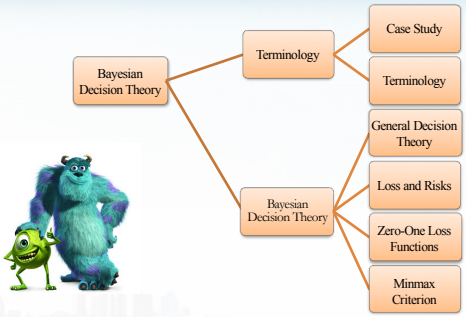
Lecture Notes on

Pattern Recognition

Session 02(A): Bayesian Decision Theory (I)

Gang Li
School of Information Technology
Deakin University, VIC 3125, Australia

Road map



Terminology

- A Case Study
- Terminology

Dormitory Guard (A Pattern Classification System)

Dormitory Guard (A Pattern Classification System)

Dormitory Guard (A Pattern Classification System)

- If the Dormitory Guard is presbyopia "blind", how will she make decision?

Terminology

- **Random variable**: state of nature ω
  - e.g., ω_1 for girl, ω_2 for boy
- **Priors**: probabilities $P(\omega_1)$ and $P(\omega_2)$
  - e.g., prior knowledge of how likely is to meet a girl or a boy nearby

Decision Rule Using Priors Only

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

$$P(error) = \begin{cases} P(\omega_1) \text{ if we decide } \omega_2 \\ P(\omega_2) \text{ if we decide } \omega_1 \end{cases} = \min(p(\omega_1), p(\omega_2))$$

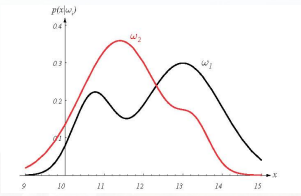
- This rule makes the same decision all times!
  - Unless the prior probability changes
- Favours the most likely class
  - optimum if no other info is available

Decision Rule Upon Evidence

- **Evidence**: probability density function p(x)
  - how frequently we will measure a pattern with feature value x (e.g., x is ***length of the hair***)

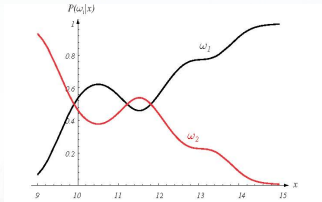
Decision Rule Upon Evidence

- **Likelihood**: conditional probability density $p(x|\omega_i)$
  - how frequently we will measure a pattern with feature value x given that the pattern belongs to class ω_i



Decision Rule Upon Evidence

- **Posterior**: conditional probability density $P(\omega_i|x)$
  - the probability that the person belongs to class ω_i given measurement x.



$$P(\omega_1) = \frac{2}{3}$$

$$P(\omega_2) = \frac{1}{3}$$

Decision Rule Upon Evidence

- **Bayes Rule**:
  - the posterior probability of category ω_i given measurement x is given by

$$P(\omega_j / x) = \frac{p(x/\omega_j)P(\omega_j)}{p(x)} = \frac{likelihood \times prior}{evidence}$$

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

- **Decide** ω_1 if $P(\omega_1/x) > P(\omega_2/x)$; otherwise **decide** ω_2
- **Decide** ω_1 if $p(x/\omega_1)P(\omega_1) > p(x/\omega_2)P(\omega_2)$ otherwise **decide** ω_2

Decision Rule Upon Evidence

- In this case, the probability of error is:

$$P(error/x)=\begin{cases}P(\omega_1/x) & if\ we\ decide\ \omega_2\\ P(\omega_2/x) & if\ we\ decide\ \omega_1\end{cases}$$

- What is the average probability error?

