 and inner product $\langle \rangle_{HK_1}$ and K_2 has its feature map ϕ_2 and inner product $\langle \rangle_{HK_2}$.

So, $K(x, z) = \langle \sqrt{c_1}\, \phi_1(x), \sqrt{c_1}\, \phi_1(z) \rangle_{HK_1} +$

$\langle \sqrt{c_2}\, \phi_2(x), \sqrt{c_2}\, \phi_2(z) \rangle_{HK_2}$.

$= \langle [\sqrt{c_1}\, \phi_1(x), \sqrt{c_2}\, \phi_2(z)],$

$\sqrt{c_1}\, \phi_1(z), \sqrt{c_2}\, \phi_2(z)] \rangle_{HK_{new}}$

6.2) $K(x, z) = K_1(x, z) \cdot K_2(x, z)$

$$= \phi_1(x)^T \phi_1(z)\, \phi_2(x)^T \phi_2(z)$$

$$= \sum_i \phi_1(x)_i \phi_1(z)_i \sum_j \phi_2(x)_j \phi_2(z)_j$$

$$= \sum_{ij} \phi_1(x)_i \phi_1(z)_i \phi_2(x)_j \phi_2(z)_j$$

$$= \sum_{ij} \phi_1(x)_i \phi_2(x)_j \phi_1(z)_i \phi_2(z)_j$$

$$= \underline{\phi(x)^T \cdot \phi(y)}$$ where

$\phi(x) = \phi_1(x) \otimes \phi_2(x)$ (Kronecker product).