VIETNAM NATIONAL UNIVERSITY, HO CHI MINH CITY
UNIVERSITY OF TECHNOLOGY
FACULTY OF COMPUTER SCIENCE AND ENGINEERING

MACHINE LEARNING (CO3117)

Homework 1:

Statistics Review

Team LHPD2

Semester 2, Academic Year 2024 - 2025

Teacher: Nguyen An Khuong
Students: Nguyen Quang Phu - 2252621 (***Leader***)
Nguyen Thanh Dat - 2252145 (Member)
Pham Huynh Bao Dai - 2252139 (Member)
Nguyen Tien Hung - 2252280 (Member)
Nguyen Thien Loc - 2252460 (Member)

HO CHI MINH CITY, FEBRUARY 2025

Contents

Chapter 1

Introduction

1.1 Purpose of This Document

This chapter serves as an introduction to our team, LHPD2, for the Machine Learning (CO3117) course in Semester 2, Academic Year 2024 - 2025. The purpose of this writing is to formally present our team members and confirm our participation in solving the given homework exercise, "Homework 1: Statistics Review".

1.2 Workload

Specifically, our contributions included:

- **Understanding the Problem:** Each member participated in analyzing and discussing the given exercise to ensure a shared understanding of the requirements.
- **Exploring Concepts:** We collectively researched and reviewed relevant concepts, theories, and methods necessary for solving the problems.
- **Reasoning for the Solution:** The team worked together to identify the rationale behind each approach, ensuring that all steps were logical and well-justified.
- **Implementing the Solution:** The solutions were implemented with equal collaboration, where every member contributed to coding, calculations, and documentation.
- **Final Review:** The team jointly reviewed the solutions to verify their correctness, clarity, and coherence before submission.

We believe that this exercise has enhanced our understanding of the concepts involved and strengthened our ability to work collaboratively as a team.

Chapter 2

Problem Description and Our Team Solution

2.1 Exercise 4.5 [BIC for a 2D Discrete Distribution]

(Source: Jaakkola.)

2.1.1 Description

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	$1 - \theta_2$
$x = 1$	$1 - \theta_2$	θ_2

Assume that θ_2 is independent of x and θ_1.

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

b. Suppose we have the following dataset:

$$\boldsymbol{x} = (1, 1, 0, 1, 1, 0, 0), \quad \boldsymbol{y} = (1, 0, 0, 0, 1, 0, 1).$$

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? (You may leave your answer in fractional form if you wish.)

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 MLE of $\boldsymbol{\theta}$? What is $p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4)$, where M_4 denotes this 4-parameter model?

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:

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

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

Hint: Notice how the table of counts changes when you omit each training case one at a time.

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

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

where 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?

2.1.2 Solution

a. The **joint probability distribution** $p(x, y|\boldsymbol{\theta})$ as a 2×2 table

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

b. **Take the logarithm** on both sides:

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

The $\sum log(p(\boldsymbol{x}|\theta_1))$ part:

$$\sum log(p(\boldsymbol{x}|\theta_1)) = \sum log(\theta_1) + \sum log(1 - \theta_1)$$

So from the question we have that $\boldsymbol{x} = (1, 1, 0, 1, 1, 0, 0)$
The number $x_i = 1$: 4
The number $x_i = 0$: 3
7 time coin toss: 4 heads and 3 tails
Hence,

$$log(p(\boldsymbol{x}|\theta_1)) = 4log(\theta_1) + 3log(1 - \theta_1)$$

Taking the derivative respect to θ_1:

$$\frac{\delta}{\delta\theta_1} log(p(x|\theta_1)) = \frac{\delta}{\delta\theta_1} 4log(\theta_1) + \frac{\delta}{\delta\theta_1} 3log(1-\theta_1) = 4 \times \frac{1}{\theta_1} + 3 \times \frac{-1}{1-\theta_1}$$

To find the MLE, set the derivative equal to zero:

$$\frac{\delta}{\delta\theta_1} log(p(x|\theta_1)) = 0$$

This gives:

$$4 \times \frac{1}{\theta_1} + 3 \times \frac{-1}{1-\theta_1} = 0$$

Multiply through:

$$4 \times (1-\theta_1) = 3 \times \theta_1$$

Expand:

$$4 - 4\theta_1 = 3\theta_1$$

Final result:

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

The $\sum log(p(y|x, theta_2))$ part:

$$\sum log(p(y|x,\theta_2)) = \sum log(\theta_2) + \sum log(1-\theta_2)$$

So from the question we have that $\boldsymbol{y} = (1, 0, 0, 0, 1, 0, 1)$.
The number $y_i =$ x: 3
The number $y_i \neq$ x: 4 Do the same as the previous one:

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

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_2) = (\frac{4}{7})^4 + (1-\frac{4}{7})^3 + (\frac{4}{7})^4 + (1-\frac{4}{7})^3 \approx 7.04 \ \times 10^{-5}$$

c. So we can think it as a **multinomial distribution with 4 states**

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

$$p(\mathcal{D}|\hat{\boldsymbol{\theta}}, M_4) = (\frac{1}{7})^1 + (\frac{2}{7})^2 + (\frac{2}{7})^2 + (\frac{2}{7})^2 \approx 7.7 \ \times 10^{-5}$$

d. With:
$\boldsymbol{x} = (1, 1, 0, 1, 1, 0, 0)$
$\boldsymbol{y} = (1, 0, 0, 0, 1, 0, 1)$

2 - parameter:

$$L(m2) = 8 \times log(\frac{3}{6}) + 6 \times log(\frac{2}{6}) \approx -12.14$$

4 - parameter:

$$L(m4) = 3 \times log(\frac{2}{6}) + log(\frac{0}{6}) = -\infty$$

CV will select the 2-parameter model because the 4-parameter model is prone to over-fitting.