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

- **Bayes rule is optimum**
  - it minimizes the average probability error since

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

Decision Rule Upon Evidence

$$P(\omega_j/x)=\frac{p(x/\omega_j)P(\omega_j)}{p(x)}=\frac{likelihood\ \times\ prior}{evidence}$$

- **Bayes Rule Special Case (1)**: ***Uniform priors***
  - $P(\omega_1)=P(\omega_2)=\ldots=P(\omega_c)=1/c$
  - Then the decision only depends on the ***likelihood***

- **Decide** ω_1 if $p(x/\omega_1)>p(x/\omega_2)$ otherwise **decide** ω_2 (binary)

Decision Rule Upon Evidence

$$P(\omega_j/x)=\frac{p(x/\omega_j)P(\omega_j)}{p(x)}=\frac{likelihood\ \times\ prior}{evidence}$$

- **Bayes Rule Special Case (2)**: ***Uniform likelihood***
  - $P(x|\omega_1)=\ldots=P(x|\omega_c)$
  - Then the decision only depends on the ***priors***

- **Decide** ω_1 if $p(\omega_1)>p(\omega_2)$ otherwise **decide** ω_2 (binary)

Where do Probabilities Come From?

- In order to use Bayes Rule, it is necessary to know
  - **Priors**: $P(\omega_i)$
  - **Likelihood**: $P(x|\omega_i)$
- Two competitive answers to this question:
  - ***Relative frequency*** (objective) approach.
    - Probabilities can only come from experiments.
  - ***Bayesian*** (subjective) approach.
    - Probabilities may reflect degree of belief and can be based on opinion as well as experiments.
- Density Estimation (covered in later sessions)

Dormitory Guard

(A Pattern Classification System)

- Suppose in a dorm area, the priors of girls is 2/3, the priors of boys is 1/3.
  - If the observation is on "***with long hair or not***", what is the chance that a particular long hair person is a "girl"?



Dormitory Guard

(A Pattern Classification System)

- Suppose in a dorm area, the priors of girls is 2/3, the priors of boys is 1/3.
  - If the observation is on "***wearing glasses or not***", what is the chance that a glasses-wearing person is a "girl"?



Dormitory Guard

(A Pattern Classification System)

- Suppose in a dorm area, the priors of girls is 2/3, the priors of boys is 1/3.
  - If the observation is on "***wearing skirts or not***", what is the chance that a glasses-wearing person is a "girl"?

Dormitory Guard

(A Pattern Classification System)

- Now suppose the mistakes/errors in decisions exist, but with different cost
  - If letting a girl enter a boy's dorm building, then the guard will be ***fined*** for ***2 days' salary***
  - If letting a boy enter a girl's dorm building, then the guard will be ***fined*** for ***4 days' salary***
  - If rejecting a boy enter a boy's dorm, or rejecting a girl enter the girl's dorm, then the guard will be ***fined*** for ***1 day's salary***

Bayesian Decision Theory

- General Decision Theory
- Loss and Risks
- Zero-One Loss Functions
- MinMax Criterion

Bayesian Decision Theory

- Fundamental statistical approach that quantifies the trade-offs between various classification decisions using probabilities and the costs associated with such decisions.
  - Design classifiers to recommend actions that ***minimize some total expected "risk".***
  - The simplest risk is the classification error
    - i.e., all costs are equal

General Decision Theory

- Use more than one features.
- Allow more than two categories.
- Allow actions other than classifying the input to one of the possible categories (e.g., rejection).
- Employ a more general error function which penalizes some errors more than others.
  - **loss function**
    - i.e., associate "costs" with actions

General Decision Theory

- Features form a vector
- A finite set of c categories $\omega_1, \omega_2, \ldots, \omega_c$
- Bayes rule (i.e., using vector notation):

$$P(\omega_j / \mathbf{x}) = \frac{p(\mathbf{x} / \omega_j) P(\omega_j)}{p(\mathbf{x})} \quad \text{where} \quad p(\mathbf{x}) = \sum_{j=1}^{c} p(\mathbf{x} / \omega_j) P(\omega_j)$$

- A finite set of l actions $\alpha_1, \alpha_2, \ldots, \alpha_l$
- A **loss function** $\lambda(\alpha_i \mid \omega_j)$
  - the loss or cost incurred for taking action α_i when the classification category is ω_j

Expected Loss (or Conditional Risk)

- Suppose we observe x and take action α_i
- If the true classification category is ω_j, we incur the loss $\lambda(\alpha_i \mid \omega_j)$
- The expected loss (or conditional risk) with taking action α_i :

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

Overall Risk

- Overall risk R

$$R = \int R(a(\mathbf{x}) / \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

where $\alpha(x)$ is a general decision rule that determines which action $\alpha_1, \alpha_2, \ldots, \alpha_l$ to take for every x.

