SET-3

PIYUSHI

Piyushi

Murphy '12

3.6)

$$\lambda_{MLE} = \underset{\lambda > 0}{\operatorname{argmax}} \sum_{i=1}^{m} \log\left(e^{-\lambda} \frac{\lambda^{x_i}}{x_i!}\right)$$

$$= \underset{\lambda > 0}{\operatorname{argmax}} \sum_{i=1}^{m} \left(-\lambda + x_i \log \lambda\right)$$

$$= \underset{\lambda > 0}{\operatorname{argmin}} \sum_{i=1}^{m} \left(\lambda - x_i \log \lambda\right)$$

Feasibility set is open &

$$\frac{d^2}{d\lambda^2}\left(\lambda - x_i \log \lambda\right)$$

$$= \frac{x_i}{\lambda^2}$$

as $x_i \geq 0$

$$\frac{d^2}{d\lambda^2} \geq 0$$

$\therefore f(\lambda) = \lambda - x_i \log \lambda$

with domain $\{\lambda \mid \lambda > 0\}$, a convex set, is a convex function

$\therefore$ Putting gradient to 0 gives

$$\lambda_{MLE} = \sum_{i=1}^{m} \frac{x_i}{m}$$

3.8) (a) $\hat{a} = \underset{a \in \mathbb{R}}{\operatorname{argmax}} \frac{1}{(2a)^m} \prod_{i=1}^{m} I(x \in [-a, a])$

As the objective decreases with a, we take a as the least value such that $x_i \in [-a, a]$ $\forall$ $i \in [1, m]$

$$\hat{a} = \max_{i \in [1,m]} |x_i|$$

(b) $p(x_{n+1}) = \begin{cases} 0 & \text{if } x_{n+1} \notin [-\hat{a}, \hat{a}] \\ \frac{1}{2\hat{a}} & \text{otherwise} \end{cases}$

(c) The problem is that all datapoints outside of $[-\hat{a}, \hat{a}]$ gets 0 prob.

3.11 (a) $\theta^* = \underset{\theta > 0}{\operatorname{argmax}} \sum_{i=1}^{N} \log(\theta e^{-\theta x_i})$

$$= \underset{\theta > 0}{\operatorname{argmax}} \; N \log \theta - \theta \sum_{i=1}^{N} x_i$$

After checks like convexity as done in 3.6, we can put gradient to 0.

$$\frac{N}{\theta^*} - \sum_{i=1}^{N} x_i = 0$$

$$\Rightarrow \theta^* = \frac{1}{\bar{x}}$$

3.11 (b) $\theta^* = \frac{3}{5+6+4} = \frac{3}{15} = \frac{1}{5}$

$= 0.2$

3.20)

(a) $p(x|y=c)$ when there is no factorization assumption

$p(x|y=c) =$

$p(x_1|y=c)\,p(x_2|y=c,x_1)\,p(x_3|y=c,x_1,x_2)$
$\ldots \quad p(x_D|y=c,x_1,\ldots,x_{D-1})$

We can model each of these factors as Bernoulli.

Total parameters

$$C + 2C + 2^2 C + \dots + 2^{D-1} C$$

$$= C(2^{D-1} - 1)$$

$$O(2^D C)$$

(b) Naive Bayes should give lesser test set error as for full Bayes we have more parameters to estimate & N is small here.

(c) Full Bayes should give lesser test set error as we are doing a richer modeling by not assuming conditional independence

(d) Computational Complexity for fitting as a function of N&D:

Naive Bayes $O(ND)$

Full Bayes, we can maintain count of $x_1, (x_1, x_2), \dots (x_1 \dots x_D)$. Then for each sample we need $1 + 2 + \dots + D$ updates. $\therefore$ total computational complexity $= O(ND^2)$

(e) After training, for applying to a single test case, complexity of Naive Bayes $= O(CD)$. If we can access the parameter from D-bit sample in $O(D)$ then $O(CD)$ for Full Bayes.

(f) Naive Bayes $O(VC)$.
Full Bayes $O(2^h C)$ also need to marginalize out x_h.

