Pattern Recognition

Lecture 3. Bayesian Decision Theory

Dr. Shanshan ZHAO

shanshan.zhao@xjtlu.edu.cn

School of AI and Advanced Computing

Xi’an Jiaotong-Liverpool University

Academic Year 2021-2022

Table of Contents

Bayesian Decision theory

Design classifiers to recommend decisions that minimize some total expected ”risk”.

- fundamental statistical approach to the problem of pattern classification
- ideal case, optimal classifier
- compare with all other classifiers

XJTLU

Fish example

- Reconsider the problem: Classify two fish as salmon or seabass.

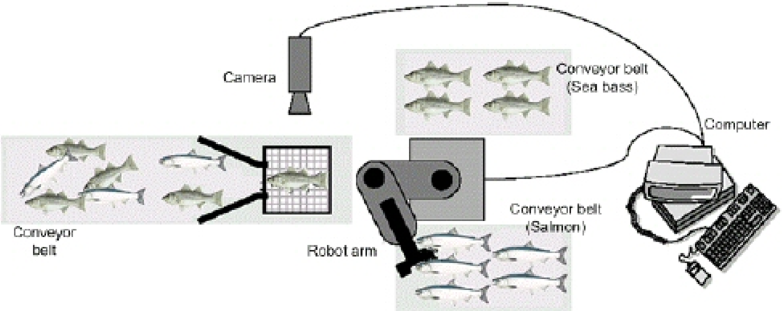


Figure: The fish packing system.

Fish example

- Assume any given fish is either a salmon or a sea bass.
- Let's define a variable ω that describes the *state of nature*

$$\begin{array}{ll} \omega = \omega_1 & \text{for sea bass} \\ \omega = \omega_2 & \text{for salmon} \end{array}$$

(a)

(b)

Figure: The objects to be classified: a. salmon; b. sea bass

Prior Probability

If sea bass is produced as much as salmon, we would say that the next fish is equally like to be sea bass or salmon.
More generally, we assume there is a **prior probability**.

- The a priori or prior probability reflects our knowledge of how likely we expect a certain category before we can actually observe.
- The priors must exhibit exclusivity and exhaustivity. For c states of nature, or classes:
  $\sum_{i=1}^{c} P(\omega_i) = 1$
  In the fish example, $P(\omega_1) + P(\omega_2) = 1$

Decision Rule with Only Priors

A **decision rule** prescribes what action to take based on observed input.
Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

- Favours the most likely class
- This rule will be making the same decision all times
  -i.e., optimum if no other information is available

What can we say about this decision rule?

Decision Rule with Only Priors

A **decision rule** prescribes what action to take based on observed input.
Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2

- Favours the most likely class
- This rule will be making the same decision all times
  -i.e., optimum if no other information is available

What can we say about this decision rule?
Seems reasonable, but it will always make the same choice. It doesn't make sense if we were to judge many fish.

XJTLU

Features and Feature Spaces

Mostly, we are not asked to make decisions with such little information. We might use some measurements or features to improve our classifier.

- A **feature** is an observable variable.
- A **feature space** is a set from which we can sample or observe values.
- Examples of features: Length, Width, Lightness, etc.
- For simplicity, we assume our features are all continuous values.
- Denote a scalar feature as x and a vector features as x. For a d-dimensional feature space, $\mathsf{x} \in \mathcal{R}^{d}$.

XJTLU

Conditional probability density

- The Conditional probability density $p(x|\omega_i)$is also called *likelihood*. It shows the probability density of feature x, given that it belongs to class ω_i.
- Example

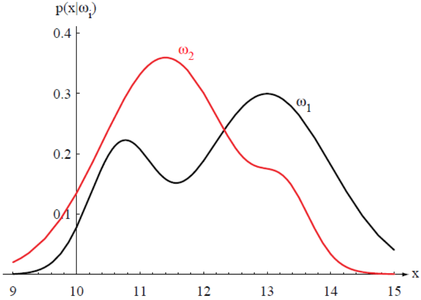


Figure: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

Posterior Probability

- Now that we know the prior distribution as well as the conditional density, how does this affect our decision rule?
- Posterior probability is the probability of a class given our observations: $P(\omega|x)$.
- Bayes Formula:

$$P(\omega, x) = P(\omega|x)p(x) = p(x|\omega)P(\omega) \tag{1}$$

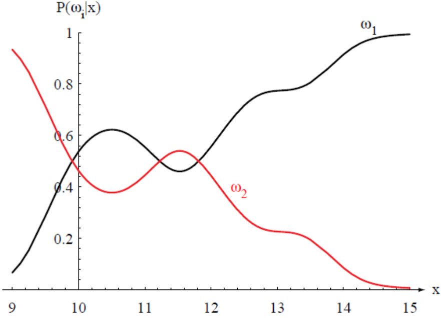
$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{p(x)} \tag{2}$$

$$= \frac{p(x|\omega)P(\omega)}{\sum_i p(x|\omega_i)P(\omega_i)} \tag{3}$$

- Notice that the likelihood and the prior govern the posterior. The $p(x)$ evidence term is a scale-factor to normalize the density.

