5.8)

a) Joint distribution can be factorized into

$$p(x, y/\theta) = p(y/x, \theta_2)\, p(x/\theta_1) \qquad \theta = (\theta_1, \theta_2)$$

So, table becomes

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

Data

x	y
1	1
1	0
0	0
1	0
1	1
0	0
0	1

b) MLE Solution for likelihood given

$x = (1,1,0,1,1,0,0)$ $y = (1,0,0,0,1,0,1)$ is

$$\hat{\theta} = \underset{\theta_1,\theta_2}{\arg\max} \log \prod_{i=1}^{m} p(D/\theta)$$

$$= \underset{\theta}{\arg\max} . \log\big((\theta_1 . \theta_2)(\theta_1(1-\theta_2))(\theta_2(1-\theta_1))(\theta_1 . (1-\theta_2))$$

$$\cdot (\theta_1 . \theta_2) . (\theta_2 . (1-\theta_1)) . (1-\theta_1)(1-\theta_2)\big)$$

$$= \underset{\theta}{\arg\max} . \log\ \theta_1^4 . (1-\theta_1)^3 . \theta_2^4 (1-\theta_2)^3$$

$$= \underset{\theta}{\arg\max} . (4 \log\theta_1 + 3\log(1-\theta_1) + 4\log\theta_2 + 3\log(1-\theta_2))$$

Lets call this maximization as $J(\theta)$ and take Partial derivatives w.r.t θ_1, θ_2 and set 0

$$\frac{\partial J}{\partial \theta_1} = \frac{4}{\theta_1} - \frac{3}{1-\theta_1} = 0 \Rightarrow 4(1-\theta_1) - 3\theta_1 = 0$$

$$\Rightarrow \theta_1 = 4/7$$

$$\frac{\partial J}{\partial \theta_2} = \frac{4}{\theta_2} - \frac{3}{1-\theta_2} = 0 \Rightarrow \theta_2 = 4/7$$

The marginal likelihood ~~for~~ this data is

$$P(D/\theta, M_2) = \prod_{i=1}^{m} P(x_i, y_i/\theta)$$

$$= \left(\frac{4}{7}\cdot\frac{4}{7}\right)\left(\frac{4}{7}\cdot\frac{3}{7}\right)\left(\frac{4}{7}\cdot\frac{3}{7}\right)\cdot\left(\frac{4}{7}\cdot\frac{3}{7}\right)\cdot\left(\frac{4}{7}\cdot\frac{4}{7}\right)$$

$$\left(\frac{4}{7}\cdot\frac{3}{7}\right)\left(\frac{3}{7}\cdot\frac{3}{7}\right)$$

c) MLE with 4 parameters representing all combination

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \log \prod_{i=1}^{m} P(D/\theta) \quad \text{s.t.} \sum_{i=0}^{1} \theta_{i,i} = 1$$

$$= \underset{\theta}{\operatorname{argmax}} \log\left(\theta_{1,1}\cdot\theta_{1,0}\cdot\theta_{0,0}\cdot\theta_{1,0}\cdot\theta_{1,1}\cdot\theta_{0,0}\cdot\theta_{0,1}\right)$$

$$\text{s.t.} \quad \theta_{0,0}+\theta_{1,1}+\theta_{1,0}+\theta_{0,1}=1$$

$$= \underset{\theta}{\operatorname{argmax}} \left(2\log\theta_{1,1} + 2\log\theta_{1,0} + 2\log\theta_{0,0} + 1\log\theta_{0,1}\right)$$

~~let call this as J(θ) and do same as~~ above,

~~$\frac{\partial J}{\partial \theta_{0,0}} = \frac{2}{\theta_{0,0}}$~~

lets write this using lagrangian multiplier

$$J(\theta) = 2\log\theta_{1,1} + 2\log\theta_{1,0} + 2\log\theta_{0,0} + 1\log\theta_{0,1} - \lambda\left(\theta_{0,0}+\theta_{0,1}+\theta_{1,0}+\theta_{1,1}-1\right) = 0.$$

$$\frac{\partial J}{\partial \theta_{0,0}} = \frac{2}{\theta_{0,0}} - \lambda = 0 \Rightarrow \theta_{0,0} = \frac{2}{\lambda} \rightarrow ①$$