3.22)

$$\theta_{spam} = \frac{3}{7}$$

$$\theta_{secret|spam} = \frac{3}{9} = \frac{1}{3} \qquad \theta_{secret|nonspam} = \frac{1}{15}$$

$$\theta_{sports|non\ spam} = \frac{2}{15} \qquad \theta_{dollar|spam} = \frac{1}{9}$$

4.18) a) $x = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

Bayes Rule $\quad p(y_c|x) = \dfrac{p(x|y_c)\, p(y_c)}{\sum_c p(x, y_c)}$ —— ①

Using Naive Bayes assumption,

$$p(x|y_c) = p(x_1=0|y_c)\, p(x_2=0|y_c)$$

Using the parameter values given:

$$p(x_1=0|y_c) = (1-\theta_c)$$
$$= 0.5 \quad \forall \; c \in [0,2]$$

$$p(x_2=0|y_c) = \frac{\exp\left(-\frac{(0-\mu_c)^2}{2\sigma_c^2}\right)}{\sqrt{2\pi}\,\sigma_c}$$

$$p(x_2=0|y_0) = \exp\left(-\frac{1}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}$$

$$p(x_2=0|y_1) = \exp(0)\cdot\frac{1}{\sqrt{2\pi}}$$

$$p(x_2=0|y_2) = \exp\left(-\frac{1}{2}\right)\cdot\frac{1}{\sqrt{2\pi}}$$

$$p(y_0) = 0.5$$
$$p(y_1) = 0.25$$
$$p(y_2) = 0.25$$

$$p(y_0|x) = \frac{e^{-1/2}}{1+2\cdot e^{-1/2}}$$

$$p(y_1|x) = \frac{1}{1+2\cdot e^{-1/2}}$$

$$p(y_2|x) = \frac{e^{-1/2}}{1+2\cdot e^{-1/2}}$$

b) $p(y_c | x_1 = 0)$

$$= \frac{p(x_1 = 0 | y_c)\, p(y_c)}{p(x_1 = 0)}$$

$$= \frac{(1-\theta_c) \times \pi_c}{(1-\theta_0) \times 0.5 + (1-\theta_1) \times 0.25 + (1-\theta_2) \times 0.25}$$

$$= \pi_c$$

$$p(y \mid x_1 = 0) = \pi$$

c) $p(y_c | x_2 = 0) =$

$$\frac{\frac{\exp}{\sqrt{2\pi}\,\sigma_c}\left(-\frac{(0-\mu_c)^2}{2\sigma_c^2}\right)\pi_c}{\sum_c \frac{\exp}{\sqrt{2\pi}\,\sigma}\left(-\frac{(0-\mu_c)^2}{2\sigma_c^2}\right)\pi_c}$$

$$= \frac{\exp\left(-\frac{\mu_c^2}{2}\right)\pi_c}{\exp\left(-\frac{1}{2}\right)0.5 + 0.25\left(1+\exp\left(-\frac{1}{2}\right)\right)}$$

$$p(y_0 \mid x_2 = 0) = \frac{0.5\, e^{-1/2}}{0.75\, e^{-1/2} + 0.25}$$

$$= \frac{2\, e^{-1/2}}{3\, e^{-1/2} + 1}$$

$$p(y_1 \mid x_2 = 0) = \frac{1}{3\, e^{-1/2} + 1}$$

$$p(y_2 \mid x_2 = 0) = \frac{e^{-1/2}}{3\, e^{-1/2} + 1}$$

d) As θ is uniform

$p(y \mid x_1 = 0)$ is same as the prior distribution over y, i.e. π

but $p(y \mid x_2 = 0)$ is not same as π, where $x_2 \sim \mathcal{N}(\mu_c, \sigma_c^2)$

4.19) Let $y \sim \text{Ber}(\theta)$

$$p(y \mid x) \propto p(x \mid y)\, p(y)$$

$$p(x \mid y=0) = \frac{\exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)}{\sqrt{2\pi}\,|\Sigma_0|}$$

$$p(x \mid y=1) = \frac{\exp\left(-\frac{1}{2K}(x-\mu_1)^T \Sigma_0^{-1} (x-\mu_1)\right)}{\sqrt{2\pi}\,K|\Sigma_0|}$$

$$p(y=0 \mid x) \propto \frac{(1-\theta)\exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)}{\sqrt{2\pi}\,|\Sigma_0|}$$

$$p(y=1|x) \propto \frac{\theta \exp\left(-\frac{1}{2k}(x-\mu_1)^T \Sigma_0^{-1} (x-\mu_1)\right)}{\sqrt{2\pi}\, k\, |\Sigma_0|}$$

$$p(y=1|x) =$$

$$\frac{\theta \exp\left(-\frac{1}{2k}(x-\mu_1)^T \Sigma_0^{-1} (x-\mu_1)\right)}{\left(\theta \exp\left(-\frac{1}{2k}(x-\mu_1)^T \Sigma_0^{-1} (x-\mu_1)\right) + k(1-\theta) \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1} (x-\mu_0)\right)\right)}$$

