Chapter 2:
Bayesian Decision Theory (Part 1)

Introduction:

- Bayesian decision theory is a fundamental statistical approach to the problem of pattern classification. This approach is based on quantifying the tradeoffs between various classification decisions using probability and the costs that accompany such decisions.

- The sea bass/salmon example

  - State of nature, prior

    - State of nature is a random variable
    - The catch of salmon and sea bass is equiprobable

      $\omega = \omega_1$ for see bass and $\omega = \omega_2$ for salmon

      $P(\omega_1)$ *a priori probability* that the next fish is sea bass

      $P(\omega_1) = P(\omega_2)$ (uniform priors)

      $P(\omega_1) + P(\omega_2) = 1$ (exclusivity and exhaustivity)

- **Decision rule with only the prior information**
  - **Decide ω_1 if $P(\omega_1) > P(\omega_2)$ otherwise decide ω_2**
- In most circumstances we are not asked to make decisions with so little information.
  - We might for instance use a lightness measurement x to improve our classifier.

- **Use of the class – conditional information**

- The probability density function $p(x|\omega_1)$ should be written as $p_X(x|\omega_1)$ to indicate that we are speaking about a particular density function for the random variable X.

- $p(x|\omega_1)$ and $p(x|\omega_2)$ describe the difference in lightness between populations of sea and salmon

We generally use an upper-case $P(\cdot)$ to denote a *probability mass function* and a lower-case $p(\cdot)$ to denote a *probability density function*.

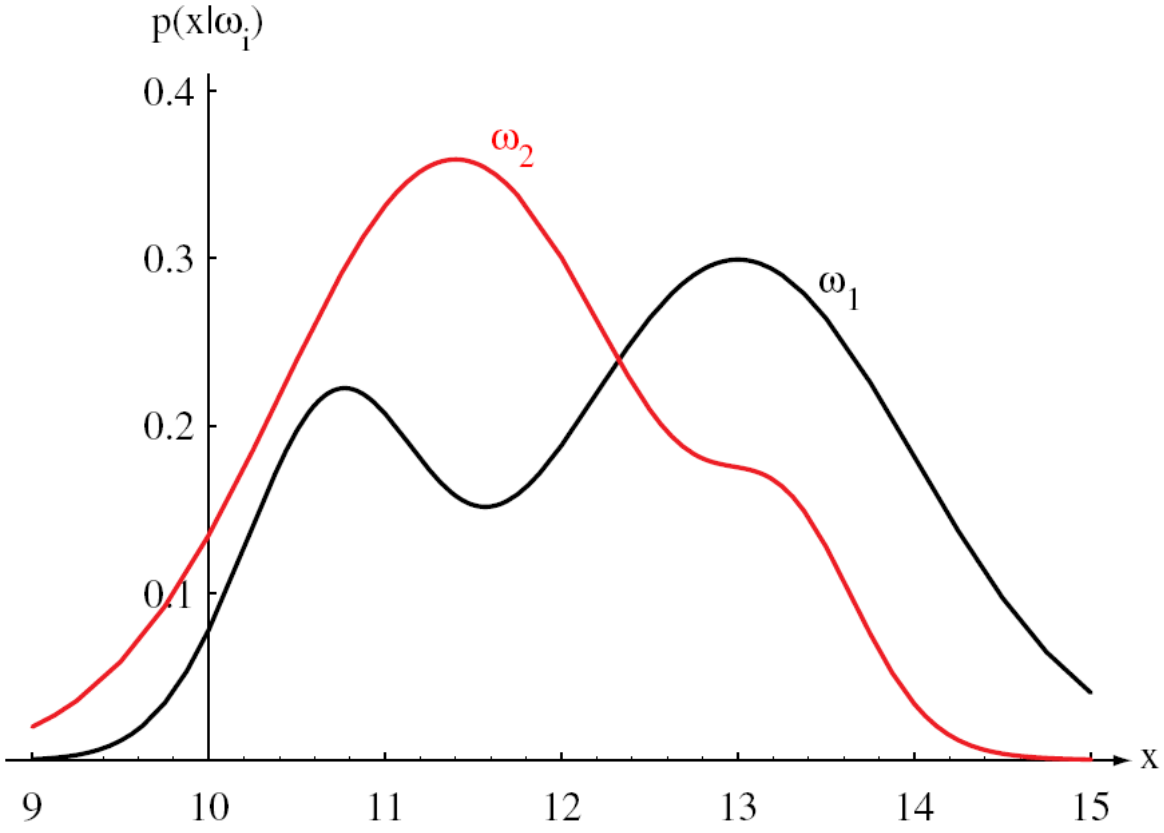


Figure 2.1: Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value *x* given the pattern is in category ω_i. If *x* represents the length of a fish, the two curves might describe the difference in length of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0.