Posterior Probability

For the case of $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$, the posterior is



For a given observation x, we would be inclined to let the posterior decide the decision:

$$\omega = arg \max_i P(\omega_i|x) \tag{4}$$

XJTLU

Decision Rule Using Posteriors

Recap: Using Bayes' rule, the posterior probability of category ω_j given measurement x is given by: Probabilities *Bayes rule*.

$$P(\omega_i|x) = \frac{p(x|\omega_i)P(\omega_i)}{p(x)}$$

The *Bayes classification rule* can be stated as

- Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; or
- Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$;
  Decision making relies on both the priors and the likelihoods and Bayesian Decision Rule combines them to achieve the minimum probability of error.

Error Probability

For the two class situation, we have

$$P(error|x) = \begin{cases} P(\omega_1|x) & \text{if we decide } \omega_2 \\ P(\omega_2|x) & \text{if we decide } \omega_1 \end{cases} \tag{5}$$

We can minimize the probability of error by :

$$\text{Decide} \quad \omega_1 \quad \text{if} \quad P(\omega_1|x) > P(\omega_2|x); \quad \text{otherwise decide} \quad \omega_2 \tag{6}$$

$$P(error|x) = \min[P(\omega_1|x), P(\omega_2|x)] \tag{7}$$

And , this minimizes the average probability of error too:

$$P(error) = \int_{-\infty}^{\infty} P(error|x)p(x)dx \tag{8}$$

Because the integral will be minimized when we can ensure each $P(error|x)$is as small as possible.

XJTLU

Minimizing the misclassification rate

Goal: To make as a few misclassifications as possible. A mistake occurs when an input vector belongs to class ω_1 is assigned to class ω_2 or vice versa.

$$\begin{aligned} P(error) &= P(x \in \mathcal{R}_2, \omega_1) + P(x \in \mathcal{R}_1, \omega_2) \\ &= \int_{\mathcal{R}_2} p(x, \omega_1) dx + \int_{\mathcal{R}_1} p(x, \omega_2) dx \\ &= \int_{\mathcal{R}_2} p(x|\omega_1) P(\omega_1) dx + \int_{\mathcal{R}_1} p(x|\omega_2) P(\omega_2) dx \end{aligned}$$

Question: It is true that the probability of misclassification is minimised by assigning each point to the class with the maximum posterior probability?

Minimizing the misclassification rate

Example 1. The two regions R_1 and R_2 formed by the Bayesian classifier for the case of two equiprobable classes.

$$P_e = \frac{1}{2}\int_{-\infty}^{x_0} p(x|\omega_2)dx + \frac{1}{2}\int_{x_0}^{+\infty} p(x|\omega_1)dx \tag{9}$$

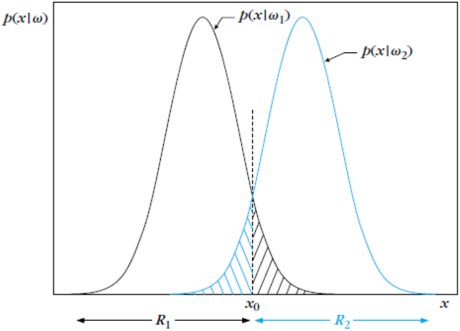


Figure: P_e is equal to the total shaded area under the curves

Minimizing the misclassification rate

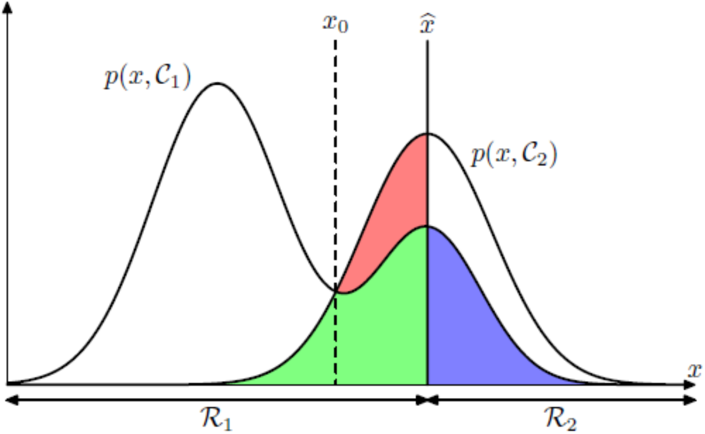


Figure: Schematic illustration of the joint probabilities $p(x, C_k)$ for each of two classes plotted against x, together with the decision boundary$x = \hat{x}$. (from Bishop's PRML, page 40)

Minimizing the misclassification rate

It is possible to extend this justification for a decision rule based on **maximum posterior probability**.

