Problem 1

Thursday, October 26, 2017 6:25 PM

(a) $\nabla_w l(w) = \nabla\left[\sum_{i=1}^{n} -y_i \log h(x_i) - (1-y_i)\log(1-h(x_i))\right]$

$= \sum_{i=1}^{n} -y_i \cdot \nabla \log h(x_i) - (1-y_i)\nabla \log(1-h(x_i))$

$= \sum_{i=1}^{n} -y_i \cdot \frac{\nabla h(x_i)}{h(x_i)} - (1-y_i)\frac{\nabla(1-h(x_i))}{1-h(x_i)}$ ①

$\nabla h(x_i) = \nabla_w \frac{1}{1+\exp(-w^T x_i)} = \frac{-1\cdot(-x_i)\exp(-w^T x_i)}{(1+\exp(-w^T x_i))^2}$

$= \frac{x_i \cdot \exp(-w^T x_i)}{(1+\exp(-w^T x_i))^2} = h^2(x_i)\cdot\exp(-w^T x_i)\cdot x_i$

$① = \sum_{i=1}^{n} -y_i \frac{\nabla h(x_i)}{h(x_i)} + (1-y_i)\frac{\nabla h(x_i)}{1-h(x_i)}$

$= \sum_{i=1}^{n} -y_i \frac{x_i \exp(-w^T x_i)}{1+\exp(-w^T x_i)} + (1-y_i)\frac{x_i}{1+\exp(-w^T x_i)}$

$= \sum_{i=1}^{n} -y_i x_i + \frac{x_i}{1+\exp(-w^T x_i)}$

$= \sum_{i=1}^{n} -x_i\,(y_i - h(x_i))$

$= -X \cdot (\underline{y} - h(X))$

where $h(X)$ is a vector of $h(x_i)$, X is $[x_1 \cdots x_n] \in \mathbb{R}^{m\times n}$

(b) $\nabla^2 l(w) = \nabla_w\left[-X^T\cdot(\underline{y} - h(X))\right]$

$= \nabla_w X^T \cdot h(X)$

$$= X S X^T$$

where S is a diagonal matrix, with diagonal element

$$S_{ii} = h^2(x_i) \exp(-w^T x_i).$$

$X \in \mathbb{R}^{m \times n}$

Since $h(x_i) > 0$ and $\exp(-w^T x_i) > 0$, it is positive semi-definite and thus $L(w)$ is convex & has no local minima other than the global one.

Problem 2

Sunday, October 29, 2017 9:14 PM

(a) $\log L(\alpha; \underline{X}) = -\frac{n}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\sum_{i=1}^{n}(X_i-\alpha)^2$

$$\frac{\partial \log L(\alpha; \underline{X})}{\partial \alpha} = \frac{1}{\delta^2}\sum_{i=1}^{n}(X_i-\alpha)$$

$$= 0$$

$$\Rightarrow \sum_{i=1}^{n}(X_i - \alpha) = 0$$

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\log L(\alpha, \delta^2; \underline{X}) = -\frac{n}{2}\log(2\pi\delta^2) - \frac{1}{2\delta^2}\sum_{i=1}^{n}(X_i-\alpha)^2.$$

$$\frac{\partial \log L(\alpha, \delta^2; \underline{X})}{\partial(\delta^2)} = -\frac{n}{2\delta^2} + \frac{1}{2\delta^4}\sum_{i=1}^{n}(X_i-\alpha)^2$$

According to previous one for α, the MLE of $[\hat{\alpha}, \hat{\delta^2}]^T$ solves the system of equations

$$\frac{1}{\hat{\delta}^2}\sum_{i=1}^{n}(X_i - \hat{\alpha}) = 0 \quad (1)$$

$$-\frac{n}{2\hat{\delta}^2} + \frac{1}{2(\hat{\delta}^2)^2}\sum_{i=1}^{n}(X_i - \hat{\alpha})^2 = 0 \quad (2)$$

(1) $\Rightarrow$ $\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} X_i$

$$\hat{\delta}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \hat{\alpha})^2.$$

