

CHAPTER 3:

BAYESIAN DECISION THEORY

國立雲林科技大學 資訊工程研究所

張傳育(Chuan-Yu Chang) 博士

Office: EB 212

TEL: 05-5342601 ext. 4516

E-mail: chuanyu@yuntech.edu.tw

Website: <http://MIPL.yuntech.edu.tw>

Probability and Inference

2

- Result of tossing a coin is $\in \{\text{Heads}, \text{Tails}\}$
- Random var $x \in \{1, 0\}$
 - $x = 1$ the outcome of a toss is head
 - $x = 0$ the outcome of a toss is tails
- Bernoulli distribution:
 $P\{x = 1\} = p_0, P(x = 0) = 1 - P(x = 1) = 1 - p_0$
- The probability mass function f of this distribution, over possible outcomes k , is
- $$f(k, p_0) = \begin{cases} p_0 & \text{if } x = 1 \\ 1 - p_0 & \text{if } x = 0 \end{cases}$$
- $f(k, p_0) = p_0^x (1 - p_0)^{(1 - x)}$

Probability and Inference

3

- Sample: $\mathbf{X} = \{x^t\}_{t=1}^{t=N}$
- Estimation: $p_0 = \# \{\text{Heads}\} / \#\{\text{Tosses}\} = \sum_t x^t / N$
$$\hat{p}_0 = \frac{\#\{\text{tosses with outcome heads}\}}{\#\{\text{tosses}\}}$$
- Ex: Given the sample { heads, heads, heads, tails, heads, tails, tails, heads, heads }
 $X = \{1, 1, 1, 0, 1, 0, 0, 1, 1\}$
$$\hat{p}_0 = \frac{\sum_{t=1}^N x^t}{N} = \frac{6}{9}$$
- Prediction of next toss:
Heads if $p_0 > 1/2$, Tails otherwise

Classification

- Credit scoring:
Inputs are *income* and *savings*.
Output is *low-risk* vs *high-risk*
- Input: $x = [x_1, x_2]^T$, Output: $C \in \{0, 1\}$
- Prediction:

$$\text{choose } \begin{cases} C = 1 & \text{if } P(C = 1 | x_1, x_2) > 0.5 \\ C = 0 & \text{otherwise} \end{cases}$$

or

$$\text{choose } \begin{cases} C = 1 & \text{if } P(C = 1 | x_1, x_2) > P(C = 0 | x_1, x_2) \\ C = 0 & \text{otherwise} \end{cases}$$

Bayes' Rule

5

Bayes' rule relates the odds of event X_1 to the odds of event X_2 , before (prior to) and after (posterior to) conditioning on another event C .

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

posterior → $P(C | \mathbf{x})$ *prior* → $P(C)$ *evidence* → $p(\mathbf{x})$ *likelihood* → $p(\mathbf{x} | C)$

$$P(C = 0) + P(C = 1) = 1$$

$$p(\mathbf{x}) = p(\mathbf{x} | C = 1)P(C = 1) + p(\mathbf{x} | C = 0)P(C = 0)$$

$$P(C = 0 | \mathbf{x}) + P(C = 1 | \mathbf{x}) = 1$$

Bayes' Rule: $K > 2$ Classes

6

$$\begin{aligned} P(C_i | \mathbf{x}) &= \frac{p(\mathbf{x} | C_i)P(C_i)}{p(\mathbf{x})} \\ &= \frac{p(\mathbf{x} | C_i)P(C_i)}{\sum_{k=1}^K p(\mathbf{x} | C_k)P(C_k)} \end{aligned}$$

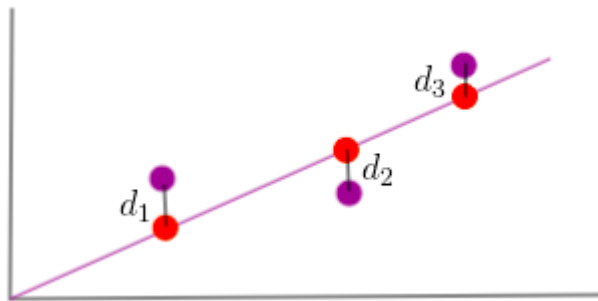
$$P(C_i) \geq 0 \text{ and } \sum_{i=1}^K P(C_i) = 1$$

choose C_i if $P(C_i | \mathbf{x}) = \max_k P(C_k | \mathbf{x})$

Losses and Risks

7

- **Loss** is a function that measures the badness of some particular guess, d , that you made for some value that was actually y .
- **Risk** is the expected value of your loss over all of the guesses that you made.