- To minimize R, we need a decision rule $\alpha(x)$ that chooses the action with the minimum conditional risk $R(a_i \mid x)$ for every x.

Bayes Risk

- ***Bayes decision rule*** minimizes R by:
  - Computing $R(\alpha_i \mid x)$ for every α_i given an input x
  - Choosing the action α_i with the minimum $R(\alpha_i \mid x)$
- The resulting minimum overall risk is called ***Bayes risk*** and is the best (i.e., optimum) performance that can be achieved.

$$R^* = \min R$$

Two-category classification

- Define
  - α_1: decide ω_1
  - α_2: decide ω_2
  - $\lambda_{ij} = \lambda(\alpha_i \mid \omega_j)$
- The conditional risks are:

$$R(a_i / \mathbf{x}) = \sum_{j=1}^{c} \lambda(a_i / \omega_j) P(\omega_j / \mathbf{x}) \qquad \text{(c=2)}$$

$$R(a_1/\mathbf{x}) = \lambda_{11} P(\omega_1/\mathbf{x}) + \lambda_{12} P(\omega_2/\mathbf{x})$$
$$R(a_2/\mathbf{x}) = \lambda_{21} P(\omega_1/\mathbf{x}) + \lambda_{22} P(\omega_2/\mathbf{x})$$

Two-category classification

- Minimum risk decision rule:

Decide ω_1 if $R(a_1/\mathbf{x}) < R(a_2/\mathbf{x})$; otherwise decide ω_2

Decide ω_1 if $(\lambda_{21} - \lambda_{11}) P(\omega_1/\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(\omega_2/\mathbf{x})$; otherwise decide ω_2

  - In general, the cost for error action (decision) is larger than the cost for right action (decision)

$$\lambda_{21} - \lambda_{11} > 0 \qquad \lambda_{12} - \lambda_{22} > 0$$

Zero-One Loss Function

- Assign the same loss to all errors:

$$\lambda(\alpha_i, \omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, \ldots, c$$

- The conditional risk corresponding to this loss function:

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

"The risk corresponding to the 0-1 loss function is the average probability of error"

Zero-One Loss Function

- The overall risk in this case is the average probability error! Minimizing the risk requires maximizing the posterior probability $P(\omega_i|x)$

$$R(\alpha_i|x) = 1 - P(\omega_i|x)$$

- For *minimum* error rate
  - Decide ω_i, if $P(\omega_i|x) > P(\omega_j|x)$, $\forall j \neq i$

Decision Boundaries and Decision Regions

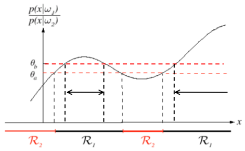
- **Likelihood Ratio**

$$Let \quad \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\cdot\frac{P(\omega_2)}{P(\omega_1)} = \theta_\lambda \text{ then decide } \omega_1 \text{ if: } \frac{P(x|\omega_1)}{P(x|\omega_2)} > \theta_\lambda$$

- If λ is the 0-1 loss function then the threshold involves only the priors:

$$if\ \lambda = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix},\ then\ \theta_\lambda = \frac{P(\omega_2)}{P(\omega_1)} = \theta_a$$

$$if\ \lambda = \begin{pmatrix} 0 & 2 \\ 1 & 0 \end{pmatrix},\ then\ \theta_\lambda = \frac{2P(\omega_2)}{P(\omega_1)} = \theta_b$$



Dormitory Guard

(A Pattern Classification System)

- Now suppose the mistakes/errors in decisions exist, but with different cost
  - If letting a girl enter a boy's dorm building, then the guard will be ***fined*** for ***2 days' salary***
  - If letting a boy enter a girl's dorm building, then the guard will be ***fined*** for ***4 days' salary***
  - If rejecting a boy enter a boy's dorm, or rejecting a girl enter the girl's dorm, then the guard will be ***fined*** for ***1 day's salary***

How does this affect the guard's decision?