(b) $L(\mu, \Sigma; X_1, \ldots, X_n) = \frac{1}{(2\pi)^{np/2}|\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{k=1}^{n}(\underline{X}_k - \underline{\mu})^T \Sigma^{-1}(\underline{X}_k - \underline{\mu})}$

(b) $$l(\underline{\mu}; \underline{X}_1, \cdots \underline{X}_n) = \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2}\sum_{k=1}^{n} (\underline{X}_k - \underline{\mu})^T \Sigma^{-1} (\underline{X}_k - \underline{\mu})}$$

$$= \frac{1}{(2\pi)^{np/2} |\Sigma|^{n/2}} e^{-\frac{1}{2} tr(\Sigma^{-1} S)} e^{-\frac{1}{2} n (\bar{\underline{X}} - \underline{\mu})^T \Sigma (\bar{\underline{X}} - \underline{\mu})}$$

where $\underline{X}_k$ $k \in \{1, 2, \cdots, n\} \in \mathbb{R}^p$, and where

$S = \sum_{k=1}^{n} (\underline{X}_k - \bar{\underline{X}})(\underline{X}_k - \bar{\underline{X}})^T$ where $\bar{\underline{X}}$ is the sample mean of $\underline{X}_k$.

Then $$\log l(\underline{\mu}; \underline{X}_1, \cdots \underline{X}_n) = c - \frac{n}{2} \log |\Sigma| - \frac{1}{2} tr(\Sigma^{-1} S)$$

$$- \frac{1}{2} n \cdot (\bar{\underline{X}} - \underline{\mu})^T \Sigma^{-1} (\bar{\underline{X}} - \underline{\mu})$$

where c does not depend on the parameter $\underline{\mu}$.

$$\frac{\partial \log l(\underline{\mu}; \underline{X}_1 \cdots \underline{X}_n)}{\partial \underline{\mu}} = 0$$

$$\Rightarrow \quad \bar{\underline{X}} - \underline{\mu} = 0$$

$$\therefore \quad \hat{\underline{\mu}} = \bar{\underline{X}} = \frac{1}{n} \sum_{i=1}^{n} \underline{X}_i$$

Problem 3

Sunday, October 29, 2017 9:41 PM

(a) $H(X) - H(X|Y) = -\int p(x) \ln p(x)\, dX + \int p(x,y) \ln p(x|y)\, dx\, dy$

$= -\int p(x) \ln p(x)\, dX + \int p(x,y) \ln p(x|y)\, dy\, dx$

$= \int \left(\int p(x,y) \ln p(x|y)\, dy \right) - p(x) \ln p(x)\, dx. \quad (1)$

$\ln p(x|y) = \ln \frac{p(x,y)}{p(y)} = \ln \frac{p(y|x)\, p(x)}{p(y)}$

$= \ln p(y|x) + \ln p(x) - \ln p(y)$

$\int p(x,y) \cdot \ln p(x|y)\, dy$

$= \int p(x,y) \ln p(y|x)\, dy + \int p(x,y) \ln p(x)\, dy - \int p(x,y) \ln p(y)\, dy. \quad (2)$

Plug in (2) into (1) we get

$\int \left(\int p(x,y) \ln p(y|x) \right) dy + \int p(x,y)\, dy \cdot \ln p(x) - \int p(x,y) \ln p(y)\, dy$
$- p(x) \ln p(x)\, dx$

$= -H(Y|X) + \iint p(x,y) \ln \frac{p(x)}{p(y)} - p(x) \ln p(x)\, dy\, dx \quad (3)$

Since $p(x) = \int p(x,y)\, dy$

Then the second term of (3) becomes

$\int \int p(x,y) \ln \frac{p(x)}{p(y)} - p(x,y) \ln p(x)\, dy\, dx$

$= -\int \int p(x,y) \ln p(y)\, dy\, dx$

$$= -\int p(Y) \ln p(Y)\, dY$$

$$= H(Y)$$

Finally, ① $= H(Y) - H(Y|X)$

Therefore, we have $H(X) - H(X|Y) = H(Y) - H(Y|X) = I(X,Y)$

(b) $I(X,Y) = H(X) - H(X|Y)$

$$= -\int p(x) \ln p(x)\, dX + \int p(X,Y) \ln(X|Y)\, dX dY$$

plug in $X = f(Y)$, we have

$$= -\int p(f|Y)) \ln p(f(Y))\, df(Y) + \int p(f|Y), Y) \ln p(f(Y)|Y)\, df(Y)\, dY$$