absolute loss: $L(y,d)=|y-d|$

squared loss : $L(y,d)=(y-d)^2$

Risk= $d_1+d_2+d_3$ (absolute loss)

Risk= $d_1^2+d_2^2+d_3^2$ (squared loss)

Losses and Risks

- Let us define
 - ▣ Actions: α_i (the decision to assign the input to class C_i)
 - ▣ Loss of α_i : λ_{ik} (the loss incurred for taking action α_i when the input belongs to C_k)
 - ▣ Expected risk for taking action α_i (Duda and Hart, 1973)

$$R(\alpha_i | \mathbf{x}) = \sum_{k=1}^K \lambda_{ik} P(C_k | \mathbf{x})$$

choose α_i if $R(\alpha_i | \mathbf{x}) = \min_k R(\alpha_k | \mathbf{x})$

We choose the action with minimum risk

Losses and Risks: 0/1 Loss

9

Define K actions α_i , $i=1, \dots, K$, α_i is the action of assigning \mathbf{x} to C_i .

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ 1 & \text{if } i \neq k \end{cases}$$

正確分類的Loss=0
錯誤分類的Loss=1

The risk of taking action α_i is

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{k=1}^K \lambda_{ik} P(C_k | \mathbf{x}) \\ &= \sum_{k \neq i} P(C_k | \mathbf{x}) && \text{正確分類的Loss=0,} \\ & && \text{僅需統計錯誤分類的部分} \\ &= 1 - P(C_i | \mathbf{x}) && \text{總風險=1-正確分類的部分} \end{aligned}$$

For minimum risk, choose the most probable class

Losses and Risks: Reject

Define action of reject, α_{K+1} , with α_i , $i=1, \dots, K$, being the usual actions of deciding on classes C_i , $i=1, \dots, K$.

A possible loss function is

$$\lambda_{ik} = \begin{cases} 0 & \text{if } i = k \\ \lambda & \text{if } i = K + 1, \quad 0 < \lambda < 1 \\ 1 & \text{otherwise} \end{cases}$$

正確分類的Loss=0
Reject的Loss = λ Eq(3.10)
錯誤分類的Loss=1

The risk of reject is

$$R(\alpha_{K+1} | \mathbf{x}) = \sum_{k=1}^K \lambda P(C_k | \mathbf{x}) = \lambda$$

The risk of choosing class C_i is

$$R(\alpha_i | \mathbf{x}) = \sum_{k \neq i} P(C_k | \mathbf{x}) = 1 - P(C_i | \mathbf{x})$$

Losses and Risks: Reject

11

- The optimal decision rule is to

choose C_i if $R(\alpha_i | \mathbf{x}) < R(\alpha_k | \mathbf{x})$ for $\forall k \neq i$ and $R(\alpha_i | \mathbf{x}) < R(\alpha_{K+1} | \mathbf{x})$
reject if $R(\alpha_{K+1} | \mathbf{x}) < R(\alpha_i | \mathbf{x}), i = 1, \dots, K$

- Given the loss function of Eq(3.10),

choose C_i if $P(C_i | \mathbf{x}) > P(C_k | \mathbf{x}) \forall k \neq i$ and $P(C_i | \mathbf{x}) > 1 - \lambda$
reject otherwise

Example

12

- In a two-class, two-action problem, if the loss function is $\lambda_{11}=\lambda_{22}=0$, $\lambda_{12}=10$, $\lambda_{21}=5$, write the optimal decision rule. How does the rule change if we add a third action of reject with $\lambda=1$?
- Solution:
 - The loss table