e. for dof(2):

$$\text{BIC}(M, \mathcal{D}) = 8 \times log(\frac{4}{7}) + 6 \times log(\frac{3}{7}) - \frac{2}{2} log7 \approx -11.5$$

for dof(4):

$$\text{BIC}(M, \mathcal{D}) = 6 \times log(\frac{24}{7}) + log(\frac{1}{7}) - \frac{3}{2} log7 \approx -12.38$$

BIC prefer 2-parameter model

2.2 Exercise 4.6 [A mixture of conjugate priors is conjugate *]

2.2.1 Description

Consider a mixture prior

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

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

2.2.2 Solution

Consider a mixture prior defined as:

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

where each $p(\theta|z = k)$ is conjugate to the likelihood $p(x|\theta)$.

Step 1: Apply Bayes' Theorem

The posterior distribution is given by Bayes' theorem:

$$p(\theta|x) \propto p(x|\theta) \cdot p(\theta)$$

Substituting the mixture prior:

$$p(\theta|x) \propto p(x|\theta) \cdot \sum_k p(h = k)p(\theta|z = k)$$

We can distribute the likelihood inside the summation:

$$p(\theta|x) \propto \sum_k p(h = k)p(x|\theta)p(\theta|z = k)$$

Step 2: Use the Conjugacy Property

Since $p(\theta|z = k)$ is conjugate to the likelihood $p(x|\theta)$, we have:

$$p(x|\theta)p(\theta|z = k) \propto p(\theta|x, z = k)$$

where $p(\theta|x, z = k)$ is the posterior distribution belonging to the same family as $p(\theta|z = k)$.

Thus, we can rewrite:

$$p(\theta|x) \propto \sum_k p(h=k)p(\theta|x, z=k)$$

This represents a mixture of posterior distributions, each of which is from the same family as its corresponding prior $p(\theta|z=k)$.

Conclusion

Since $p(\theta|x)$ retains the same functional form as the prior mixture (a weighted sum of conjugate distributions), we conclude:

A mixture of conjugate priors is itself a conjugate prior.

2.3 Exercise 4.7 [ML estimator σ^2_{MLE} is biased]

2.3.1 Description

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

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

Hint: Note that $X_1,\ldots,X_N$ are independent, and use the fact that the expectation of a product of independent random variables is the product of the expectations.

2.3.2 Solution

To demonstrate that $\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(X_n - \hat{\mu})^2$ is biased, we compute its expectation:

Expand the estimator

The estimator can be rewritten as:

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(X_n^2 - 2X_n\hat{\mu} + \hat{\mu}^2),$$

where $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} X_n$ is the sample mean.

$$\hat{\sigma}^2_{\text{MLE}} = \frac{1}{N}\sum_{n=1}^{N}(X_n^2 - 2X_n\hat{\mu} + \hat{\mu}^2) \tag{2.1}$$

$$= \frac{1}{N}\sum_{n=1}^{N} X_n^2 - \frac{2}{N}\sum_{n=1}^{N} X_n\hat{\mu} + \frac{1}{N}\sum_{n=1}^{N}\hat{\mu}^2 \tag{2.2}$$

$$= \frac{1}{N}\sum_{n=1}^{N} X_n^2 - 2\hat{\mu}^2 + \hat{\mu}^2 \tag{2.3}$$

$$= \frac{1}{N}\sum_{n=1}^{N} X_n^2 - \hat{\mu}^2 \tag{2.4}$$

Compute key expectations

- For X_n^2:

$$\mathbf{E}[X_n^2] = \text{Var}(X_n) + (\mathbf{E}[X_n])^2 = \sigma^2 + \mu^2 \quad (\text{since } X_n \sim \mathcal{N}(\mu, \sigma^2)).$$

  Thus:

$$\mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n^2\right] = \frac{1}{N}\sum_{n=1}^{N}(\sigma^2 + \mu^2) = \sigma^2 + \mu^2.$$

- For $\hat{\mu}^2$: The sample mean $\hat{\mu}$ has variance $\text{Var}(\hat{\mu}) = \frac{\sigma^2}{N}$. Therefore:

$$\mathbf{E}[\hat{\mu}^2] = \text{Var}(\hat{\mu}) + (\mathbf{E}[\hat{\mu}])^2 = \frac{\sigma^2}{N} + \mu^2.$$

Combine expectations

The expectation of the MLE estimator is:

$$\mathbf{E}[\hat{\sigma}^2_{\text{MLE}}] = \mathbf{E}\left[\frac{1}{N}\sum_{n=1}^{N} X_n^2\right] - \mathbf{E}[\hat{\mu}^2].$$

Substituting the results:

$$\mathbf{E}[\hat{\sigma}^2_{\text{MLE}}] = (\sigma^2 + \mu^2) - \left(\frac{\sigma^2}{N} + \mu^2\right) = \sigma^2 - \frac{\sigma^2}{N}.$$

Simplifying:

$$\mathbf{E}[\hat{\sigma}^2_{\text{MLE}}] = \frac{N-1}{N}\sigma^2.$$

Conclusion

Since $\mathbf{E}[\hat{\sigma}^2_{\text{MLE}}] \neq \sigma^2$, the **MLE estimator is biased**. The bias arises because the sample mean $\hat{\mu}$ (not the true mean μ) is used, reducing the expected value by a factor of $\frac{N-1}{N}$.

$$\boxed{\mathbf{E}[\hat{\sigma}^2_{\text{MLE}}] = \frac{N-1}{N}\sigma^2 \neq \sigma^2}$$