Team ML_ShibaInu

Group Members:

1. 2211649 - Trần Đình Đăng Khoa
2. 2252938 - Phan Chí Vỹ
3. 2252734 - Nguyễn Đức Tâm
4. 2252338 - Huỳnh Kiệt Khải
5. 2252289 - Hà Tuấn Khang

Homework 1: Statistics Review

Exercise 4.5

Bayesian Information Criterion (BIC) for a 2D Discrete Distribution

Let $x \in \{0, 1\}$ denote the result of a coin toss ($x = 0$ for tails, $x = 1$ for heads). The coin is potentially biased, so that heads occurs with probability θ_1. Suppose that someone else observes the coin flip and reports to you the outcome, y. But this person is unreliable and only reports the result correctly with probability θ_2; i.e., $p(y|x, \theta_2)$ is given by:

	$y = 0$	$y = 1$
$x = 0$	(θ_2)	(1 - θ_2)
$x = 1$	(1 - θ_2)	(θ_2)

Assume that θ_2 is independent of x and θ_1.
a. Write down the joint probability distribution $p(x, y|\theta)$ as a 2×2 table, in terms of $\boldsymbol{\theta} = (\theta_1\theta_2)$

ANSWER:

The joint distribution is $p(x, y|\boldsymbol{\theta}) = p(x|\theta_1)p(y|x, \theta_2)$

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

b. Suppose have the following dataset: $x=(1,1,0,1,1,0,0), \quad y=(1,0,0,0,1,0,1)$. What are the MLEs for θ_1 and θ_2? Justify your answer. Hint: note that the likelihood function factorizes,

$$p(x,y|\boldsymbol{\theta}) = p(y|x,\theta_2)p(x|\theta_1)$$

What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2)$ where M_2 denotes this 2-parameter model?

ANSWER:

The log likelihood is

$$\log p(\mathcal{D}|\boldsymbol{\theta}) = \sum_i \log p(x_i|\theta_1) + \sum_i \log p(y_i|x_i,\theta_2)$$

Hence we can optimize each term separately. For θ_1,we have

$$\hat{\theta_1} = \frac{\sum_i I(x_i=1)}{n} = \frac{N(x=1)}{N} = \frac{4}{7} = 0.5714$$

For θ_2 we have

$$\hat{\theta_2} = \frac{\sum_i I(x_i=y_i)}{n} = \frac{N(x=y)}{N} = \frac{4}{7}$$

The likelihood is

$$\begin{aligned} p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2) &= \left(\frac{4}{7}\right)^{N(x=1)} \left(\frac{3}{7}\right)^{N(x=0)} \left(\frac{4}{7}\right)^{N(x=y)} \left(\frac{3}{7}\right)^{N(x\neq y)} \\ &= \left(\frac{4}{7}\right)^4 \left(\frac{3}{7}\right)^3 \left(\frac{4}{7}\right)^4 \left(\frac{3}{7}\right)^3 \\ &= \left(\frac{4}{7}\right)^8 \left(\frac{3}{7}\right)^6 \\ &\approx 7.04 \times 10^{-5} \end{aligned}$$

c. Now consider a model with 4 parameters, $\boldsymbol{\theta} = (\theta_{0,0}, \theta_{0,1}, \theta_{1,0}, \theta_{1,1})$, representing $p(x,y|\boldsymbol{\theta}) = \theta_{x,y}$
(Only 3 of these parameters are free to vary, since they must sum to one.)
What is the MLEs of $\boldsymbol{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4)$ where M_4 denotes this 4-parameter model?

ANSWER:

The table of joint counts is:

	$y=0$	$y=1$
$x=0$	2	1
$x=1$	2	2

We can view this as a multinomial distribution with 4 states. By normalizing the counts, we obtain the MLE:

	$y=0$	$y=1$
$x=0$	$2/7$	$1/7$
$x=1$	$2/7$	$2/7$

The likelihood is

$$\begin{aligned} p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4) &= \theta_{00}^{N(x=0,y=0)}\theta_{01}^{N(x=0,y=1)}\theta_{10}^{N(x=1,y=0)}\theta_{11}^{N(x=1,y=1)} \\ &= \left(\frac{2}{7}\right)^2\left(\frac{1}{7}\right)^1\left(\frac{2}{7}\right)^2\left(\frac{2}{7}\right)^2 \\ &= \left(\frac{2}{7}\right)^6\left(\frac{1}{7}\right)^1 \\ &\approx 7.77 \times 10^{-5} \end{aligned}$$

