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

MACHINE LEARNING (CO3117)

Homework 4:

Bayesian Inference

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 4: Bayesian Inference".

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 - Our Solution

2.1 Question 5.1

2.1.1 Description

(Source: [**DHS01, Q2.13**]) In many classification problems one has the option either of assigning $\mathbf{x}$ to class j or, if you are too uncertain, of choosing the **reject option**. If the cost for rejection is less than the cost of falsely classifying the object, it may be the optimal action. Let α_i mean you choose action i, for $i = 1, \ldots, C+1$, where C is the number of classes and $C+1$ is the reject action. Let $Y = j$ be the true (but unknown) **state of nature**. Define the loss function as follows:

$$\lambda(\alpha_i|Y=j) = \begin{cases} 0 & \text{if } i = j \text{ and } i, j \in \{1, \ldots, C\} \\ \lambda_r & \text{if } i = C+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

In other words, you incur 0 loss if you correctly classify, you incur λ_r loss (cost) if you choose the reject option, and you incur λ_s loss (cost) if you make a substitution error (misclassification).

(a) Show that the minimum risk is obtained if we decide $Y = j$ if $p(Y = j|\mathbf{x}) \geq p(Y = k|\mathbf{x})$ for all k (i.e., j is the most probable class) and if $p(Y = j|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$; otherwise, we decide to reject.

(b) Describe qualitatively what happens as λ_r/λ_s is increased from 0 to 1 (i.e., the relative cost of rejection increases).

2.1.2 Solution of Question a

To determine the decision rule that minimizes the expected risk, we analyze the costs associated with rejecting a classification versus assigning the most probable class.

Step 1: Define the Risk of Each Action

We need to choose between rejecting the classification with risk λ_r or selecting the most probable class, $j_{\max}$, where:

$$j_{\max} = \arg\max_j p(Y = j|\mathbf{x})$$

If we choose $j_{\max}$, the expected risk due to misclassification is:

$$\lambda_s \sum_{j \neq j_{\max}} p(Y = j|\mathbf{x}) = \lambda_s(1 - p(Y = j_{\max}|\mathbf{x}))$$

Thus, selecting $j_{\max}$ is preferable if the expected risk of rejection λ_r is greater than or equal to the expected misclassification risk:

$$\lambda_r \geq \lambda_s(1 - p(Y = j_{\max}|\mathbf{x}))$$

Rearranging, we obtain the decision criterion:

$$p(Y = j_{\max}|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$$

If this condition is met, we classify $\mathbf{x}$ as $j_{\max}$; otherwise, we reject.

Step 2: Justification for Picking $j_{\max}$

If we decide to classify instead of rejecting, we must always pick the most probable class. Suppose we choose a non-maximal class $k \neq j_{\max}$. The misclassification risk in this case is:

$$\lambda_s \sum_{j \neq k} p(Y = j|\mathbf{x}) = \lambda_s(1 - p(Y = k|\mathbf{x}))$$

Since $p(Y = k|\mathbf{x}) \leq p(Y = j_{\max}|\mathbf{x})$, it follows that:

$$\lambda_s(1 - p(Y = k|\mathbf{x})) \geq \lambda_s(1 - p(Y = j_{\max}|\mathbf{x}))$$

This confirms that selecting $j_{\max}$ minimizes risk when classification is chosen.

2.1.3 Solution of Question b

Effect of λ_r/λ_s on Decision Rule

- If $\lambda_r/\lambda_s = 0$, rejecting has no cost, so we always reject.
- As λ_r/λ_s increases, rejection becomes more costly, making classification more favorable.
- When $\lambda_r/\lambda_s \to 1$, the rejection threshold approaches $p(Y = j_{\max}|\mathbf{x}) \geq 0$, meaning we always classify and never reject.

Thus, increasing λ_r/λ_s reduces the likelihood of rejection, favoring classification.

Conclusion

The optimal decision rule is:

- Classify $\mathbf{x}$ as $j_{\max}$ if $p(Y = j_{\max}|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$.
- Otherwise, reject.

As the relative rejection cost λ_r/λ_s increases, rejection becomes less frequent, and we classify more often.

