# EE5907/EE5027 Week 6: Bayesian Statistics Solutions

### Exercise 5.1

Given that $p(\theta) = \sum_k p(z = k)p(\theta|z = k) = \sum_{z=1}^{K} p(\theta, z)$, where $p(\theta|\mathcal{D})$ is conjugate, and $p(z = k)$ are the (prior) mixing weights, we get

$$\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\
&= \frac{p(\mathcal{D}|\theta)\sum_z p(\theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}|\theta)p(\theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}|\theta, z)p(\theta, z)}{p(\mathcal{D})} \quad \text{because } z \text{ and } D \text{ are conditionally independent given } \theta \\
&= \sum_z \frac{p(D, \theta, z)}{p(\mathcal{D})} \\
&= \sum_z \frac{p(\mathcal{D}, z)p(\theta|\mathcal{D}, z)}{p(\mathcal{D})} \\
&= \sum_z p(z|\mathcal{D})p(\theta|\mathcal{D}, z)
\end{aligned}$$

### Exercise 5.3

a. The minimum risk is obtained if the posterior expected loss

$$\begin{aligned}
\rho(a|x) = \mathbb{E}_{p(y|x)}[L(y, a)] &= \sum_y L(y, a)p(y|x) \\
&= \begin{cases} 0 \cdot p(y = a|x) + \sum_{y \neq a} \lambda_s p(y|x) & 1 \leq a \leq C \\ \sum_y \lambda_r p(y|x) & a = C + 1 \end{cases} \\
&= \begin{cases} \sum_{y \neq a} \lambda_s p(y|x) & 1 \leq a \leq C \\ \lambda_r & a = C + 1 \end{cases}
\end{aligned}$$

is minimized over $a$. Therefore, for optimal $a$ to be $j$, where $1 \leq j \leq C$, we require

$$\rho(a = j|x) \leq \rho(a = i|x) \text{ for all } i \neq j,$$

which can be split into two cases: (1) $1 \leq i \leq C$ and $i \neq j$, and (2) $i = C + 1$ and $i \neq j$:

Case 1: $1 \leq i \leq C$ and $i \neq j$, then

$$\begin{aligned}
\rho(a = j|x) &\leq \rho(a = i|x) \\
\sum_{y \neq j} \lambda_s p(y|x) &\leq \sum_{y \neq i} \lambda_s p(y|x) \\
\lambda_s p(y = i|x) &\leq \lambda_s p(y = j|x) \\
p(y = i|x) &\leq p(y = j|x)
\end{aligned}$$

Case 2: $i = C + 1$ and $i \neq j$

$$\begin{aligned}
\rho(a = j|x) &\leq \rho(a = i|x) \\
\sum_{y \neq j} \lambda_s p(y|x) &\leq \sum_{y} \lambda_r p(y|x) \\
\lambda_s(1 - p(y = j|x)) &\leq \lambda_r \\
1 - p(y = j|x) &\leq \frac{\lambda_r}{\lambda_s} \\
p(y = j|x) &\geq 1 - \frac{\lambda_r}{\lambda_s}
\end{aligned}$$

b. For $\frac{\lambda_r}{\lambda_s} = 0$, there is zero cost in rejection, so we should always reject. As $\lambda_r/\lambda_s \to 1$, the cost of rejection is the same as the cost of guessing the wrong label. Therefore we should always guess the class label with the highest posterior probability and never guess the reject option.

## Exercise 5.7

Given that the expectation over $\Delta$ is with respect to

$$p(\Delta|\mathcal{D}) = \sum_{m \in M} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D})$$

We have for Bayes model averaging:

$$\begin{aligned}
&\mathbb{E}_{p(\Delta|\mathcal{D})} \left[ L\left(\Delta, p^{BMA}\right) \right] \\
&= \mathbb{E}_{p(\Delta|\mathcal{D})} \left[ -\log p^{BMA}(\Delta) \right] \\
&= -\int \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \log \sum_{m \in \mathcal{M}} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \, d\Delta
\end{aligned}$$

Similarly, for plugin approximation:

$$\begin{aligned}
&\mathbb{E}_{p(\Delta|\mathcal{D})} \left[ L\left(\Delta, p^{M}\right) \right] \\
&= \mathbb{E} \left[ -\log p^{M}(\Delta) \right] \\
&= \mathbb{E}_{p(\Delta|\mathcal{D})} \left[ -\log p(\Delta|m', \mathcal{D}) \right] \\
&= -\int \sum_{m \in M} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \log p(\Delta|m', \mathcal{D}) \, d\Delta
\end{aligned}$$

