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 4: Bayesian Inference

Exercise 5.1

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:

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

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).

ANSWER

a.

We need to choose between rejecting the classification with risk λ_r or choosing the most probable class, $j_{\max} = \arg\max_j p(Y = j|\mathbf{x})$, 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:

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

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

Hence, we obtain the equation:

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

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

For completeness, we should prove that when we decide to choose a class (and not reject), we always pick the most probable one. Suppose we choose a non-maximal class $k \neq j_{\max}$. The misclassification risk in this case is:

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

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

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

which is always bigger than picking $j_{\max}$.
This should confirm that by selecting $j_{\max}$, it will minimize the risk of the chosen classification.

b.

Effect of λ_r/λ_s on Decision Rule

- If $\lambda_r/\lambda_s = 0$, there is no cost to rejecting, so we always reject.
- As λ_r/λ_s increases to 1, the cost of rejecting increase, making classification more favorable.

Thus, increasing λ_r/λ_s reduces the likelihood of rejection, favoring classification. We can see that $p(Y = j_{\max}|\mathbf{x}) \geq 1 - \frac{\lambda_r}{\lambda_s}$ is always satisfied, so we always accept the most probable class.

Exercise 5.2

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

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

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

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

ANSWER

The expected profit for buying quantity Q is

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

We simplify each term:

$$\int_Q^\infty (P - C)Qf(D)dD = (P - C)Q[1 - F(Q)].$$

$$\int_0^Q (P - C)Df(D)dD = (P - C)\int_0^Q Df(D)dD.$$

$$\int_0^Q C(Q - D)f(D)dD = CQ\int_0^Q f(D)dD - C\int_0^Q Df(D)dD = CQF(Q) - C\int_0^Q Df(D)dD.$$

Putting the equations (10), (11), (12) into equation (9):

$$E\pi(Q) = (P - C)Q[1 - F(Q)] + (P - C)\int_0^Q Df(D)dD$$

$$- \left[CQF(Q) - C\int_0^Q Df(D)dD\right]$$

$$= (P - C)Q[1 - F(Q)] + (P - C)\int_0^Q Df(D)dD - CQF(Q) + C\int_0^Q Df(D)dD$$

$$= (P-C)Q[1-F(Q)] + \underbrace{[(P-C)+C]}_{=P}\int_0^Q Df(D)dD - CQF(Q)$$

$$= (P-C)Q[1-F(Q)] + P\int_0^Q Df(D)dD - CQF(Q).$$

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:

$$\begin{aligned}\frac{d}{dQ}E\pi(Q) &= (P-C)\,[1-F(Q)] - (P-C)Qf(Q) + PQf(Q) - CF(Q) - CQF(Q)\\ &= (P-C) - PF(Q)\\ &= 0\end{aligned}$$

So,

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

Conclusion

The optimal order quantity Q^* (which can be called *critical fractile*) is the inverse CDF evaluated by

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

Exercise 5.3

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.

ANSWER

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.

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.

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, and therefore cannot affect the shape of the ROC curve. This makes sense intuitively, since, in the two class case, it should not matter whether we threshold the ratio $p(H_1|D)$ / $p(H_0|D)$, or the posterior, $p(H_1|D)$, since they contain the same information, just measured on different scales.

Exercise 5.4

Posterior median is optimal estimate under L1 loss

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

ANSWER

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 $P(\theta \leq a|x) = 0.5$.