For second term, since $\ln p(f(Y)|Y) = 1$ then it becomes

$$= -\int p(f(Y)) \ln p(f(Y))\, df(Y)$$

$$= -\int p(Y) \ln p(Y)\, dY$$

$$= H(Y)$$

Also, since $-\int p(f(Y)) \ln p(f(Y))\, df(Y)$

$$= H(f(Y))$$
$$= H(X)$$

We have $I(X,Y) = H(X) = H(Y)$

(c) $$\min_\theta D_{KL}(\hat{p} \| q) \triangleq \min_\theta -\int \hat{p}(x) \ln \frac{q(x|\theta)}{\hat{p}(x)} dx$$

$$= \min_\theta -\int \hat{p}(x) \ln q(x|\theta) dx + \int \hat{p}(x) \ln \hat{p}(x) dx$$

$$\propto \min_\theta -\int \hat{p}(x) \ln q(x|\theta) dx$$

$$= \min_\theta -\int \frac{1}{N} \sum_{i=1}^{N} I[x = x_i] \ln q(x|\theta) dx$$

$$= \min_\theta -\frac{1}{N} \sum_{i=1}^{N} \int \delta(x - x_i) \ln q(x|\theta) dx$$

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

$$\propto \max_\theta \sum_{i=1}^{N} \ln q(x_i|\theta)$$

$$= \max_\theta q(\underline{x}|\theta)$$ where it is exactly the maximum likelihood estimate θ_{ML} given dat D.

(d) To maximize entropy for a continuous variable, we need to constrain the first and second moments of $p(x)$ as well as preserve the normalization constraint in order for this maximum well-defined.

Then the three constraint are the following:

$$\int_{-\infty}^{\infty} p(x)\,dx = 1$$

$$\int_{-\infty}^{\infty} x\,p(x)\,dx = u$$

$$\int_{-\infty}^{\infty} (x-u)^2 p(x)\,dx = \delta^2$$

The objective function is given by

$$\max\ -\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx = H(x)$$

Then using Lagrange multipliers to perform the constrained maximization

$$-\int_{-\infty}^{\infty} p(x)\ln p(x)\,dx + \lambda_1\left(\int_{-\infty}^{\infty} p(x)\,dx - 1\right) + \lambda_2\left(\int_{-\infty}^{\infty} x\,p(x)\,dx - u\right) + \lambda_3\left(\int_{-\infty}^{\infty} (x-u)^2 p(x)\,dx - \delta^2\right) = F(p(x))$$

Let $\dfrac{\partial F(p(x))}{p(x)} = 0$

we have $p(x) = \exp\{-1 + \lambda_1 + \lambda_2 x + \lambda_3 (x-u)^2\}$.

Since $\begin{cases} \int p(x)\,dx = 1 \\ \int x\,p(x)\,dx = u \\ \int (x-u)^2 p(x)\,dx = \delta^2 \end{cases}$

we solve λ_1, λ_2 & λ_3 and plug into $p(x)$ we get

$$\hat{p}(x) = \frac{1}{(2\pi\delta^2)^{\frac{1}{2}}} \exp\left\{-\frac{(x-u)^2}{2\delta^2}\right\}$$

where $\hat{p}(x)$ is exactly Gaussian distributed.

Therefore Gaussian distribution has maximum entropy among all distribution

Therefore Gaussian distribution has maximum entropy among all distribution of the same variance.