As a result, this likelihood is greater than the previous one, due to the model having a higher number of parameters.

d. Suppose we are not sure which model is correct. We compute the leave-one-out cross validated log likelihood of the 2-parameter model and the 4-parameter model as follows:

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

and $\hat{\theta}(\mathcal{D}_{-i})$ denotes the MLE computed on $\mathcal{D}$ excluding row i. Which model will CV pick and why?

ANSWER:

For M_4, if case 7 is omitted, $\hat{\theta}_{01} = 0$, resulting in $p(x_7, y_7 | m_4, \hat{\theta}) = 0$, and thus $L(m_4) = -\infty$. In contrast, $L(m_2)$ remains finite, as all counts stay non-zero when a single case is excluded. Therefore, CV will favor M_2 because M_4 is overfitting.

e. Recall that an alternative to CV is to use the BIC score, defined as

$$\mathrm{BIC}(M, \mathcal{D}) \triangleq \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}) - \frac{\mathrm{dof}(M)}{2} \log N$$

where $\mathbf{dof}(M)$ is the number of free parameters in the model. Compute the BIC scores for both models (use log base e). Which model does BIC prefer?

ANSWER:

The BIC score is

$$\mathrm{BIC}(m) = \log p(\mathcal{D}|\hat{\boldsymbol{\theta}}, m) - \frac{\mathrm{dof}(m)}{2} \log n$$

where $n = 7$. For M_2, we have $\mathbf{dof} = 2$, so

$$\mathrm{BIC}(m_2) = 8 \log\left(\frac{4}{7}\right) + 6 \log\left(\frac{3}{7}\right) - \frac{2}{2} \log(7) = -11.5066$$

For M_4, we have $\mathbf{dof} = 3$ because of the sum-to-one constraint, so that

$$\mathrm{BIC}(m_4) = 6 \log\left(\frac{2}{7}\right) + 1 \log\left(\frac{1}{7}\right) - \frac{3}{2} \log(7) = -12.3814$$

So BIC prefers m_2.

Exercise 4.6

A mixture of conjugate priors is conjugate †

Consider a mixture prior

$$p(\theta) = \sum_k p(z = k) p(\theta | z = k)$$

where each $p(\theta|z = k)$ is conjugate to the likelihood. Prove that this is a conjugate prior.

ANSWER:

If the prior distribution is a mixture, the posterior distribution will also be a mixture.

$$
\begin{aligned}
p(\theta|\mathcal{D}) &= \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \\
&= \frac{p(\mathcal{D}|\theta)\sum_k p(Z=k)p(\theta|Z=k)}{\int p(\mathcal{D}|\theta')\sum_{k'} p(Z=k')p(\theta'|Z=k')d\theta'} \\
&= \frac{\sum_k p(Z=k)p(\mathcal{D},\theta|Z=k)}{\sum_{k'} p(Z=k')\int p(\mathcal{D},\theta'|Z=k')d\theta'} \\
&= \frac{\sum_k p(Z=k)p(\theta|\mathcal{D},Z=k)p(\mathcal{D}|Z=k)}{\sum_{k'} p(Z=k')p(\mathcal{D}|Z=k')} \\
&= \sum_k \left[\frac{p(Z=k)p(\mathcal{D}|Z=k)}{\sum_{k'} p(Z=k')p(\mathcal{D}|Z=k')}\right] p(\theta|\mathcal{D},Z=k) \\
&= \sum_k p(Z=k|\mathcal{D})p(\theta|\mathcal{D},Z=k)
\end{aligned}
$$

where $p(Z=k)=\pi_k$ are the prior mixing weights, and $p(Z=k|\mathcal{D})$ are the posterior mixing weights given by

$$
p(Z=k|\mathcal{D}) = \frac{p(Z=k)p(\mathcal{D}|Z=k)}{\sum_{k'} p(Z=k')p(\mathcal{D}|Z=k')}
$$