$$= \frac{1}{1 + \frac{(1-\theta)}{\theta} k \exp\left(-\frac{1}{2k} g(x)\right)}$$

where $g(x) =$

$$-x^T \Sigma_0^{-1} x - \mu_1^T \Sigma_0^{-1} \mu_1 + 2x^T \Sigma_0^{-1} \mu_1$$

$$+k\left(x^T \Sigma_0^{-1} x + \mu_0^T \Sigma_0^{-1} \mu_0 - 2x^T \Sigma_0^{-1} \mu_0\right)$$

$$= (K-1)\, x^T \Sigma_0^{-1} x - 2x^T \Sigma_0^{-1} (K\mu_0 - \mu_1) + K\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_0^{-1} \mu_1$$

To plot decision boundary, we notice that region for class 1 needs

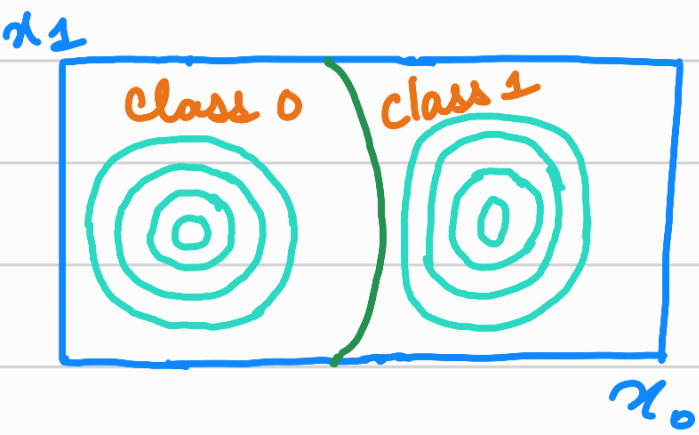
$$\left(\frac{1-\theta}{\theta}\right) K \exp\left(-\frac{1}{2K} g(x)\right) \leq 1$$

$\Rightarrow$

$$-\frac{1}{2K} g(x) \leq \ln \frac{\theta}{(1-\theta)K}$$

$$(K-1)\, x^T \Sigma_0^{-1} x - 2x^T \Sigma_0^{-1} (K\mu_0 - \mu_1) + K\mu_0^T \Sigma_0^{-1} \mu_0 - \mu_1^T \Sigma_0^{-1} \mu_1 \geq c$$

where $c = 2K \ln \frac{(1-\theta)K}{\theta}$



Quadratic decision boundary

4.21)

a) $p(x \mid \mu_1, \sigma_1) = \dfrac{\exp}{\sqrt{2\pi}}\left(-\dfrac{x^2}{2}\right)$

$p(x \mid \mu_2, \sigma_2) = \dfrac{\exp}{10^3\sqrt{2\pi}}\left(-\dfrac{(x-1)^2}{2\times 10^6}\right)$

$R_1 = \left\{x : \exp\left(-\dfrac{x^2}{2}\right) \geq \dfrac{1}{10^3}\exp\left(-\dfrac{(x-1)^2}{2\times 10^6}\right)\right\}$

Solving

$$-\frac{x^2}{2} \geq -3\ln 10 - \frac{(x-1)^2}{2\times 10^6}$$

$$\Rightarrow -\frac{x^2}{2} \geq -3\ln 10 - \frac{x^2+1-2x}{2\times 10^6}$$

$$\Rightarrow 10^6 x^2 \leq 6\times 10^6 \ln 10 + x^2 + 1 - 2x$$

$$\Rightarrow (1+10^6)x^2 + 2x - 6\times 10^6 \ln 10 - 1 \leq 0$$

Roots of quadratic =

$$\frac{-2 \pm \sqrt{4 + 4(1+10^6)(6\times10^6 \ln 10 + 1)}}{2(1+10^6)}$$

Denoting these by α, β

$$R_1 = \{x \mid x \in [\alpha, \beta]\}$$

b) when $\sigma_2 = 1$

$$R_1 = \left\{x : \exp\left(-\frac{x^2}{2}\right) \geq \exp\left(-\frac{(x-1)^2}{2}\right)\right\}$$

Solving

$$\cancel{x^2} \leq \cancel{x^2} + 1 - 2x$$

$$R_1 = \{x : x \leq 1/2\}$$

4.22) $p(Y=c|x) \propto p(x|y=c)$ here.