The difference between the two approaches is given by

$$\begin{aligned}
&\mathbb{E}\left[L\left(\Delta, p^{M}\right)\right] - \mathbb{E}\left[L\left(\Delta, p^{BMA}\right)\right] \\
&= \int \sum_{m \in M} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \left[ -\log p(\Delta|m', \mathcal{D}) + \log \sum_{m \in M} p(\Delta|m, \mathcal{D}) p(m|\mathcal{D}) \right] d\Delta \\
&= \int p^{BMA}(\Delta) \log \frac{p^{BMA}(\Delta)}{p^{M}(\Delta)} \, d\Delta \\
&= \mathbb{KL}\left(p^{BMA} || p^{M}\right) \geq 0
\end{aligned}$$

Hence we have

$$\mathbb{E}\left[L\left(\Delta, p^{M}\right)\right] \geq \mathbb{E}\left[L\left(\Delta, p^{BMA}\right)\right]$$

## Exercise 5.8

a. $p(x, y|\theta) = p(x|\theta)p(y|x, \theta) = p(x|\theta_1)p(y|x, \theta_2)$ Given that

$$p(x|\theta_1) = \begin{cases} 1 - \theta_1 & x = 0 \\ \theta_1 & x = 1 \end{cases}$$

and $p(y|x, \theta_2)$ is given by

| | $y = 0$ | $y = 1$ |
|---|---|---|
| $x = 0$ | $\theta_2$ | $1 - \theta_2$ |
| $x = 1$ | $1 - \theta_2$ | $\theta_2$ |

Thus $p(x, y|\theta)$ is

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

b. In the dataset, (0,0) appears 2 times, (0,1) appears 1 times, (1,0) appears 2 times and (1,1) appears 2 times, i.e.,

| | $y = 0$ | $y = 1$ |
|---|---|---|
| $x = 0$ | 2 | 1 |
| $x = 1$ | 2 | 2 |

thus

$$\begin{aligned}
p(\mathcal{D}|\theta_1, \theta_2) &= [(1 - \theta_1)\theta_2]^2 \times [(1 - \theta_1)(1 - \theta_2)] \times [\theta_1(1 - \theta_2)]^2 \times [\theta_1\theta_2]^2 \\
&= (1 - \theta_1)^3\theta_1^4 \times (1 - \theta_2)^3\theta_2^4 \\
\log p(\mathcal{D}|\theta_1, \theta_2) &= 3\log(1 - \theta_1) + 4\log\theta_1 + 3\log(1 - \theta_2) + 4\log\theta_2
\end{aligned}$$

Take the partial derivatives and set to 0, we get

$$\begin{aligned}
\frac{\partial \log p(\mathcal{D}|\theta_1, \theta_2)}{\partial \theta_1} &= -\frac{3}{1 - \theta_1} + \frac{4}{\theta_1} = 0 \\
\frac{\partial \log p(\mathcal{D}|\theta_1, \theta_2)}{\partial \theta_2} &= -\frac{3}{1 - \theta_2} + \frac{4}{\theta_2} = 0
\end{aligned}$$

We have $\theta_1 = \theta_2 = \frac{4}{7}$. Hence

$$p(\mathcal{D}|\hat{\theta}, M_2) = \left(1 - \frac{4}{7}\right)^3 \left(\frac{4}{7}\right)^4 \left(1 - \frac{4}{7}\right)^3 \left(\frac{4}{7}\right)^4 = \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6$$

c. In the model with 4 parameters, we have

$$\hat{\theta}^{ML} = \operatorname*{argmax}_{\theta} p(\mathcal{D}|\theta) = \operatorname*{argmax}_{\boldsymbol{\theta}} \ \theta_{11}^2 \cdot \theta_{10}^2 \cdot \theta_{00}^2 \cdot \theta_{01}$$