Exercise 4.7

ML estimator σ^2_{mle} is biased

Show that $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$ is a biased estimator of σ^2, i.e., show

$$
\mathbf{E}_{X_1,,\ldots,X_N \sim \mathcal{N}(\mu,\sigma)}[\hat{\sigma}^2(X_1,\ldots,X_n)] \neq \sigma^2
$$

ANSWER:

Because of the variance of any random variable R is given by $var(R) = E[R^2] - (E[R])^2$, the expected value of the square of a Gaussian random variable X_i with mean μ and variance σ^2 is $E[X_i^2] = var(X_i) + (E[X_i])^2 = \sigma^2 + \mu^2$.

$$
\begin{aligned}
E_{X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma)}[\sigma^2(X_1,\ldots,X_n)] &= E\left[\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \frac{\sum_{j=1}^{n} X_j}{n}\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} nE\left[\left(X_i - \frac{\sum_{j=1}^{n} X_j}{n}\right)^2\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} nE\left[\left(X_i - \frac{\sum_{j=1}^{n} X_j}{n}\right)\left(X_i - \frac{\sum_{j=1}^{n} X_j}{n}\right)\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} nE\left[X_i^2 - \frac{2}{n}X_i\sum_{j=1}^{n} X_j + \frac{1}{n^2}\sum_{j=1}^{n}\sum_{k=1}^{n} X_jX_k\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} nE[X_i^2] - \frac{2}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] + \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{k=1}^{n} E[X_jX_k]
\end{aligned}
$$

Consider the two summations

$$
\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] \quad \text{and} \quad \sum_{j=1}^{n}\sum_{k=1}^{n} E[X_jX_k].
$$

Among the n^2 terms in each summation, n terms meet the condition $i = j$ or $j = k$, hence these terms are of the form $E[X_i^2]$. Due to the linearity of expectation, these terms contribute $nE[X_i^2]$ to the total. The other $n^2 - n$ terms take the form $E[X_iX_j]$ or $E[X_jX_k]$ for $i \neq j$ or $j \neq k$. Given that the X_i are independent samples, linearity of expectation implies that these terms add $(n^2 - n)E[X_i]E[X_j]$ to the summation.

$$
\begin{aligned}
\sum_{i=1}^{n}\sum_{j=1}^{n} E[X_iX_j] &= \sum_{j=1}^{n}\sum_{k=1}^{n} E[X_jX_k] \\
&= nE[X_i^2] + (n^2 - n)E[X_i]E[X_j] \\
&= n(\sigma^2 + \mu^2) + (n^2 - n)\mu\mu \\
&= n\sigma^2 + n\mu^2 + n^2\mu^2 - n\mu^2 \\
&= n\sigma^2 + n^2\mu^2
\end{aligned}
$$

$$
\begin{aligned}
E_{X_1,\ldots,X_n\sim\mathcal{N}(\mu,\sigma)}[\sigma^2(X_1,\ldots,X_n)] &= \frac{1}{n}\sum_{i} n(\sigma^2 + \mu^2) - \frac{2}{n^2}(n\sigma^2 + n^2\mu^2) + \frac{1}{n^3}\sum_{i=1}^{n}(n\sigma^2 + n^2\mu^2) \\
&= \frac{1}{n}(n\sigma^2 + n\mu^2) - \frac{2\sigma^2}{n} - 2\mu^2 + \frac{1}{n^3}(n^2\sigma^2 + n^3\mu^2) \\
&= \sigma^2 + \mu^2 - \frac{2\sigma^2}{n} - 2\mu^2 + \frac{\sigma^2}{n} + \mu^2 \\
&= \sigma^2 - \frac{\sigma^2}{n} \\
&= \frac{n-1}{n}\sigma^2
\end{aligned}
$$

Since the expected value of $\hat{\sigma}^2(X_1,\ldots,X_n)$ is different from the actual variance σ^2, it indicates that σ^2 is not an unbiased estimator. In fact, the maximum likelihood estimator often underestimates the variance. This is expected: with only a single sample, no variance can be

detected. With multiple samples, variance will be observed, but the mean estimate is likely to be skewed towards the sample values, leading to an underestimation of the variance.