(a) $x = [-0.5, 0.5]$

$$p(x|y=1) = \frac{1}{\sqrt{2\pi}(0.49)} \exp\left(-\frac{(0.5)^2}{0.7}\right)$$

$$\Sigma_2^{-1} = \frac{1}{|\Sigma_2|} \begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}$$

$$p(x|y=2) = \frac{1}{\sqrt{2\pi}(0.6)} \times$$

$$\exp\left(-\frac{1}{2\times 0.6}[-1.5, -0.5]\begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}\begin{bmatrix} -1.5 \\ -0.5 \end{bmatrix}\right)$$

$$p(x|y=3) = \frac{1}{\sqrt{2\pi}(0.6)} \times$$

$$\exp\left(-\frac{1}{2\times 0.6}[0.5 \quad -0.5]\begin{bmatrix} 0.8 & -0.2 \\ -0.2 & 0.8 \end{bmatrix}\begin{bmatrix} 0.5 \\ -0.5 \end{bmatrix}\right)$$

Predicted class label for $x = 1$

(b) Proceeding as above. Predicted class label for the new x is also 1.

4.23)

a) $\mu_m = \dfrac{67 + 79 + 71}{3}$

$= 69$

$\mu_f = \dfrac{68 + 67 + 60}{3}$

$= 65$

$\sigma_m = 2.67$

$\sigma_f = 12.67$

$$\pi_m = 3/6 = 0.5$$

$$\pi_f = 0.5$$

b) $p(y=m \mid x, \hat{\theta}) =$

$$\frac{p_{\hat{\theta}}(x \mid y=m)\, p_{\hat{\theta}}(y=m)}{\sum_{c \in \{m,f\}} p_{\hat{\theta}}(x \mid y=c)\, p_{\hat{\theta}}(y=c)}$$

$$= \frac{1}{1 + \dfrac{p_{\hat{\theta}}(x \mid y=f)}{p_{\hat{\theta}}(x \mid y=m)}}$$

$$= \frac{1}{1 + \dfrac{\exp\left(-\dfrac{(72-65)^2}{2(12.67)^2}\right)}{\exp\left(-\dfrac{(72-69)^2}{2(2.67)^2}\right)}}$$

$= 0.38$

(c) we can extend it to multiple attributes by including Naive Bayes assumption.

$$p(x \mid y=c) = \prod_{i=1}^{nf} p(x_i \mid y=c)$$

where nf is the no. of features.

$x_i \sim N(\mu_c, \sigma_c)$

7.5) $Y \mid X = x \sim N(w_0 + w^T x, I)$

$p(y \mid x) =$

$$\exp\left(-\frac{(y-\omega_0-\omega^T x)^2}{2}\right)\frac{1}{\sqrt{2\pi}}$$

min. MCLE

$$\min_{\omega_0,\omega} \sum_{i=1}^{N}(y_i-\omega_0-\omega^T x_i)^2$$

On putting gradients wrt ω_0 & ω to 0

$$2\sum_{i=1}^{N}(y_i-\hat{\omega}_0-\hat{\omega}^T x_i)=0 \quad \text{— (1)}$$

$$-2\sum_{i=1}^{N}(y_i-\hat{\omega}_0-\hat{\omega}^T x_i)x_i=0 \quad \text{— (2)}$$

From (1) $\hat{\omega}_0 = \bar{y} - \hat{\omega}^T \bar{x}$

Using in (2),

$$\hat{\omega}_0 \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i x_i^T \hat{\omega} = \sum_{i=1}^{N} x_i y_i$$

$$(\bar{y} - \hat{\omega}^T \bar{x}) \sum_{i=1}^{N} x_i + \sum_{i=1}^{N} x_i x_i^T \hat{\omega} = \sum_{i=1}^{N} x_i y_i$$

$$\left(\sum_{i=1}^{N} x_i x_i^T - \sum_{i=1}^{N} x_i \bar{x}^T \right) \hat{\omega} = \sum_{i=1}^{N} x_i y_i - \bar{y} \sum_{i=1}^{N} x_i$$