- **Posterior, likelihood, evidence**
- Suppose that we know both the prior probabilities $P(\omega_j)$ and the conditional densities $p(x|\omega_j)$. Suppose further that we measure the lightness of a fish and discover that its value is x. How does this measurement influence our attitude concerning the true state of nature — that is, the category of the fish?
- The (joint) probability density of finding a pattern that is in category ω_j and has feature value x can be written two ways:
- $p(\omega_j, x) = P(\omega_j|x)p(x) = p(x|\omega_j)P(\omega_j)$

Bayes' formula

$$\boxed{P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},}$$

$$\text{posterior} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

Where in case of two categories

$$p(x) = \sum_{j=1}^{2} p(x|\omega_j)P(\omega_j).$$

Notice that in Bayes' formula the product of the likelihood and the prior probability that is most important in determining the posterior probability; the evidence factor, $p(x)$, can be viewed as merely a scale factor that guarantees that the posterior probabilities sum to one

If we have an observation x for which $P(\omega_1|x)$ is greater than $P(\omega_2|x)$, we would naturally be inclined to decide that the true state of nature is ω_1.

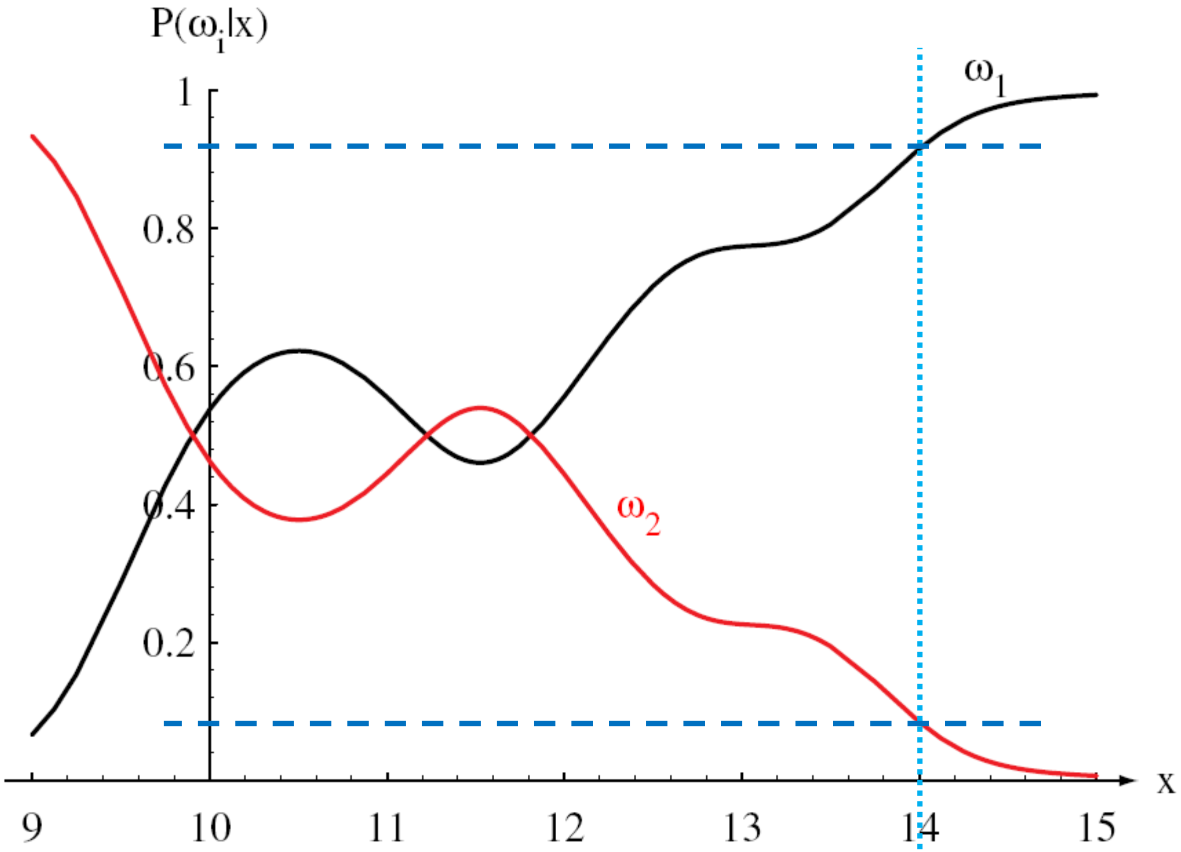


Figure 2.2: Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) =1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0