Problem 4

Wednesday, November 1, 2017 4:53 PM

(a) $\underline{w} = \arg\min_{\beta, b} \sum_{i=1}^{n} c_i (y_i - \beta^T x_i - b)^2$

$= \arg\min_{\underline{w}} \| C^{\frac{1}{2}} (\underline{y} - \underline{\underline{X}}\, \underline{w}) \|_2^2$

where $\underline{\underline{C}}$ is diagonal matrix with $C_{ii} = c_i$, and $C = C^{\frac{1}{2}} \cdot C^{\frac{1}{2}}$

$$\frac{\partial L(\underline{w})}{\partial \underline{w}} = -2 (C^{\frac{1}{2}} X)^T C^{\frac{1}{2}} (y - Xw) = 0$$

$$\Rightarrow (CX)^T (y - Xw) = 0$$

$$\hat{\underline{w}} = ((CX)^T X)^{-1} \cdot (CX)^T y$$

$$= (X^T C \cdot X)^{-1} X^T C\, y$$

When $c_i = 1 \ \forall i$, C is identity matrix.

$$\hat{\underline{w}} = (X^T X)^{-1} X^T y \quad , \quad \hat{\underline{w}} = \begin{bmatrix} b \\ \underline{\beta} \end{bmatrix}$$

For $y_i = \beta^T x_i + b + \varepsilon_i$ it can be written as

$$\underline{y} = \underline{\underline{X}}\, \underline{w} + \underline{\varepsilon} \quad \text{where } \underline{\varepsilon} \mid \underline{\underline{X}} \sim N(0, \delta^2 I)$$

$$\hat{\underline{w}}_{ML} = \max_{\underline{w}} p(\underline{y} \mid \underline{w}, \underline{\underline{X}})$$

$\because \underline{y} \mid \underline{w}, \underline{\underline{X}} \sim N(Xw, \delta^2 I)$

$\therefore \max_{\underline{w}} p(\underline{y} \mid \underline{w}, \underline{X}) \quad \max_{\underline{w}} \log p(\underline{y} \mid \underline{w})$

$$= \max_{\underline{w}} -(\underline{y} - \underline{X}\underline{w})^T (\delta^2 I)^{-1} (\underline{y} - \underline{X}\underline{w})$$

$$= \min_{\underline{w}} \frac{1}{\delta^2} \| \underline{y} - \underline{X}\underline{w} \|_2^2$$

$$= \min_{\underline{w}} \| \underline{y} - \underline{X}\underline{w} \|_2^2$$

which is a least square regression problem, and therefore

$$\hat{\underline{w}}_{ML} = (\underline{X}^T \underline{X})^{-1} \underline{X}^T \underline{y} = \hat{\underline{w}}_{LS}.$$

$\therefore$ When $c_i = 1 \ \forall i$, the weighted least squares regression problem is the same as finding the MLE of $\underline{w} = [b, \beta^T]^T$ with y_i being modeled as $y_i = \beta^T x_i + b + \varepsilon_i \ \forall i$.

(b) Now $\underline{y} \mid \underline{X}, \underline{w} \sim N(\underline{X}\underline{w}, \Sigma)$, where Σ is diagonal and $\Sigma_{ii} = \delta_i^2$

$$\max_{\underline{w}} p(\underline{y} \mid \underline{X}, \underline{w})$$

$$= \min_{\underline{w}} (\underline{y} - \underline{X}\underline{w})^T \underline{\underline{\Sigma}}^{-1} (\underline{y} - \underline{X}\underline{w})$$

$$= \min_{\underline{w}} \| \underline{\underline{\Sigma}}^{-\frac{1}{2}} (\underline{y} - \underline{X}\underline{w}) \|_2^2, \text{ which is exactly weighted LS problem.}$$

According to part a, we have

$$\hat{\underline{w}}_{WLS} = (\underline{X}^T \underline{\underline{\Sigma}}^{-1} \underline{X})^{-1} \underline{X}^T \underline{\underline{\Sigma}}^{-1} \underline{y}$$

where we change C_i in part a to Σ^{-1}

where we change $\underline{\underline{C}}$ in part a to $\underline{\underline{\Sigma}}^{-1}$.

∴ The MLE of $\underline{W}$ with different noise variance for each i is equivalent to weighted LS solution with weight matrix $\underline{\underline{C}} = \underline{\underline{\Sigma}}^{-1}$.