Action	Truth	
	C ₁	C ₂
α_1 : choose C ₁	0	10
α_2 : choose C ₂	5	0

Example

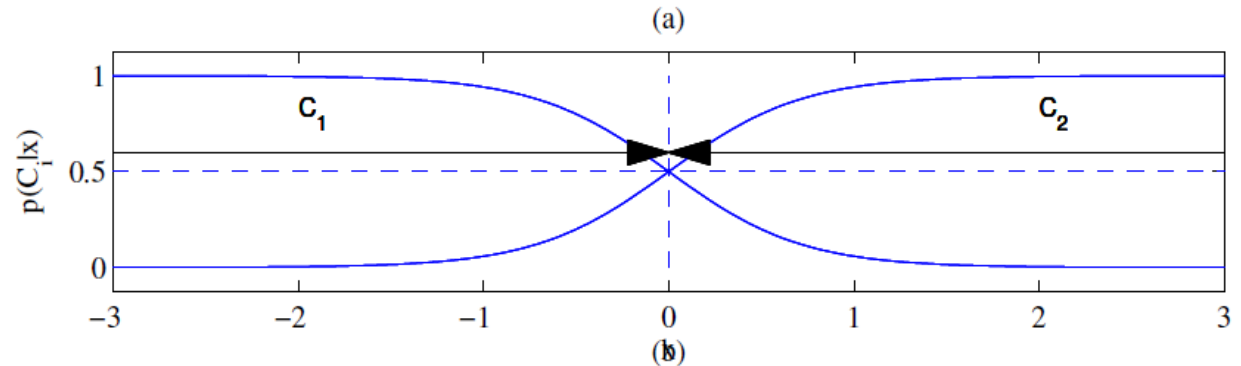
13

- The expected risks of the two actions:
 - ▣ $R(\alpha_1 | x) = 0 \cdot P(C_1 | x) + 10 \cdot P(C_2 | x) = 10 \cdot (1 - P(C_1 | x))$
 - ▣ $R(\alpha_2 | x) = 5 \cdot P(C_1 | x) + 0 \cdot P(C_2 | x) = 5 \cdot (1 - P(C_1 | x))$
- We choose α_1 if
 - ▣ $R(\alpha_1 | x) < R(\alpha_2 | x)$
 - ▣ $10 \cdot (1 - P(C_1 | x)) < 5 \cdot (1 - P(C_1 | x))$
 - ▣ $P(C_1 | x) > 2/3$
- If the two misclassification were equally costly, the decision threshold would be at $1/2$, but because the cost of wrongly choosing C_1 is higher, we want to choose C_1 only when we are really certain.