2.2 Question 5.2

2.2.1 Description

Newsvendor problem †

Consider the following classic problem in decision theory / economics. Suppose you are trying to decide how much quantity Q of some product (e.g., newspapers) to buy to maximize your profits. The optimal amount will depend on how much demand D you think there is for your product, as well as its cost to you C and its selling price P. Suppose D is unknown but has pdf $f(D)$ and cdf $F(D)$. We can evaluate the expected profit by considering two cases: if $D > Q$, then we sell all Q items, and make profit $\pi = (P - C)Q$; but if $D < Q$,

we only sell D items, at profit $(P - C)D$, but have wasted $C(Q - D)$ on the unsold items. So the expected profit if we buy quantity Q is

$$E\pi(Q) = \int_Q^\infty (P - C)Qf(D)dD + \int_0^Q (P - C)Df(D)dD - \int_0^Q C(Q - D)f(D)dD \tag{5.112}$$

Simplify this expression, and then take derivatives with respect to Q to show that the optimal quantity Q^* (which maximizes the expected profit) satisfies

$$F(Q^*) = \frac{P - C}{P} \tag{5.113}$$

2.2.2 Solution

2.2.2.1 Problem Setup

- **Demand:** Random variable D with PDF $f(D)$ and CDF $F(D)$.
- **Cost per unit:** C.
- **Selling price:** P.
- **Order quantity:** Q, chosen before demand is realized.

If $D > Q$, then all Q units sell at profit $(P - C)\,Q$. If $D < Q$, only D units sell (profit $(P - C)\,D$), and the remaining $Q - D$ units are unsold, losing cost $C\,(Q - D)$.

2.2.2.2 Expected Profit

The expected profit for ordering Q is

$$E\pi(Q) \;=\; \int_Q^\infty (P-C)\,Q\,f(D)\,dD \;+\; \int_0^Q (P-C)\,D\,f(D)\,dD \;-\; \int_0^Q C\,(Q-D)\,f(D)\,dD.$$

2.2.2.3 Simplification

We simplify term by term:

$$\int_Q^\infty (P-C)\,Q\,f(D)\,dD = (P-C)\,Q\,[1-F(Q)]. \tag{1}$$

$$\int_0^Q (P-C)\,D\,f(D)\,dD = (P-C)\int_0^Q D\,f(D)\,dD. \tag{2}$$

$$\int_0^Q C\,(Q-D)\,f(D)\,dD = C\,Q\int_0^Q f(D)\,dD - C\int_0^Q D\,f(D)\,dD = C\,Q\,F(Q) - C\int_0^Q D\,f(D)\,dD. \tag{3}$$

Putting these together:

$$\begin{aligned}
E\pi(Q) =\; & (P-C)\,Q\,[1-F(Q)] \;+\; (P-C)\int_0^Q D\,f(D)\,dD \\
& - \Big[C\,Q\,F(Q) \;-\; C\int_0^Q D\,f(D)\,dD\Big] \\
=\; & (P-C)\,Q\,[1-F(Q)] \;+\; (P-C)\int_0^Q D\,f(D)\,dD \;-\; C\,Q\,F(Q) \;+\; C\int_0^Q D\,f(D)\,dD \\
=\; & (P-C)\,Q\,[1-F(Q)] \;+\; \underbrace{\big[(P-C)+C\big]}_{=P}\int_0^Q D\,f(D)\,dD \;-\; C\,Q\,F(Q) \\
=\; & (P-C)\,Q\,[1-F(Q)] \;+\; P\int_0^Q D\,f(D)\,dD \;-\; C\,Q\,F(Q).
\end{aligned}$$

2.2.2.4 Maximizing the Expected Profit

We find the optimal order quantity Q^* by taking the derivative of $E\pi(Q)$ with respect to Q and setting it to zero:

$$\frac{d}{dQ}\Big((P-C)\,Q\,[1-F(Q)]\Big) = (P-C)\Big([1-F(Q)] - Q\,f(Q)\Big),$$

$$\frac{d}{dQ}\Big(P\int_0^Q D\,f(D)\,dD\Big) = P\,Q\,f(Q),$$

$$\frac{d}{dQ}\Big(-C\,Q\,F(Q)\Big) = -\,C\,\Big(F(Q)+Q\,f(Q)\Big).$$

Summing and equating to zero:

$$(P-C)\Big([1-F(Q)]-Q\,f(Q)\Big)\;+\;P\,Q\,f(Q)\;-\;C\Big(F(Q)+Q\,f(Q)\Big)=0.$$

