

# Chapter 5

stevenjin8

August 5, 2021

## Exercise 1

a. Since  $x$  and  $y$  are independent, we have

$$p(x, y|\theta) = p(y|x, \theta)p(x|\theta).$$

The author gives both  $p(y|x, \theta)$  and  $p(x|\theta)$ . Plugging in the given values, we get

$$p(x = 0, y = 0|\theta) = \theta_2(1 - \theta_1) \quad (1)$$

$$p(x = 0, y = 1|\theta) = (1 - \theta_2)\theta_1 \quad (2)$$

$$p(x = 0, y = 1|\theta) = (1 - \theta_2)(1 - \theta_1) \quad (3)$$

$$p(x = 1, y = 1|\theta) = \theta_2\theta_1. \quad (4)$$

b. We find the likelihood by plugging in (2) (3) (4) and (5) into the likelihood function:

$$\begin{aligned} p(\mathcal{D}|\hat{\theta}) &= \prod p(x_i, y_i|\theta) \\ &= \theta_1^4\theta_2^4(1 - \theta_1)^3(1 - \theta_2)^3. \end{aligned} \quad (5)$$

The MLE is given by

$$\begin{aligned} \hat{\theta} &= \underset{\theta}{\operatorname{argmax}} p(\mathcal{D}|\theta) \\ &= \underset{\theta}{\operatorname{argmax}} \theta_1^4\theta_2^4(1 - \theta_1)^3(1 - \theta_2)^3. \end{aligned}$$

Since  $p(\mathcal{D}|\theta)$  is differentiable with respect to  $\theta_1$  and  $\theta_2$ , we can differentiate and set equal to zero to obtain  $\hat{\theta}$ :

$$\begin{aligned} \frac{\partial}{\partial \theta_1} p(\mathcal{D}|\theta) &= \frac{\partial}{\partial \theta_1} \theta_1^4\theta_2^4(1 - \theta_1)^3(1 - \theta_2)^3 \\ &= 4\theta_1^3\theta_2^4(1 - \theta_1)^3(1 - \theta_2)^3 - 3\theta_1^4\theta_2^4(1 - \theta_1)^2(1 - \theta_2)^3 \\ &= 0. \end{aligned}$$

Solving for  $\theta_1$ , we get that  $\hat{\theta}_1 = \frac{4}{7}$ . A similar process yields that  $\hat{\theta}_2 = \frac{4}{7}$ . This result is not too surprising as  $\theta_1$  is how often  $x = 1$  in the data and  $\theta_2$

is how often our observer was correct, both of which are  $\frac{4}{7}$ . Plugging in the MLE's of  $\hat{\theta}_1$  and  $\hat{\theta}_2$  into equation (5) gives

$$\begin{aligned} p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2) &= \left(\frac{4}{7}\right)^4 \left(\frac{4}{7}\right)^4 \left(1 - \frac{4}{7}\right)^3 \left(1 - \frac{4}{7}\right)^3 \\ &= \frac{4^8 3^6}{7^{14}}. \end{aligned}$$

c. We take a different approach than part b by maximizing the log-likelihood (rather than the likelihood) and by using Lagrange multipliers (rather than differentiating and setting equal to zero). To put it in math notation, we are trying to maximize

$$\begin{aligned} \log p(\mathcal{D}|\boldsymbol{\theta}, M_4) &= \log \theta_{00}^2 \theta_{01} \theta_{10}^2 \theta_{11}^2 \\ &= 2 \log \theta_{00} + \log \theta_{01} + 2 \log \theta_{10} + 2 \log \theta_{11} \end{aligned} \quad (6)$$

over the constraint

$$G(\boldsymbol{\theta}) = \theta_{00} + \theta_{01} + \theta_{10} + \theta_{11} = 1. \quad (7)$$