- **Decision given the posterior probabilities**

x is an observation for which:

if $P(\omega_1 \mid x) > P(\omega_2 \mid x) \Rightarrow$ True state of nature = ω_1

if $P(\omega_1 \mid x) < P(\omega_2 \mid x) \Rightarrow$ True state of nature = ω_2

Therefore:

whenever we observe a particular x, the probability of error is:

$$P(error \mid x) = P(\omega_1 \mid x) \text{ if we decide } \omega_2$$

$$P(error \mid x) = P(\omega_2 \mid x) \text{ if we decide } \omega_1$$

- **Minimizing the probability of error**

Decide ω_1 if $P(\omega_1 \mid x) > P(\omega_2 \mid x)$; otherwise decide ω_2

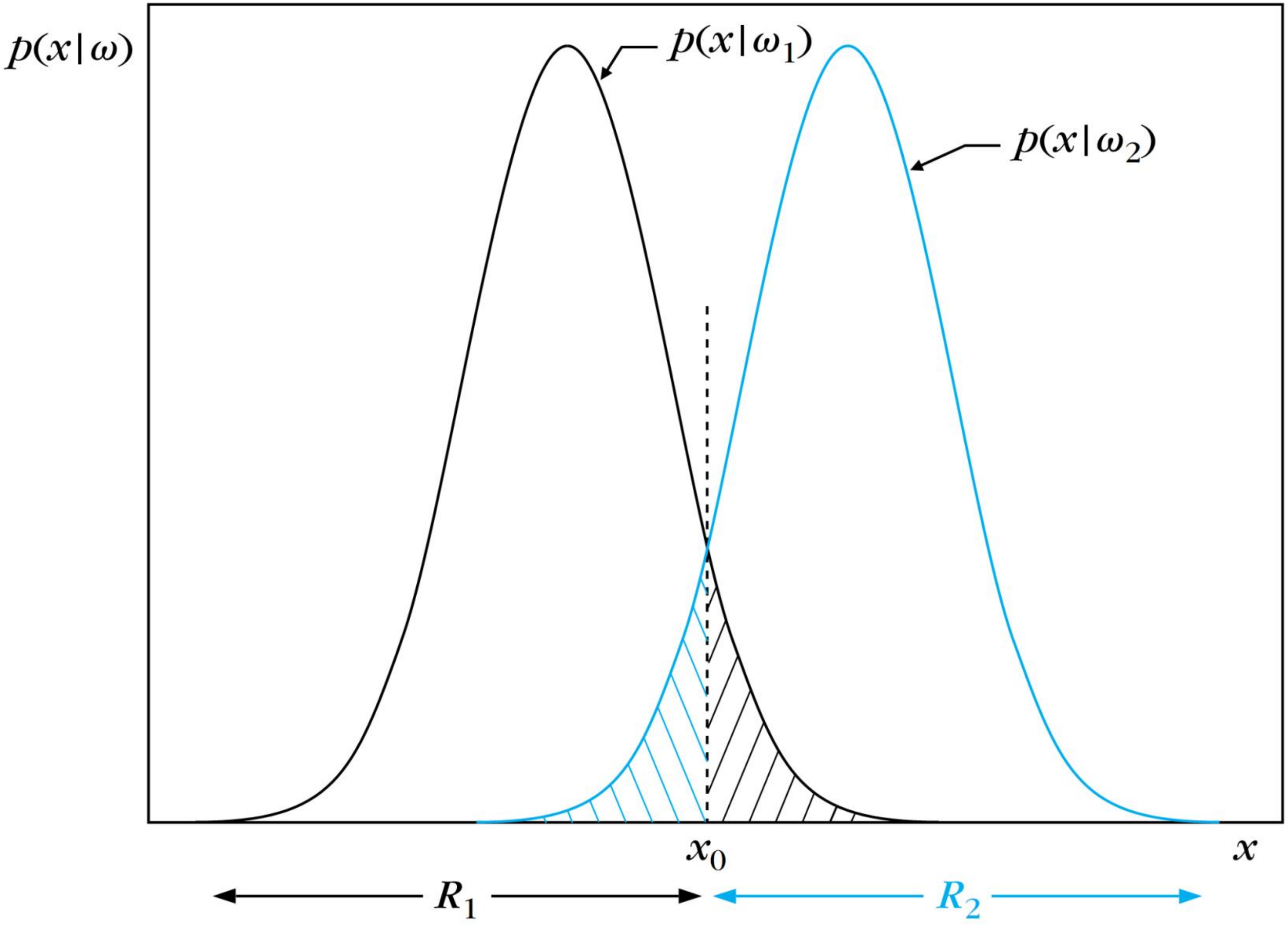
$$\boxed{P(error) = \int_{-\infty}^{\infty} p(error, x)dx = \int_{-\infty}^{\infty} P(error \mid x)\, p(x)dx}$$

If for every x we insure that $P(error|x)$ is as small as possible, then the integral must be as small as possible.