Same like this $\theta_{1,0} = \theta_{1,1} = 2/\lambda$, $\theta_{0,1} = 1/\lambda \rightarrow ②$

we can solve λ by using $\sum_{i=0}^{1} \theta_{i,i} = 1$

$$\Rightarrow 7/\lambda = 1 \Rightarrow \lambda = 7$$

By substituting $\lambda = 7$ in above

$\theta_{0,0} = \theta_{1,0} = \theta_{1,1} = 2/7$ and $\theta_{0,1} = 1/7$

and marginal likelihood with 4-parameters is,

$$P(D/\hat{\theta}, M_4) = \prod_{i=1}^{4} p(x_i, y_i/\theta)$$

$$= \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{2}{7} \cdot \frac{1}{7} = \frac{8}{7^4}$$

d) For leave-one-out cross validation, we have to remove one data point at every time and see how well the new parameters explain remaining data.

Lets see <u>first for model 1:</u>

$$L(1) = \sum_{i=1}^{m} \log p(x_i, y_i / \hat{\theta}(D_{-i}))$$

Here samples size is $7 \Rightarrow m = 7$

So, for $i=1$, $\hat{\theta}_{D_{-1}} = [\theta_1 = 3/6, \theta_2 = 3/6]$ and $p(x_1, y_1/\theta_{D_{-1}}) = \log \frac{3}{6} \cdot \frac{3}{6}$

$= \log 1/4$

for $i=2$, $\hat{\theta}_{D_{-2}} = [\theta_1 = 3/6, \theta_2 = 4/6]$ and $p(x_2, y_2/\theta_{D_{-2}}) = \log 3/6 \cdot 4/6 = \log 1/3$

for $i=3$, $\hat{\theta}_{D_{-3}} = [\theta_1 = 4/6, \theta_2 = 3/6]$ and $p(x_3, y_3/\theta_{D_{-3}}) = \log 1/3$.

for $i=4$, $\hat{\theta}_{D_{-4}} = [\theta_1 = 3/6, \theta_2 = 4/6]$ and $p(x_4, y_4/\theta_{D_{-4}}) = \log 1/3$

for $i=5$, $\hat{\theta}_{D_{-5}} = [\theta_1 = 3/6, \theta_2 = 3/6]$ and $p(x_5, y_5/\theta_{D_{-5}}) = \log 1/4$

for $i=6$, $\hat{\theta}_{D_{-6}} = [\theta_1 = 4/6, \theta_2 = 3/6]$ and $p(x_6, y_6/\theta_{D_{-6}}) = \log 1/3$

for $i=7$, $\hat{\theta}_{D-7} = [\theta_1 = 4/6, \theta_2 = 4/6]$ and $p(x_7, y_7 / \theta_{D-7}) =$

$= \log 4/6 \cdot 4/6 = \log 4/9$

so, the Cross Validated likelihood for mode M_2 is

$L(1) = 2\log 1/4 + 4\log 1/3 + \log 4/9$

lets do same for model 2:

$$L(2) = \sum_{i=1}^{n} \log p(x_i, y_i / \hat{\theta}_{D-i})$$

for $i=1$, $\hat{\theta}_{D-1} = [\theta_{1,1} = 1/6, \theta_{1,0} = 2/6, \theta_{0,0} = 2/6, \theta_{0,1} = 1/6]$

~~and~~ $p(x_1, y_1 / \hat{\theta}_{D-1}) = \log 1/6$

for $i=2$, $\hat{\theta}_{D-1} = [\theta_{1,1} = 2/6, \theta_{1,0} = 1/6, \theta_{0,0} = 2/6, \theta_{0,1} = 1/6]$

and $p(x_2, y_2 / \hat{\theta}_{D-2}) = \log 1/6$

like this for $i = 3,4,5,6$ we get $p(x_i, y_i / \theta_{D-i}) = \log 1/6$

for $i=7$, $\hat{\theta}_{D-7} = [\theta_{1,1} = 2/6, \theta_{1,0} = 2/6, \theta_{0,0} = 2/6, \theta_{0,1} = 0]$

and $p(x_7, y_7 / \hat{\theta}_{D-7}) = \log 0$
$\Rightarrow$ undefined.

So, for mode M_4, as we saw ~~training~~ examples in predictions that are not there in training samples, solution is undefined. This is because in mode M_4 we overfit the data and set $\theta_{0,1} = 0$, as we didn't see $x=0, y=1$ in training samples. We'll choose M_2 for this reason.

b.ii) We have completely random dataset with N_1 examples from class 1, and N_2 from class ~~1~~ 2, (with equal proportions i.e $N_1 = N_2$). Based on this intuition as there is no learning here,

$$P(\hat{y} = \text{class } 1) = \frac{N_1}{N_1+N_2} = 1/2$$

$$P(\hat{y} = \text{class } 2) = \frac{N_2}{N_1+N_2} = 1/2$$

Misclassification rate $E(\text{error}) = 1 \cdot P(\hat{y} \neq \text{correct class}) + 0 \cdot P(\hat{y} = \text{correct class})$

$$= 1/2$$

The best misclassification rate is 1/2.