We first find the critical points by finding  $\boldsymbol{\theta}$  such that

$$\nabla G(\boldsymbol{\theta}) = \lambda \nabla \log p(\mathcal{D}|\boldsymbol{\theta}, M_4)$$

for some non-zero  $\lambda$ . Finding the gradients is fairly straightforward, we just differentiate with respect to each  $\theta_{ij}$ :

$$\begin{aligned} \nabla \log p(\mathcal{D}|\boldsymbol{\theta}, M_4) &= \left( \frac{2}{\theta_{00}}, \frac{2}{\theta_{10}}, \frac{1}{\theta_{01}}, \frac{2}{\theta_{11}} \right) \\ \nabla G(\boldsymbol{\theta}) &= (1, 1, 1, 1) \end{aligned}$$

Now we solve for  $\lambda$ :

$$\lambda \left( \frac{2}{\theta_{00}}, \frac{1}{\theta_{01}}, \frac{2}{\theta_{10}}, \frac{2}{\theta_{11}} \right) = (1, 1, 1, 1) \quad (8)$$

$$\lambda(2, 1, 2, 2) = \theta_{00}, \theta_{01}, \theta_{10}, \theta_{11}. \quad (9)$$

Plugging into our constraint yields

$$\begin{aligned} 2\lambda + \lambda + 2\lambda + 2\lambda &= 1 \\ \lambda &= \frac{1}{7}. \end{aligned} \quad (10)$$

It follows that  $\hat{\boldsymbol{\theta}}$  is given by

$$\hat{\theta}_{00} = \frac{2}{7}, \hat{\theta}_{01} = \frac{1}{7}, \hat{\theta}_{10} = \frac{2}{7}, \hat{\theta}_{11} = \frac{2}{7}.$$

This result is not too surprising. Each  $\theta_{ij}$  is the probability of an event happening. Intuitively,  $\hat{\theta}_{ij}$  would be how often the corresponding even happened in the given data.

Now that we have found  $\hat{\theta}$ , we can find  $p(\mathcal{D}|\hat{\theta}, M_4)$ :

$$\begin{aligned} p(\mathcal{D}|\hat{\theta}, M_4) &= \theta_{00}^2 \theta_{01} \theta_{10}^2 \theta_{11}^2 \\ &= \left(\frac{2}{7}\right)^2 \left(\frac{1}{7}\right) \left(\frac{2}{7}\right)^2 \left(\frac{2}{7}\right)^2 \\ &= \frac{2^6}{7^7}. \end{aligned}$$

d. We use the same methods as above to find  $p(x_i, y_i | M_j, D_{-i})$ . For the two parameter model, we have:

$i$	$x_i$	$y_i$	$\hat{\theta}_1$	$\hat{\theta}_2$	$p(x_i, y_i   M_2, \hat{\theta}(\mathcal{D}_{-i}))$
0	1	1	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{9}{36}$
1	1	0	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{6}{36}$
2	0	0	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{6}{36}$
3	1	0	$\frac{3}{6}$	$\frac{4}{6}$	$\frac{6}{36}$
4	1	1	$\frac{3}{6}$	$\frac{3}{6}$	$\frac{9}{36}$
5	0	0	$\frac{4}{6}$	$\frac{3}{6}$	$\frac{6}{36}$
6	0	1	$\frac{4}{6}$	$\frac{4}{6}$	$\frac{4}{36}$

$$L(M_2) = \log \frac{9 \cdot 6 \cdot 6 \cdot 6 \cdot 9 \cdot 6 \cdot 4}{36^7} \approx -5.271.$$

As for the four parameter model, we have:

$i$	$x_i$	$y_k$	$\hat{\theta}_{00}$	$\hat{\theta}_{01}$	$\hat{\theta}_{10}$	$\hat{\theta}_{11}$	$p(x_i, y_i   M_4, \hat{\theta}(\mathcal{D}_{-i}))$
0	1	1	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
1	1	0	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
2	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
3	1	0	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
4	1	1	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{1}{6}$	$\frac{1}{6}$
5	0	0	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$
6	0	1	$\frac{2}{6}$	$\frac{0}{6}$	$\frac{2}{6}$	$\frac{2}{6}$	$\frac{1}{6}$

$$L(M_4) = \log \frac{1 \cdot 0 \cdot 1 \cdot 0 \cdot 1 \cdot 1 \cdot 1}{6} = \log \frac{0}{6} = -\infty.$$

Since  $L(M_2) \approx -5.271 > L(M_4) = -\infty$ , CV will pick  $M_2$ .

e. We can use part b and c to answer this question. For model  $M_2$ ,

$$BIC(M_2, \mathcal{D}) = \log \frac{4^8 3^7}{7^{14}} - \frac{2}{2} \log 7 \approx -4.520.$$

For model  $M_4$ ,

$$BIC(M_4, \mathcal{D}) = \log \frac{2^6}{7^7} - \frac{3}{2} \log 7 \approx -5.377.$$

Recall that  $M_4$ , despite having four parameters, only has 3 free parameters because all the parameters must sum to 1.