Therefore, we consider the probability for a pattern being correctly classified $P(correct)$.

Exercises:

- 1. $P(correct) = ?$
- 2. Prove that the maximum posterior probability decision rule is equivalent to minimising the probability of misclassification.

Minimizing the Average Risk(Expected loss)

The classification error probability is not always the best criterion to be adopted for minimization. $\Longrightarrow$ assign a penalty term to weigh each error.

For the sake of generality, this *risk* or *loss* associated with $\omega_k (k = 1, 2, ..., M)$ is defined as

$$R_k = \sum_{i=1}^{M} \lambda_{ki} \int_{\mathcal{R}_k} p(x, \omega_i) dx \tag{10}$$

where, λ_{ki} is the loss incurred when a value of x, whose true class is ω_i, but we assign it to class ω_k. $\mathcal{R}_1$ denotes that region in feature space where the classifier decides ω_1 and likewise for $\mathcal{R}_2$ and ω_2, etc.

The risk is (**Note that we use r_k to denote the term within the integral**):

$$R = \sum_{k=1}^{M} R_k = \sum_{k=1}^{M} \int_{\mathcal{R}_k} \left(\sum_{i=1}^{M} \lambda_{ki} p(x, \omega_i) \right) dx = \sum_{k=1}^{M} \int_{\mathcal{R}_k} r_k dx$$

Minimizing the risk is achieved if each of the integrals is minimized, which is equivalent to selecting partitioning regions so that

$x \in \mathcal{R}_a \quad$ if $\quad r_a \equiv \sum_{i=1}^{M} \lambda_{ai} p(x, \omega_i) \quad < \quad r_b \equiv \sum_{i=1}^{M} \lambda_{bi} p(x, \omega_i) \quad \forall b \neq a$

Minimizing the Average Risk(Expected loss)

The two-class case

$$
\begin{aligned}
r_1 &= \lambda_{11} p(x, \omega_1) + \lambda_{12} p(x, \omega_2) \\
r_2 &= \lambda_{21} p(x, \omega_1) + \lambda_{22} p(x, \omega_2)
\end{aligned} \tag{11}
$$

We assign x to ω_1 if $r_1 < r_2$, that is,

$$
\begin{aligned}
(\lambda_{12} - \lambda_{22}) p(x, \omega_2) &< (\lambda_{21} - \lambda_{11}) p(x, \omega_1) \\
(\lambda_{12} - \lambda_{22}) p(x|\omega_2) P(\omega_2) &< (\lambda_{21} - \lambda_{11}) p(x|\omega_1) P(\omega_1)
\end{aligned} \tag{12}
$$

The decision rule now becomes,

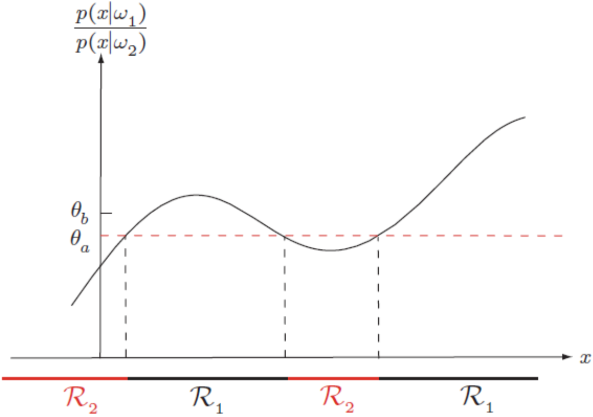
$$
x \in \omega_1 \quad \textit{if} \quad r_{12} \equiv \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \tag{13}
$$

(It is natural to assume that $\lambda_{ij} > \lambda_{ii}$.)

Minimizing the Average Risk(Expected loss)

The decision rule now becomes,

$$x \in \omega_1 \quad \textit{if} \quad r_{12} \equiv \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)} \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \tag{14}$$



Exercise

In a two-class problem with a single feature x the pdfs are Gaussians with variance $\sigma^2 = 1/2$ for both classes and mean values 0 and 1, respectively, that is,

$$\begin{aligned} p(x|\omega_1) &= \frac{1}{\sqrt{\pi}} exp(-x^2) \\ p(x|\omega_2) &= \frac{1}{\sqrt{\pi}} exp(-(x-1)^2) \end{aligned} \tag{15}$$

If $P(\omega_1) = P(\omega_2) = 1/2$, compute the threshold value x_0

- (i) for minimum error probability
- (ii) for minimum risk if the loss matrix is

$$L = \begin{bmatrix} 0 & 0.5 \\ 1 & 0 \end{bmatrix} \tag{16}$$

Thank You !

$\mathcal{Q}$ & $\mathcal{A}$

XJTLU