Suppose we use leave one out - cross validation then

~~for all -Nf,~~

if ith ~~class~~ example belongs to class 1, then prediction for ~~i~~

~~ex class 1 ⇒ $P(\hat{y} = \text{class } 1 / i \in \text{class1})$~~ $= \frac{N_1 - 1}{N - 1}$

and for class 2 ⇒ $P(\hat{y} = \text{class2} / i \in \text{class1}) = \frac{N_2}{N_1+N_2-1} = \frac{N_2}{N-1}$

and suppose if ith class belongs to class 2, then

$$P(\hat{y} = \text{class } 1 / i \in \text{class } 2) = \frac{N_1}{N-1}$$

& $P(\hat{y} = \text{class2} / i \in \text{class } 2) = \frac{N_2 - 1}{N - 1}$

Estimated misclassification rate

ith example is $p(\hat{y}/i\in class 1) = \frac{N_1-1}{N-1}$.

in same way, if ith example belongs to class 2 then

$$p(\hat{y}/i\in class 2) = \frac{N_2-1}{N-1}$$

then total misclassifcation rate can be

$$= \frac{N_2}{N}\left(1-\frac{N_2-1}{N-1}\right) + \frac{N_1}{N}\left(1-\frac{N_1-1}{N-1}\right)$$

$$= \frac{N_2\cdot(N_1-1)}{N(N-1)} + \frac{N_1\cdot(N_2-1)}{N(N-1)}$$

$$= \frac{2N_1N_2-N}{N(N-1)}$$

6.2) $p(y_i/\theta_i) \sim N(\theta_i, \sigma^2)$ and ~~δ~~ $p_\eta(\theta) \sim N(m_0, T_0^2)$

$\eta = (m_0, T_0^2)$

σ^2 is given 2500

ML-II estimate for m_0, T_0^2 is

$$\hat{\eta} = \underset{\eta}{\operatorname{argmax}} \int p(D/\theta)\cdot p_\eta(\theta)\, d\theta$$

$$= \underset{\eta}{\operatorname{argmax}} \int \cdot e^{-\frac{1}{2}\sum_{i=1}^{m}\frac{(y_i-\theta)^2}{\sigma^2}} \cdot e^{-\frac{1}{2}\cdot\left(\frac{\theta-m_0}{T_0^2}\right)^2} . d\theta.$$

$= \cancel{\underset{\eta}{\operatorname{argmax}} \int e^{-1/2}}$

~~let find out the quantity from~~ given data

$$\sum_{i=1}^{m} \frac{(y_i - \theta)^2}{\sigma^2} = \frac{1}{500}\left((1505-\theta)^2 + (15\ldots\right.$$

Based on equations from below

if, $p(D/\theta) = N(\bar{y}/\theta, \sigma^2/n)$

and $p(\theta) = N(\theta/m_0, \tau_0^2)$ then

$$p(\theta/D) = N(\theta/m_N, \tau_N^2)$$

$$\cancel{m_N = \tau_N^2 \left(\sigma^2 (N\bar{y}) + \tau_0^2\right.}$$

$$1/\tau_N^2 = 1/\tau_0^2 + \frac{N}{\sigma^2}$$

and $m_N = \tau_N^2 \left(\frac{N\bar{y}}{\sigma^2} + \frac{m_0}{\tau_0^2} \right)$

As we are trying to maximize marginal liklihood

$$p_\eta(D/\theta) = \int p(D/\theta) \cdot p_\eta(\theta)\, d\theta$$

$$= \int N(\theta/m_N, \tau_N^2)\, d\theta$$

$$= \sqrt{2\pi\tau_N^2}$$

So $\underset{\eta}{\operatorname{argmax}}\, p_\eta(D/\theta) = \underset{\eta}{\operatorname{argmax}} \sqrt{2\pi\tau_N^2}$

$$= \underset{\eta}{\operatorname{argmax}} \cdot \sqrt{2\pi} \cdot \left(\frac{1}{\tau_0^2} + \frac{N}{\sigma^2}\right)^{-1/2}$$