Observe that all terms involving $Q\,f(Q)$ cancel out, leaving

$$(P-C)\,[1-F(Q)]-\,C\,F(Q)=0 \quad\Longrightarrow\quad (P-C)-P\,F(Q)=0.$$

Hence,

$$F(Q^*)=\frac{P-C}{P}.$$

2.2.2.5 Conclusion

The optimal order quantity Q^* is the inverse CDF evaluated at

$$\frac{P-C}{P},$$

often called the *critical fractile*. In symbols,

$$Q^* = F^{-1}\Big(\tfrac{P-C}{P}\Big).$$

2.3 Question 5.3

2.3.1 Description

Bayes factors and ROC curves †

Let $B = \frac{p(D|H_1)}{p(D|H_0)}$ be the Bayes factor in favor of model 1. Suppose we plot two ROC curves, one computed by thresholding B, and the other computed by thresholding $p(H_1|D)$. Will they be the same or different? Explain why.

2.3.2 Solution

The two ROC curves will be the same because both the Bayes factor B and the posterior probability $p(H_1|D)$ induce the same ranking of instances.

2.3.2.1 Explanation

Definition of the Bayes Factor B: The Bayes factor is given by

$$B = \frac{p(D|H_1)}{p(D|H_0)}$$

which measures how much more likely the data is under H_1 compared to H_0.

Posterior Probability of H_1: Using Bayes' theorem, we have

$$p(H_1|D) = \frac{p(D|H_1)p(H_1)}{p(D)}$$

where

$$p(D) = p(D|H_1)p(H_1) + p(D|H_0)p(H_0).$$

Expressing $p(H_1|D)$ **in Terms of** B: Using the definition of B, we rewrite $p(H_1|D)$ as

$$p(H_1|D) = \frac{Bp(H_1)}{Bp(H_1) + p(H_0)}.$$

Since this is a monotonic function of B, the ranking of instances remains unchanged.

ROC Curves Depend Only on Ranking: The ROC curve is computed by ranking instances according to a thresholded score and plotting the true positive rate (TPR) against the false positive rate (FPR). Since both B and $p(H_1|D)$ provide the same ranking, the ROC curves will be identical.

2.3.2.2 Conclusion

Although B and $p(H_1|D)$ differ in their absolute values, they induce the same ranking of instances. Since ROC curves depend solely on the ranking and not the specific values, the two ROC curves will be identical.

2.4 Question 5.4

2.4.1 Description

Posterior median is optimal estimate under L1 loss

Prove that the posterior median is the optimal estimate under L1 loss.

2.4.2 Solution

We start by expressing the posterior expected L1 loss for an estimator a:

$$\rho(a|x) = \int_{-\infty}^{a} (a - \theta)p(\theta|x)\, d\theta + \int_{a}^{\infty} (\theta - a)p(\theta|x)\, d\theta$$

To find the optimal a that minimizes this loss, we differentiate $\rho(a|x)$ with respect to a. Using the Leibniz rule for differentiating under the integral sign:

$$\frac{d}{da}\int_{a}^{\infty} (\theta - a)p(\theta|x)\, d\theta = \int_{a}^{\infty} (-p(\theta|x))\, d\theta = -\int_{a}^{\infty} p(\theta|x)\, d\theta$$

$$\frac{d}{da}\int_{-\infty}^{a} (a - \theta)p(\theta|x)\, d\theta = \int_{-\infty}^{a} p(\theta|x)\, d\theta$$

Thus, the derivative of the expected loss is:

$$\rho'(a|x) = \int_{-\infty}^{a} p(\theta|x)\, d\theta - \int_{a}^{\infty} p(\theta|x)\, d\theta$$

Setting the derivative to zero for minimization:

$$\int_{-\infty}^{a} p(\theta|x)\, d\theta = \int_{a}^{\infty} p(\theta|x)\, d\theta$$

This implies:

$$P(\theta \leq a|x) = P(\theta \geq a|x)$$

Therefore, the optimal a is the posterior median, where the cumulative distribution function $P(\theta \leq a|x) = 0.5$.