Once again,  $M_2$  beats out  $M_4$  since  $BIC(M_2, \mathcal{D}) \approx -4.520 > BIC(M_4, \mathcal{D}) \approx -5.377$ .

## Exercise 9

In this question we prove that the posterior median minimizes the posterior  $\ell_1$  loss. Minimizing this loss is particularly useful when we do not want outliers in our data to skew our predictions.

First, we find the expected  $\ell_1$  loss in terms of  $\theta$ :

$$\begin{aligned} \rho(a, y) &= \mathbb{E}_y[L(a, y)] \\ &= \int_{-\infty}^{\infty} |a - y| p(y|\mathbf{x}) dy \\ &= \int_{-\infty}^a (a - y) p(y|\mathbf{x}) dy - \int_a^{\infty} (a - y) p(y|\mathbf{x}) dy \\ &= a \cdot P(a \leq y|\mathbf{x}) - \int_{-\infty}^a y \cdot p(y|\mathbf{x}) dy - a \cdot P(y > a|\mathbf{x}) + \int_a^{\infty} y \cdot p(y|\mathbf{x}) dy, \end{aligned}$$

where  $a$  is our prediction for the unknown value  $y$ .

Next, we find the critical points by differentiating with respect to  $a$  (not  $y$ ) and setting equal to zero:

$$\begin{aligned} \frac{\partial \rho}{\partial a} &= P(a \leq y|\mathbf{x}) + 2a \cdot p(a|\mathbf{x}) - P(a > y|\mathbf{x}) - 2a \cdot p(a|\mathbf{x}) \\ &= P(a \leq y|\mathbf{x}) - P(a > y|\mathbf{x}) \\ &= 0. \end{aligned}$$

Since we also know that  $P(a \leq y) + P(a > y) = 1$  it must be the case that

$$P(a \leq l|\mathbf{x}) = P(a > y|\mathbf{x}) = \frac{1}{2}.$$

### Exercise 10

I am fairly sure the question is wrong and should read along the lines of:

If  $L_{FN} = cL_{FP}$  show that we should pick  $\hat{y} = 1$  iff  $\tau < p(y = 1|\mathbf{x})$ , where  $\tau = \frac{1}{1+c}$ .

( $\implies$ ) The loss matrix is

	$y = 1$	$y = 0$
$\hat{y} = 1$	0	$L_{FP}$
$\hat{y} = 0$	$cL_{FP}$	0

We want to predict  $\hat{y} = 1$  when

$$\begin{aligned}\mathbb{E}[Loss|\hat{y} = 1] &< \mathbb{E}[Loss|\hat{y} = 0] \\ L_{TP} \cdot p(y = 1|\mathbf{x}) + L_{FP} \cdot p(y = 0|\mathbf{x}) &< L_{TN} \cdot p(y = 0|\mathbf{x}) + L_{FN} \cdot p(y = 1|\mathbf{x}) \\ L_{FP} \cdot p(y = 0|\mathbf{x}) &< cL_{FP} \cdot p(y = 1|\mathbf{x}).\end{aligned}$$

Assuming  $L_{FP} \neq 0$ , we can solve for  $p(y = 1|\mathbf{x})$ :

$$\begin{aligned}1 - p(y = 1|\mathbf{x}) &< c \cdot p(y = 1|\mathbf{x}) \\ 1 &< (c + 1)p(y = 1|\mathbf{x}) \\ \frac{1}{c + 1} &< p(y = 1|\mathbf{x}) \\ \tau &< p(y = 1|\mathbf{x}).\end{aligned}$$

( $\Leftarrow$ ) We can apply the same logic as above in reverse order.