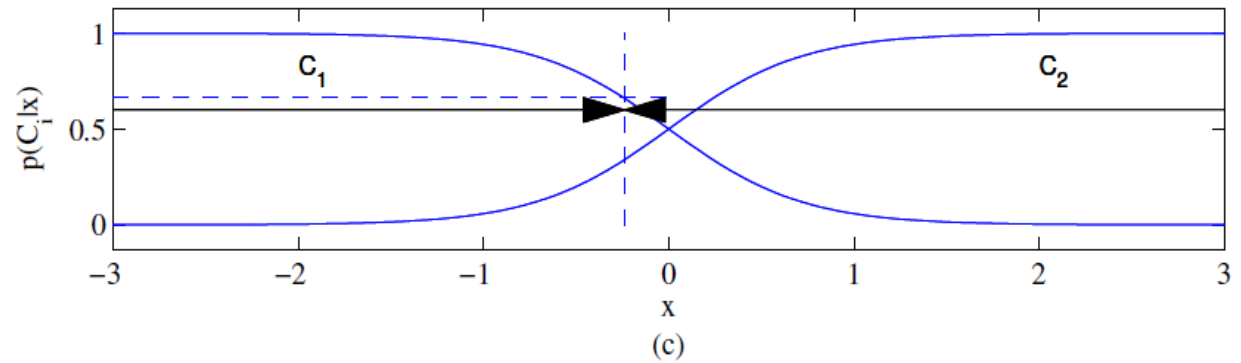
Example: Different Losses and Reject

14

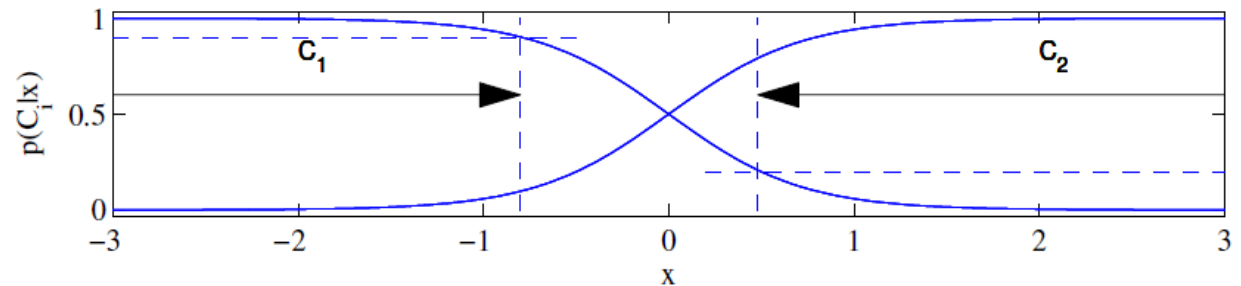
Equal losses



Unequal losses



With reject



Example

15

- If we add a reject option with a cost of 1, the loss table becomes

Action	Truth	
	C ₁	C ₂
α_1 : choose C ₁	0	10
α_2 : choose C ₂	5	0
α_r : Reject	1	1

- The expected risks of the three actions:
 - ▣ $R(\alpha_1 | x) = 0 \cdot P(C_1 | x) + 10 \cdot P(C_2 | x) = 10 \cdot (1 - P(C_1 | x))$
 - ▣ $R(\alpha_2 | x) = 5 \cdot P(C_1 | x) + 0 \cdot P(C_2 | x) = 5 \cdot P(C_1 | x)$
 - ▣ $R(\alpha_r | x) = 1$
- We choose α_1 if
 - ▣ $R(\alpha_1 | x) < 1 \Rightarrow P(C_1 | x) > 9/10$
- We choose α_2 if
 - ▣ $R(\alpha_2 | x) < 1 \Rightarrow P(C_1 | x) < 1/5$, or equivalently if $P(C_1 | x) > 4/5$
- We reject otherwise, if $1/5 < P(C_1 | x) < 9/10$

Discriminant Functions

- Classification can be seen as implementing a set of discriminant functions, $g_i(\mathbf{x}), i=1, \dots, K$, such that we choose C_i if $g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})$

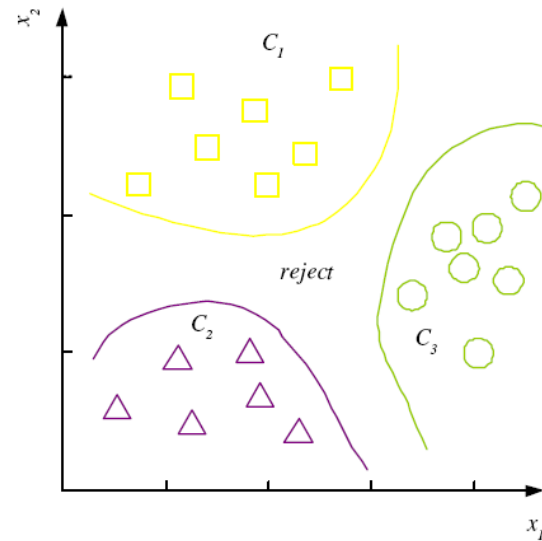
We can represent the Bayes' classifier by setting

$$g_i(\mathbf{x}) = \begin{cases} -R(\alpha_i | \mathbf{x}) \\ P(C_i | \mathbf{x}) \\ p(\mathbf{x} | C_i)P(C_i) \end{cases}$$

K decision regions $\mathcal{R}_1, \dots, \mathcal{R}_K$

$$\mathcal{R}_i = \{\mathbf{x} | g_i(\mathbf{x}) = \max_k g_k(\mathbf{x})\}$$

The regions are separated by **decision boundaries**.



K=2 Classes

- When there are two classes, we can define a single discriminant
 - ▣ $g(\mathbf{x}) = g_1(\mathbf{x}) - g_2(\mathbf{x})$

- And we

$$\text{choose } \begin{cases} C_1 & \text{if } g(\mathbf{x}) > 0 \\ C_2 & \text{otherwise} \end{cases}$$

When $K=2$ the classification system is a dichotomizer(二分器)

$K \geq 3$ is a poly-chotomizer(多分器)

Association Rules

- An association rule is an implication of the form
 $X \rightarrow Y$ where X is the antecedent and Y is the consequent of the rule.
- *People who buy/click/visit/enjoy X are also likely to buy/click/visit/enjoy Y .*
- A rule implies association, not necessarily causation.