Therefore:

$$P(error|x) = min\ [P(\omega_1|x),\ P(\omega_2|x)]$$

(Bayes decision)



Example of the two regions *R₁* and *R₂* formed by the Bayesian classifier for the case of two equiprobable classes.

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

- By eliminating this scale factor, $p(x)$, we obtain the following completely equivalent decision rule:
- Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

- Note using evidence $p(x)$ insure us that $P(\omega_1|x) + P(\omega_2|x) = 1$.

Bayesian Decision Theory – Continuous Features

- **Generalization of the preceding ideas**
  - Use of more than one feature
  - Use more than two states of nature
  - Allowing actions and not only decide on the state of nature
  - Introduce a loss function which is more general than the probability of error

- The use of more than one feature → the *feature vector* **x**, where **x** is in a d-dimensional Euclidean space $\mathbf{R}^d$, called the *feature space*.
- Allowing more feature than two states of nature provides us with a useful generalization for a small notational space expense.
- Allowing actions other than classification primarily allows the possibility of rejection, i.e., of refusing to make a decision in close cases; this is a useful option if being indecisive is not too costly.

$$R = \left\{ \boldsymbol{x} \middle| 1 - \max_i p(\omega_i|\boldsymbol{x}) > t \right\}$$

R, a *reject region*

$$A = \left\{ \boldsymbol{x} \middle| 1 - \max_i p(\omega_i|\boldsymbol{x}) \leq t \right\}$$

A, an *acceptance* or *classification region*

where *t* is a threshold.

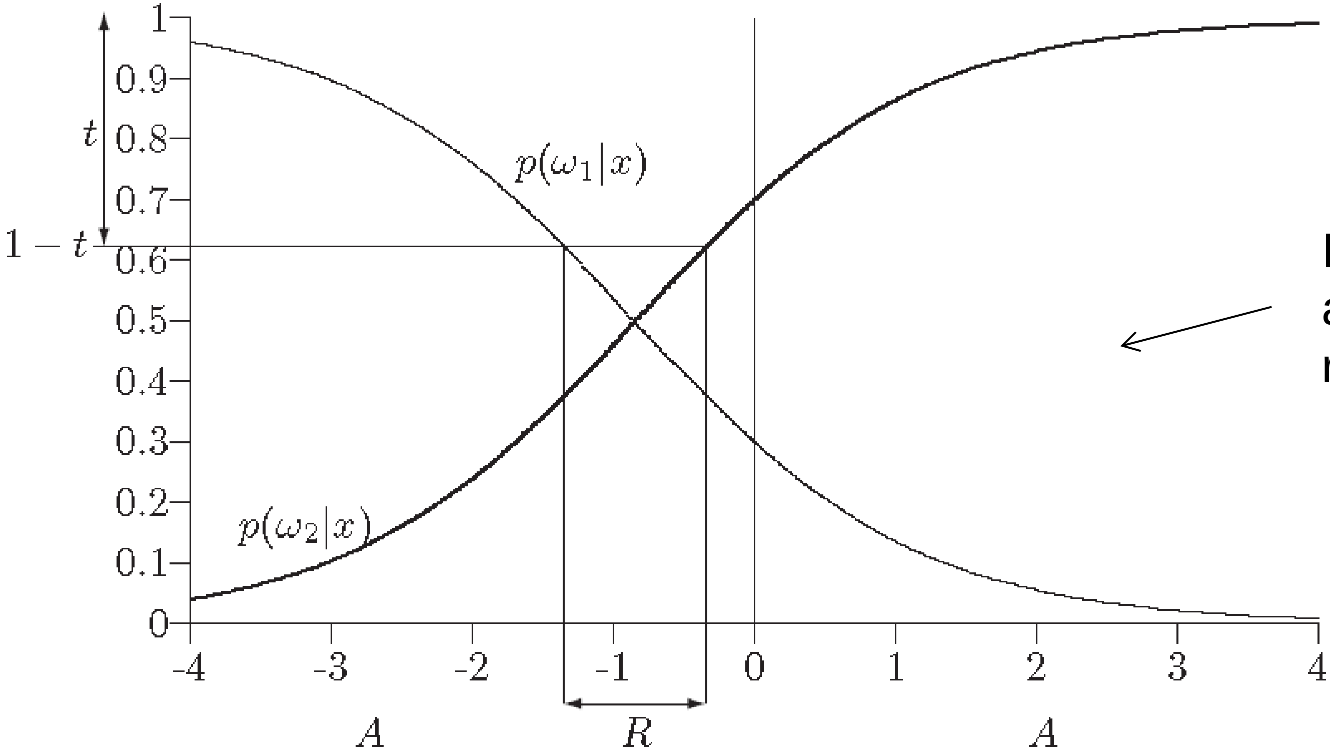


Illustration of acceptance and reject regions.