Note that RHS is same as

$$\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x}) =$$

$$\sum_{i=1}^{N} y_i x_i - \cancel{N \bar{y} \bar{x}} - N \bar{y} \bar{x} + \cancel{N \bar{y} \bar{x}}$$

$$= \sum_{i=1}^{N} y_i x_i - \bar{y} \sum_{i=1}^{N} x_i$$

Also $\left(\sum_{i=1}^{N} x_i x_i^T - \sum_{i=1}^{N} x_i \bar{x}^T \right)$ is same as $\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$

$$\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T$$

$$= \sum_{i=1}^{N} x_i x_i^T - 2N\bar{x}\bar{x}^T + N\bar{x}\bar{x}^T$$

$$= \sum_{i=1}^{N} x_i x_i^T - N\bar{x}\bar{x}^T$$

$$= \sum_{i=1}^{N} x_i x_i^T - \sum_{i=1}^{N} x_i \bar{x}^T$$

Hence,

$$\hat{\omega} = \left[\sum_{i=1}^{N} (x_i - \bar{x})(x_i - \bar{x})^T\right]^{-1} \left[\sum_{i=1}^{N} (y_i - \bar{y})(x_i - \bar{x})\right]$$

7.6) This is special case of 7.5 when input is a scalar.

$$\hat{\omega}_0 = \bar{y} - \hat{\omega}_1 \bar{x}$$

$$\hat{\omega}_1 = \frac{\sum_{i=1}^{N} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N} (x_i - \bar{x})^2}$$

7.9)

(a) $p(y|x) \propto p(x, y)$

$\propto$

$$\exp\left(-\frac{1}{2}\left((x-\mu_x)^T \Lambda_{xx} (x-\mu_x) + (y-\mu_y)^T \Lambda_{yy} (y-\mu_y) + 2(x-\mu_x)^T \Lambda_{xy} (y-\mu_y)\right)\right)$$

Ignoring terms not having y & Completing squares in y

$$p(y|x) \propto \exp\left(-\frac{1}{2}\left(\left(y-\mu_y+\Lambda_{yy}^{-1}\Lambda_{yx}(x-\mu_x)\right)^T \Lambda_{yy} \left(y-\mu_y+\Lambda_{yy}^{-1}\Lambda_{yx}(x-\mu_x)\right)\right)\right)$$

$$E[Y|x] = \mu_y - \Lambda_{yy}^{-1} \Lambda_{yx} (x - \mu_x)$$

Using Schur Complement

$$= \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x)$$

From 7.5, $\hat{\omega} = \Sigma_{xx}^{-1} \Sigma_{yx}$

$$\omega_0 = \mu_y - \mu_x^T \Sigma_{xx}^{-1} \Sigma_{yx}$$

$$\omega_0 + \hat{\omega}^T x$$

$$= \mu_y - \mu_x^T \Sigma_{xx}^{-1} \Sigma_{yx} + \Sigma_{yx} \Sigma_{xx}^{-1} x$$

$$= \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x)$$

Hence, shown that generative

linear regression also derives the same equations.

(b) Advantage that we also know $p(x|y)$ & can generate more data using it.

Disadvantage: Additional computation-al cost to compute marginal. Also, there we only estimated a vector w & a scalar w_0, but here we need to estimate matrices Σ_{xx}, Σ_{yx} & vectors μ_x, μ_y which may need more no. of training samples.

A.20) $$L(M) = \frac{1}{n} \sum_{i=1}^{n} \log p(y_i | x_i, \hat{\theta}, M)$$

(a) $L(\text{Gauss I}) \leq L(\text{LinLog})$

As Logistic regression is a

discriminative model trained by optimizing $L(M)$ but Gauss1 is a generative model trained to max joint likelihood.

(b) $L(\text{Gauss X}) \leq L(\text{QuadLog})$

Same reason as (a)

(c) $L(\text{LinLog}) \leq L(\text{QuadLog})$

QuadLog is a richer model & can subsume LinLog when weights for quad. terms are 0.

(d) $L(\text{Gauss1}) \leq L(\text{QuadLog})$

Same reason as (a)

(e) $R(M) = \frac{1}{n} \sum_{i=1}^{n} I(y_i \neq \hat{y}(x_i))$

$L(M) > L(M')$ implies $R(M) < R(M')$ with large no. of samples as $L(M)$ maximization is MCLE

$$\min_{\omega} E[KL(p^*(Y|X) \| p_\theta(Y|X))]$$

so for each x, $p_\theta(Y|x)$ will be closer to $p^*(Y|x)$

3.2.1) 3.2.1 & 3.2.2

$$f(x) = E[Y|x]$$
$$= \mu_y + \Sigma_{yx} \Sigma_{xx}^{-1} (x - \mu_x)$$

Let m_i be i^{th} row of μ_y &
a_i be i^{th} row of Σ_{yx}

$$f(x)_i = m_i + a_i^T v(x)$$
where $v(x) = \Sigma_{xx}^{-1}(x - \mu_x)$

Note that there is no i,j terms coupled $\therefore$ we can do d-generative linear scalar regression problems.