We can take the derivatives and so on, but in this case, this is essentially the multinomial distribution, and so the ML estimate corresponds to fraction of empirical count in each category, i.e., $\hat{\theta}_{11} = \hat{\theta}_{10} = \hat{\theta}_{00} = \frac{2}{7}$ and $\hat{\theta}_{01} = 1/7$. Therefore

$$p(\mathcal{D}|\hat{\theta}, M_4) = \left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2 \cdot \left(\frac{2}{7}\right)^2 \cdot \frac{1}{7} = \left(\frac{2}{7}\right)^6 \frac{1}{7}$$

d. For leave-one-out cross-validation, when $x = 0, y = 1$ is left out, model $M_4$ will assign 0 probability to $x = 0, y = 1$, and so $L(M_4) = -\infty$. On the other hand, $L(M_2)$ is finite, so CV will pick $M_2$.

e. For $M_2$ model, we have

$$\text{BIC}(M_2, \mathcal{D}) = \log p(\mathcal{D}|\hat{\theta}_{MLE}) - \frac{\text{dof}(M_2)}{2} \log N = \log \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 - \frac{2}{2} \log 7 \approx -11.51$$

For $M_4$ model, we have

$$\text{BIC}(M_4, \mathcal{D}) = \log p(\mathcal{D}|\hat{\theta}_{MLE}) - \frac{\text{dof}(M_4)}{2} \log N = \log \left(\frac{2}{7}\right)^6 \frac{1}{7} - \frac{3}{2} \log 7 \approx -12.38$$

Since $\text{BIC}(M_2, \mathcal{D}) > \text{BIC}(M_4, \mathcal{D})$. BIC will also prefer $M_2$ model.

## Exercise 5.9

L1 loss is defined as:

$$L(y, a) = |y - a|$$

The posterior expected loss is given by

$$\rho(a|x) = \mathbb{E}[L(y, a)|x] = \sum_y |y - a| p(y|x) = \sum_{y \geq a} (y - a) p(y|x) + \sum_{y < a} (a - y) p(y|x)$$

Differentiating with respect to $a$, we get

$$\frac{\partial \rho(a|x)}{\partial a} = -\sum_{y \geq a} p(y|x) + \sum_{y < a} p(y|x) = -P(y \geq a|x) + P(y < a|x) = 0$$

$$\implies P(y < a|x) = P(y \geq a|x) = 0.5$$

Thus posterior median minimizes L1 loss. This assumes $y$ is discrete. The derivation is similar for when $y$ is continuous by replacing sum with integral, except differentiation is slightly more tricky since we have to differentiate inside the integral (see https://en.wikipedia.org/wiki/Differentiation_under_the_integral_sign).

## Q6: Using an imperfect oracle

- (i) Here are the expected loss

  $$R(\alpha_0) = 0.4 * 6 = 2.4$$
  $$R(\alpha_1) = 0.6 * 6 = 3.6$$
  $$R(\alpha_h) = 2$$

  Therefore we should choose $\alpha_h$

- (ii) Here are the expected loss

  $$R(\alpha_0) = 0.1 * 6 = 0.6$$
  $$R(\alpha_1) = 0.9 * 6 = 5.4$$
  $$R(\alpha_h) = 2$$

  Therefore we should choose $\alpha_0$

- (iii) The general expected loss is given by

  $$R(\alpha_0) = 6p_1$$
  $$R(\alpha_1) = 6(1 - p_1)$$
  $$R(\alpha_h) = 2$$

  We should choose $\alpha_0$ if $R(\alpha_0) < R(\alpha_1) \implies p_1 < 0.5$ **and** $R(\alpha_0) < R(\alpha_h) \implies p_1 < 1/3$
  We should choose $\alpha_1$ if $R(\alpha_1) < R(\alpha_0) \implies p_1 > 0.5$ and $R(\alpha_1) < R(\alpha_h) \implies p_1 > 2/3$
  Therefore we should choose $\alpha_0$ if $p_1 < 1/3$, $\alpha_1$ if $p_1 > 2/3$ and $\alpha_h$ otherwise.

- (iv) If the human is correct 95% of the time, then the general expected cost of $\alpha_h$ is $0.95 \times 2 + 0.05 \times 8 = 2.3$

  Therefore, we should choose $\alpha_0$ if $p_1 < 0.5$ and $R(\alpha_0) < R(\alpha_h) \implies p_1 < 2.3/6 = 0.383$
  And we should choose $\alpha_1$ if $p_1 > 0.5$ and $R(\alpha_1) < R(\alpha_h) \implies 6(1-p_1) < 2.3 = 1-0.38 = 0.617$
  Therefore we should choose $\alpha_0$ if $p_1 < 0.383$, $\alpha_1$ if $p_1 > 0.617$ and $\alpha_h$ otherwise.