CS280 Fall 2018 Assignment 1 Part A

PAGE 1

Name: 王芊

Student ID: 2018233131

① $KL(P_{emp}||q) = \int P_{emp}(x)(\log P_{emp}(x) - \log q(x;\hat\theta))dx$

$= \int P_{emp}(x)\log P_{emp}(x)dx - \int P_{emp}(x)\log q(x;\hat\theta)dx$ --①

$\int P_{emp}(x)\log q(x;\hat\theta)dx = \int \frac{1}{n}\sum_{i=1}^{n}\delta(x,x_i)\log q(x;\hat\theta)dx = \frac{1}{n}\sum_{i=1}^{n}\log q(x;\hat\theta)\int\delta(x,x_i)dx$

$= \frac{1}{n}\sum_{i=1}^{n}\log(q(x;\theta)) \triangleq \hat\theta$

So when $\hat\theta$ is the maximum likelihood estimator, we get the minimum $KL(P_{emp}||q)$, which means $\arg\min_q KL(P_{emp}||q)$ is obtained by $q(x)$.

② "$(J(w))''$" to slove

$J(w) = -\frac{1}{|D|}\sum_{i\in D}\log\sigma(\pm 1\cdot x_i^T w) + \lambda\|w\|_2^2$, label $\log\sigma(y_i x_i^T w)$ as Ⓘ

$\cancel{\frac{\partial(\log\sigma(x_i^T w))}{\partial w_j} = -\frac{\partial(\log(1+e^{-x_i^T w}))}{\partial w_j} = \left(\frac{e^{-x_i^T w}}{1+e^{-x_i^T w}}\right)(x_{ij}) = (1-\sigma(x_i^T w))x_{ij}}$

$\cancel{\frac{-\sigma(x_i^T w)(1-\sigma(x_i^T w))\cdot x_i^T\cdot x_i^T}{1}}$

$\cancel{\frac{\partial^2(\log\sigma(x_i^T w))}{\partial w_j \partial w_k} = \sigma(x_i^T w)(1-\sigma(x_i^T w))x_{ij}x_{ik}}$

$\because \lambda\|w\|^2$ is convex, $\frac{\partial(\log\sigma(y_i x_i^T w))}{\partial w} = \frac{\sigma(y_i x_i^T w)(1-\sigma(y_i x_i^T w))y_i x_i^T}{\sigma(y_i x_i^T w)}$

$= (1-\sigma(y_i x_i^T w))y_i x_i^T$

$\therefore \frac{\partial((1-\sigma(y_i x_i^T w))y_i x_i^T)}{\partial w} = -\sigma(y_i x_i^T w)(1-\sigma(y_i x_i^T w))\cdot(y_i x_i^T)^2$

because $\sigma(x) \in (0,1)$, I'' 中 $1-\sigma(y_i x_i^T w) \in (0,1)$
$(y_i x_i^T)^2 > 0$, so I'' always < 0

when it add the coefficient $-\frac{1}{|D|}$, it will become always > 0

so the first subject is also convex [the hessian matrix ≥ 0]

convex + convex $\Rightarrow$ convex $\Rightarrow$ local optimal = global optimal

$\Rightarrow$ False

False L1 norm prefer a sparse $\hat{w}$, but L2 norm prefer a average value $\hat{w}$
if we consider $J(w)$ as a loss function, when we optimize and get the $\operatorname{argmin}_w J(w) = \hat{w}$, it's not sparse.

③ $$l(\theta) = \sum_{i=1}^{m} \log P(x_i;\theta) = \sum_{i=1}^{m} \log \sum_{z} P(x, z;\theta)$$

$$= \sum_{i} \log \sum_{z^{(i)}} Q_i(z^{(i)}) \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

$$\geq \sum_{i} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)};\theta)}{Q_i(z^{(i)})}$$

$$= l(\theta^{(t)})$$

$$w_j^{(i)} = Q_i(z^{(i)}=j) = P(z^{(i)}=j \mid x^{(i)};\phi,\mu,\Sigma)$$

$$\sum_{i=1}^{m} \sum_{z^{(i)}} Q_i(z^{(i)}) \log \frac{P(x^{(i)}, z^{(i)};\phi,\mu,\Sigma)}{Q_i(z^{(i)})} \quad \text{------ I}$$

$$= \sum_{i=1}^{m} \sum_{j=1}^{m} w_j^{(i)} \log \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_j|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x^{(i)}-\mu_j)^T \Sigma_j^{-1}(x^{(i)}-\mu_j)\right)\cdot \phi_j}{w_j^{(i)}}$$

$$\nabla_{\mu_l} \text{I} = -\nabla_{\mu_l} \sum_{i=1}^{m} \sum_{j=1}^{K} w_j^{(i)} (x^{(i)}-\mu_j)^T \cdot \frac{1}{2} \Sigma_j^{-1} (x^{(i)}-\mu_j)$$

$$= \frac{1}{2} \sum_{i=1}^{m} w_l^{(i)} \nabla_{\mu_l} 2\mu_l^T \Sigma_l^{-1} x^{(i)} - \mu_l^T \Sigma_l^{-1} \mu_l = \sum_{i=1}^{m} w_l^{(i)} (\Sigma_l^{-1} x^{(i)} - \Sigma_l^{-1} \mu_l)$$

令其为0得 $\underbrace{\mu_l = \frac{\sum_1^m w_l^{(i)} x^{(i)}}{\sum_1^m w_l^{(i)}}}$, $\Rightarrow \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j$

又 $\because \sum_{j=1}^{k} \phi_j = 1$

$\therefore$ let $L(\phi) = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} \log \phi_j + \beta(\sum_{j=1}^{k} \phi_j - 1)$

$$\frac{\partial L(\phi)}{\partial \phi_j} = \sum_{i=1}^{m} \frac{w_j^{(i)}}{\phi_j} + \beta \overset{令}{=} 0$$

$$\Rightarrow \phi_j = \frac{\sum_{i=1}^{m} w_j^{(i)}}{-\beta} \quad \Rightarrow -\beta = \sum_{i=1}^{m}\sum_{j=1}^{k} w_j^{(i)} = \sum_{i=1}^{m} 1 = m$$

$$\therefore \phi_j = \frac{1}{m}\sum_{i=1}^{m} w_j^{(i)}$$

$$\nabla_{\mu_k} \sum_{n=1}^{N}\sum_{k=1}^{K} \gamma_{nk} \log \frac{\frac{1}{(2\pi)^{\frac{n}{2}}|\Sigma_k|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x_n-\mu_k)^T \Sigma_k^{-1}(x_n-\mu_k)\right)\pi_k}{\gamma_{nk}}$$

$$= \left(\sum_{n=1}^{N} \gamma_{nk}\right)(x_n-\mu_k)\Sigma_k^{-1}$$

$$= \sum_{1}^{N} \gamma_{nk} \Sigma_k^{-1}(x_n-\mu_k)$$

'' $\sum_{k=1}^{K} \pi_k = 1$, π is the above ϕ , we have already proved that

$$\pi_k = \frac{1}{N}\sum_{n=1}^{N} \gamma_{nk}$$

''' Σ_k can be the μ above , so we get $\Sigma_k = \frac{\sum_{i=1}^{m} \gamma_{nk} x_i}{\sum_{n=1}^{m} \gamma_n}$

which is $\frac{\partial \ln P(X|\mu,\Sigma)}{\partial \mu} = -\frac{N}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left[\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T\right]\Sigma^{-1}$
$\overset{令}{=} 0$

$$\Rightarrow \Sigma_k = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu)(x_n-\mu)^T$$