Association measures

19

- Support ($X \rightarrow Y$):

$$P(X, Y) = \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers}\}}$$

- Confidence ($X \rightarrow Y$):

$$P(Y | X) = \frac{P(X, Y)}{P(X)}$$

- Lift ($X \rightarrow Y$):

$$= \frac{P(X, Y)}{P(X)P(Y)} = \frac{P(Y | X)}{P(Y)}$$

$$= \frac{\#\{\text{customers who bought } X \text{ and } Y\}}{\#\{\text{customers who bought } X\}}$$

Example

20

- Given the following data of transactions at a shop, calculate the support and confidence values of milk → banana, banana → milk, milk → chocolate, chocolate → milk

Transaction	Items in basket
1	milk, bananas, chocolate
2	milk, chocolate
3	milk, bananas
4	chocolate
5	chocolate
6	milk, chocolate

SOLUTION:

- milk → bananas : Support = 2/6, Confidence = 2/4
- bananas → milk : Support = 2/6, Confidence = 2/2
- milk → chocolate : Support = 3/6, Confidence = 3/4
- chocolate → milk : Support = 3/6, Confidence = 3/5

Apriori algorithm (Agrawal et al., 1996)

21

- **Apriori** is an algorithm for **frequent item set** mining and association rule learning over transactional databases.
- It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database.
- For (X, Y, Z) , a 3-item set, to be frequent (have enough support), (X, Y) , (X, Z) , and (Y, Z) should be frequent.
- If (X, Y) is not frequent, none of its supersets can be frequent.
- Once we find the frequent k -item sets, we convert them to rules: $X, Y \rightarrow Z, \dots$
and $X \rightarrow Y, Z, \dots$

Example

(Ref: https://en.wikipedia.org/wiki/Apriori_algorithm)

22

- Consider the following database, where each row is a transaction and each cell is an individual item of the transaction:

alpha	beta	epsilon
alpha	beta	theta
alpha	beta	epsilon
alpha	beta	theta

- The association rules that can be determined from this database are the following:
 - 100% of sets with alpha also contain beta
 - 50% of sets with alpha, beta also have epsilon
 - 50% of sets with alpha, beta also have theta

Example

(Ref: https://en.wikipedia.org/wiki/Apriori_algorithm)

23

- Let the database of transactions consist of following itemsets:
- Using the Apriori to determine the frequent item sets of this database.
 - ▣ An item set is frequent if it appears in **at least 3 transactions** of the database: the value 3 is the support threshold.
- The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. We obtain the following result

Itemsets
{1,2,3,4}
{1,2,4}
{1,2}
{2,3,4}
{2,3}
{3,4}
{2,4}

Item	Support
{1}	3
{2}	6
{3}	4
{4}	5

Example

(Ref: https://en.wikipedia.org/wiki/Apriori_algorithm)

24

- All the itemsets of size 1 have a support of at least 3, so they are all frequent.
- The next step is to generate a list of all pairs of the frequent items:
- The pairs $\{1,2\}$, $\{2,3\}$, $\{2,4\}$, and $\{3,4\}$ all meet or exceed the minimum support of 3, so they are frequent. The pairs $\{1,3\}$ and $\{1,4\}$ are not and can be pruned.
- Look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:

Item	Support
$\{1,2\}$	3
$\{1,3\}$	1
$\{1,4\}$	2
$\{2,3\}$	3
$\{2,4\}$	4
$\{3,4\}$	3

Item	Support
$\{2,3,4\}$	2

Utility Theory

- Prob of state k given evidence \mathbf{x} : $P(S_k | \mathbf{x})$
- Utility of α_i when state is k : U_{ik}
- Expected utility:
$$EU(\alpha_i | \mathbf{x}) = \sum_k U_{ik} P(S_k | \mathbf{x})$$

Choose α_i if $EU(\alpha_i | \mathbf{x}) = \max_j EU(\alpha_j | \mathbf{x})$