Minmax criterion

- When prior probabilities are not known exactly, we need to design classifiers that perform well over a range of prior probabilities.

$$R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x}$$

- Consider the case of two categories:

$$\begin{aligned} R = &\int_{\mathcal{R}_1} [\lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x})] p(\mathbf{x})\, d\mathbf{x} + \\ &\int_{\mathcal{R}_2} [\lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x})] p(\mathbf{x})\, d\mathbf{x} \\ = &\int_{\mathcal{R}_1} [\lambda_{11} p(\mathbf{x}|\omega_1) P(\omega_1) + \lambda_{12} p(\mathbf{x}|\omega_2) P(\omega_2)]\, d\mathbf{x} + \\ &\int_{\mathcal{R}_2} [\lambda_{21} p(\mathbf{x}|\omega_1) P(\omega_1) + \lambda_{22} p(\mathbf{x}|\omega_2) P(\omega_2)]\, d\mathbf{x} \end{aligned}$$

Minmax criterion

- Since $P(\omega_1) + P(\omega_2) = 1$ and $\int_{\mathcal{R}_1} p(\mathbf{x}|\omega_1)\, d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, d\mathbf{x} = 1$
- We rewrite the overall risk as:

$$R(P(\omega_1)) = \lambda_{22} + (\lambda_{12}-\lambda_{22}) \int_{R_1} p(\mathbf{x}/\omega_2) d\mathbf{x} +$$

$$P(\omega_1)[(\lambda_{11}-\lambda_{22}) - (\lambda_{21}-\lambda_{11}) \int_{R_2} p(\mathbf{x}/\omega_1) d\mathbf{x} - (\lambda_{12}-\lambda_{22}) \int_{R_2} p(\mathbf{x}/\omega_2) d\mathbf{x}]$$

  - Assuming that the decision boundary is fixed, the overall risk is a linear function of $P(\omega_1)$

Minmax criterion

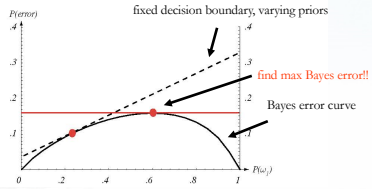
- Need to find a decision boundary that makes the coefficient of $P(\omega_1)$ zero
  - overall risk would be independent of priors!
- Minmax solution:

$$R_{mm} = \lambda_{22} + (\lambda_{12}-\lambda_{22}) \int_{\mathcal{R}_1} p(\mathbf{x}|\omega_2)\, d\mathbf{x}$$

$$R_{mm} = \lambda_{11} + (\lambda_{21}-\lambda_{11}) \int_{\mathcal{R}_2} p(\mathbf{x}|\omega_1)\, d\mathbf{x}$$

Minmax criterion

- How to find the minimax solution?
  - Search for the **prior** that **maximizes the Bayes risk**
- R_{mn} becomes equal to the worst Bayes risk



Here, we should design our decision boundary for $P(\omega_1)=0.6$

Seminar S02

- **Topic**
  - Try to solve a daily life problem using Bayesian Decision Theory, e.g.:
    - How the dorm guard will make decision on different features?
    - Will I get an A in this unit, if I speak 3 times in seminars?
- **Requirements**
  - Prepare a **15 minutes** talk on your chosen topic
  - Make **ppt** to assist your talk
  - Prepare **at least 3 questions** to ask the audience after your talk
  - Get ready to **take questions** from the audience
- **Hints**
  - Do statistics to get what you need for Bayes Formula
  - If real numbers are hard to get, make reasonable assumptions
  - Try cost-sensitive problems to make your talk more **challenging** and **interesting**

Questions?