3.2.2) Shared on Teams with the name 'set3_gaussian.pdf'

$$f(x) = E[Y|x]$$
$$= -\left(W \Sigma_1^{-1} W^T + \Sigma_2^{-1}\right)^{-1} \left(W \Sigma_1^{-1}(b - x) - \Sigma_2^{-1} \mu_2\right)$$

This doesn't seem to factor for example due to Σ_2^{-1} terms.

Shalen Schwarz

24.2) (1)

$$\min_{\theta} -\frac{1}{m}\sum_{i=1}^{m} \log P_\theta(x_i) - \frac{1}{m}\left(\log\theta + \log(1-\theta)\right)$$

If $P_\theta(x_i) = \theta^{x_i}(1-\theta)^{x_i}$

then we have

$$\min_{\theta} -\frac{1}{m}\log\left(\theta^{\sum_i x_i}(1-\theta)^{m-\sum_i x_i}\right)$$

$$-\frac{1}{m}\log\theta - \frac{1}{m}\log(1-\theta)$$

this is same as having a pseudoexample ($x_{m+1} = 1$) & another one ($x_{m+2} = 0$)

Then $\hat{\theta} = \dfrac{\left(\sum_{i=1}^{m} x_i + x_{m+1} + x_{m+2}\right)}{m+2}$

$$= \left(\sum_{i=1}^{m} x_i + 1 \right) \frac{1}{m+2}$$

Murphy 2022

4.5 (a)

	$y=0$	$y=1$
$x=0$	$\theta_2 (1-\theta_1)$	$(1-\theta_2)(1-\theta_1)$
$x=1$	$(1-\theta_2)\theta_1$	$\theta_2 \theta_1$

(b)

$p(\mathcal{D} \mid \theta, M_2) =$

$\theta_2\theta_1 \quad (1-\theta_2)\theta_1 \quad \theta_2(1-\theta_1) \quad (1-\theta_2)\theta_1$

$\theta_2\theta_1 \quad \theta_2(1-\theta_1) \quad (1-\theta_2)(1-\theta_1)$

MLE:

$$\max_{\substack{\theta_1 \in (0,1) \\ \theta_2 \in (0,1)}} \theta_2^4 \, \theta_1^4 \, (1-\theta_2)^3 \, (1-\theta_1)^3$$

Putting gradient 0

$$4\theta_2^3 (1-\theta_2)^3 - 3(1-\theta_2)^2 \theta_2^4 = 0$$

$$\Rightarrow 4(1-\theta_2) - 3\theta_2 = 0$$

$$\Rightarrow \theta_2 = \frac{4}{7}$$

Similarly $\theta_1 = \frac{4}{7}$

$$p(D \mid \hat{\theta}, M_2) = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6$$

(c) $P(D \mid \theta, M_4) =$

$$\theta_{11}^2 \, \theta_{10}^2 \, \theta_{00}^2 \, \theta_{01}$$

MLE:

$$\max_{\theta_{11},\theta_{10},\theta_{00}\in(0,1)} \theta_{11}^2\,\theta_{10}^2\,\theta_{00}^2\,(1-\theta_{11}-\theta_{10}-\theta_{00})$$

Putting gradients 0.

$$2\theta_{11}(1-\theta_{11}-\theta_{10}-\theta_{00}) - \theta_{11}^2 = 0$$

$$2\theta_{10}(1-\theta_{11}-\theta_{10}-\theta_{00}) - \theta_{10}^2 = 0$$

$$2\theta_{00}(1-\theta_{11}-\theta_{10}-\theta_{00}) - \theta_{00}^2 = 0$$

$$1-\theta_{11}-\theta_{10}-\theta_{00} = \frac{\theta_{11}}{2}$$

$$1-\theta_{11}-\theta_{10}-\theta_{00} = \frac{\theta_{10}}{2}$$

$$1-\theta_{11}-\theta_{10}-\theta_{00} = \frac{\theta_{00}}{2}$$

$$\Rightarrow \frac{1}{2} - \frac{\theta_{11}}{2} - \frac{\theta_{10}}{2} - \frac{\theta_{00}}{2} = \frac{1}{2} - 3\left(1-\theta_{11}-\theta_{10}-\theta_{00}\right)$$

$$\Rightarrow (1 - \theta_{11} - \theta_{10} - \theta_{00}) = 1 - 6(1 - \theta_{11} - \theta_{10} - \theta_{00})$$