Formally, the *loss function* states exactly how costly loss each action is, and is used to convert a probability determination into a decision.

Let $\{\omega_1, \omega_2, \ldots, \omega_c\}$ be the set of c states of nature ("categories")

Let $\{\alpha_1, \alpha_2, \ldots, \alpha_a\}$ be the set of possible actions

Let $\lambda(\alpha_i|\omega_j)$ be the loss incurred for taking action α_i when the state of nature is ω_j

Bayes' formula:

$$P(\omega_j|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_j)P(\omega_j)}{p(\mathbf{x})},$$

where the evidence is now

$$p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x}|\omega_j)P(\omega_j).$$

the expected loss associated with taking action α_i is merely

$$R(\alpha_i \,|\, \mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i \,|\, \omega_j)P(\omega_j \,|\, \mathbf{x})$$

An expected loss is called a *risk*, and $R(\alpha_i|\mathbf{x})$ is called the *conditional risk*.

We shall show that this *Bayes decision procedure* actually provides the optimal performance on an overall risk.

A general *decision rule* is a function α(**x**) that tells us which rule action to take for every possible observation. For every **x** the *decision function* α(**x**) assumes one of the *a* values α_1 , ..., α_a.

The overall risk is given by $R = \int R(\alpha(\mathbf{x})|\mathbf{x})p(\mathbf{x})\ d\mathbf{x},$

Overall risk

R = Sum of all $\underbrace{R(\alpha_i \mid \mathbf{x})}_{\textbf{Conditional risk}}$ for $i = 1,\ldots,a$

Minimizing R ⟺ Minimizing $R(\alpha_i|\mathbf{x})$ for $i = 1,\ldots, a$

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{j=c} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid \mathbf{x})$$

for $i = 1,\ldots,a$

Selecting the action α_i for which $R(\alpha_i|\mathbf{x})$ is minimum. The resulting minimum overall risk is called the *Bayes risk*, denoted R^*, and is the best performance that can be achieved.

- **Two-category classification**

α_1 : deciding ω_1

α_2 : deciding ω_2

$\lambda_{ij} = \lambda(\alpha_i|\omega_j)$ be loss incurred for deciding ω_i when the true state of nature is ω_j

Conditional risk:

$$R(\alpha_1 \mid x) = \lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x)$$

$$R(\alpha_2 \mid x) = \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x)$$

Our rule is the following:

if $R(\alpha_1 \mid x) < R(\alpha_2 \mid x)$

action α_1: “decide ω_1” is taken

This results in the equivalent rule:

decide ω_1 if: $(\lambda_{21}-\lambda_{11})\, P(\omega_1|\mathbf{x}) > (\lambda_{12}-\lambda_{22}) P(\omega_2|\mathbf{x})$

Or

$(\lambda_{21}-\lambda_{11})\, p(\mathbf{x}|\omega_1)\, P(\omega_1) > (\lambda_{12}-\lambda_{22})\, p(\mathbf{x}|\omega_2)\, P(\omega_2)$

and decide ω_2 otherwise

Likelihood ratio:

The preceding rule is equivalent to the following rule: if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)}$$

Then take action α_1 (decide ω_1)

Otherwise take action α_2 (decide ω_2)

- We can consider $p(\mathbf{x}|\omega_j)$ a function of ω_j (i.e., the likelihood function), and then form the *likelihood ratio* $p(\mathbf{x}|\omega_1)/p(\mathbf{x}|\omega_2)$.

Optimal decision property: "If the likelihood ratio exceeds a threshold value independent of the input pattern **x**, we can take optimal actions"

Exercise

Select the optimal decision where:

$\Omega = \{\omega_1, \omega_2\}$

$p(x|\omega_1) \longrightarrow N(2, 0.5)$ (Normal distribution)

$p(x|\omega_2) \longrightarrow N(1.5, 0.2)$

$P(\omega_1) = 2/3$

$P(\omega_2) = 1/3$

$$\boldsymbol{\lambda} = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$