$$\Rightarrow 1 - \theta_{11} - \theta_{10} - \theta_{00} = \frac{1}{7}$$

$$\therefore \theta_{11} = \theta_{10} = \theta_{00} = \frac{2}{7}$$

$$\theta_{01} = 1 - 3 \times \frac{2}{7} = \frac{1}{7}$$

With MLE estimates

$$p(D|\hat{\theta}, M_4) =$$

$$\left(\frac{2}{7}\right)^6 \left(\frac{1}{7}\right)$$

10.1) (a)

$$\mu_{ik} = \frac{e^{\omega_k^T x_i}}{\sum_r e^{\omega_r^T x_i}}$$

$$\frac{\partial \mu_{ik}}{\partial \eta_{ij}} = \frac{\partial \mu_{ik}}{\partial (\omega_j^T x_i)} =$$

$$\frac{e^{\omega_k^T x_i} \delta_{kj}}{\sum_r e^{\omega_r^T x_i}} - \frac{e^{\omega_k^T x_i} e^{\omega_j^T x_i}}{\left(\sum_r e^{\omega_r^T x_i}\right)^2}$$

$$= \mu_{ik}(\delta_{kj} - \mu_{ij})$$

(b) NLL:

$$-\frac{1}{N} \sum_{i=1}^{N} \log\left(\frac{e^{\omega_{y_i}^T x_i}}{\sum_r e^{\omega_r^T x_i}}\right)$$

$$= -\sum_{i=1}^{N} \log \mu_{iy_i} \frac{1}{N}$$

Gradient =

$$-\frac{1}{N}\sum_{i=1}^{N} \frac{\partial}{\partial \omega_c} \log \mu_{iy_i}$$

$$= -\sum_{i=1}^{N} \frac{1}{\mu_{iy_i}} \frac{\partial \mu_{iy_i}}{\partial \eta_{ic}} \frac{\partial \eta_{ic}}{\partial \omega_c} \frac{1}{N}$$

$$= -\sum_{i=1}^{N} \frac{1}{\mu_{iy_i}} \mu_{iy_i} (\delta_{y_i c} - \mu_{ic}) x_i \frac{1}{N}$$

$$= -\sum_{i=1}^{N} (\delta_{y_i c} - \mu_{ic}) x_i \frac{1}{N}$$

Another way:

$$\text{NLL}: \quad -\frac{1}{N}\sum_{i=1}^{N} \log\left(\frac{e^{\omega_{y_i}^T x_i}}{\sum_r e^{\omega_r^T x_i}}\right)$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\mu_{iy_i}}\left[\frac{e^{\omega_{y_i}^T x_i} x_i \delta_{y_i c}}{\sum_r e^{\omega_r^T x_i}} - \frac{e^{\omega_{y_i}^T x_i} e^{\omega_c^T x_i} x_i}{\left(\sum_r e^{\omega_r^T x_i}\right)^2}\right]$$

$$= -\frac{1}{N}\sum_{i=1}^{N} \frac{1}{\mu_{iy_i}}\left(\mu_{iy_i}\delta_{y_i c} - \mu_{iy_i}\mu_{iy_c}\right)x_i$$

$$= -\frac{1}{N}\sum_{i=1}^{N}\left(\delta_{y_i c} - \mu_{iy_c}\right)x_i$$

c) $\nabla_{\omega_c} \ell =$

$$-\frac{1}{N}\sum_{i=1}^{N} x_i\left(\delta y_{ic} - \frac{e^{\omega_c^T x_i}}{\sum_r e^{\omega_r^T x_i}}\right)$$

differentiating wrt $\omega_{c'}$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i\left(\frac{e^{\omega_c^T x_i}}{\sum_r e^{\omega_r^T x_i}} x_i \delta_{cc'} - \frac{e^{\omega_c^T x_i} e^{\omega_{c'}^T x_i}}{\left(\sum_r e^{\omega_r^T x_i}\right)^2} x_i\right)^T$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i x_i^T \mu_{ic}(\delta_{cc'} - \mu_{ic'})$$

$$= \frac{1}{N}\sum_{i=1}^{N} \mu_{ic}(\delta_{cc'} - \mu_{ic'}) x_i x_i^T$$

$$H_{cc'} = \frac{1}{N} X^T S X$$

where $S = \mathrm{diag}\big(\mu_{1c}(\delta_{cc'} - \mu_{1c'}),$

$\mu_{2c}(\delta_{cc'} - \mu_{2c'}), \ldots\big)$

We found $H_{cc'} = \frac{1}{N} X^T S X$

which is the sub-matrix correspon-ding to parameters of classes c & c'.

Here $H_{cc'}$ is not a scalar element from the matrix H.

Let $\bar{C}$ denote set of non-overlap-ping indices of the form (c, c').

TST: $\sum_{c,c' \in \bar{C}} (v^{cc'})^T H_{cc'} v^{cc'} \geq 0$

for all vectors $v^{cc'}$.

Denoting $H_{cc'} = \frac{1}{N} (X^{cc'})^T S^{cc'} X^{cc'}$

$$(H_{cc'})_{ij} = \frac{1}{N} X^{cc'T}_{:i} S^{cc'} X^{cc'}_{:j}$$

where $X_{:i}$ denotes i^{th} column of X.

with these, $\sum_{c,c' \in \bar{C}} (v^{cc'})^T H_{cc'} v^{cc'} =$

$$\sum_{c,c' \in \bar{C}} \sum_{i,j} (v^{cc'})_i X_{:i}^{cc'T} S^{cc'} X_{:j}^{cc'} (v^{cc'})_j$$

$$= \sum_{c,c' \in \bar{C}} \sum_{i,j} (v^{cc'})_i X_{:i}^{cc'T} (S^{cc'})^{1/2} (S^{cc'})^{1/2} X_{:j}^{cc'} (v^{cc'})_j$$

$$= \sum_{c,c' \in \bar{C}} \left(\sum_i (S^{cc'})^{1/2} X_{:i}^{cc'} v_i^{cc'} \right)^T \cdot \left(\sum_j (S^{cc'})^{1/2} X_{:j}^{cc'} v_j^{cc'} \right)$$

$$= \sum_{c,c' \in \bar{C}} \left\| \sum_i (S^{cc'})^{1/2} X_{:i}^{cc'} v_i^{cc'} \right\|^2$$

$$\geq 0$$

$\therefore$ Hessian is a PSD

3.2.2 $p^*(y|x) \propto p^*(x|y)\, p^*(y)$

contd. $\propto \exp\left((x - (W^T y + b))^T \Sigma_1^{-1} (x - (W^T y + b)) \right) \exp\left((y - \mu_2)^T \Sigma_2^{-1} (y - \mu_2) \right)$

Denoting Σ_1^{-1} by Λ_1 & Σ_2^{-1} by Λ_2

$$p^*(y|x) \propto \exp\left((W^T y + b - x)^T \Lambda_1 (W^T y + b - x) + (y - \mu_2)^T \Lambda_2 (y - \mu_2) \right)$$

$$= \exp\left(y^T W \Lambda_1 W^T y + (b-x)^T \Lambda_1 (b-x) + 2 y^T W \Lambda_1 (b - x) + y^T \Lambda_2 y + \mu_2^T \Lambda_2 \mu_2 - 2 y^T \Lambda_2 \mu_2 \right)$$

ignoring terms not involving y

$$\propto \exp\left(y^T M y + 2 y^T \left(W \Lambda_1 (b - x) - \Lambda_2 \mu_2 \right) + y^T \Lambda_2 y \right) \quad \text{where } M = W \Lambda_1 W^T$$

$$= \exp\left(y^T (W \Lambda_1 W^T + \Lambda_2) y + 2 y^T \left(W \Lambda_1 (b - x) - \Lambda_2 \mu_2 \right) \right)$$

As Λ_1 & Λ_2 are P.D. (because Σ_1 & Σ_2 are P.D.) & $W \Lambda_1 W^T$ is a P.D. when Λ_1 is a P.D.

and sum of two P.D. matrices is a P.D. $\therefore$ $W \Lambda_1 W^T + \Lambda_2$ is a P.D. & hence invertible.

COMPLETING SQUARES BY INSPECTION (as in set-1)

$$\propto \exp\left(\left(y + (W \Lambda_1 W^T + \Lambda_2)^{-1} (W \Lambda_1 (b - x) - \Lambda_2 \mu_2) \right)^T (W \Lambda_1 W^T + \Lambda_2) \left(y + (W \Lambda_1 W^T + \Lambda_2)^{-1} (W \Lambda_1 (b - x) - \Lambda_2 \mu_2) \right) \right)$$

Matching with the form of Gaussian, we can read off the parameters

$$Y|x \sim \mathcal{N}(\mu(x), \Sigma)$$

where $\Sigma = (W \Lambda_1 W^T + \Lambda_2)^{-1}$

& $\mu(x) = -\Sigma (W \Lambda_1 (b - x) - \Lambda_2 \mu_2)$