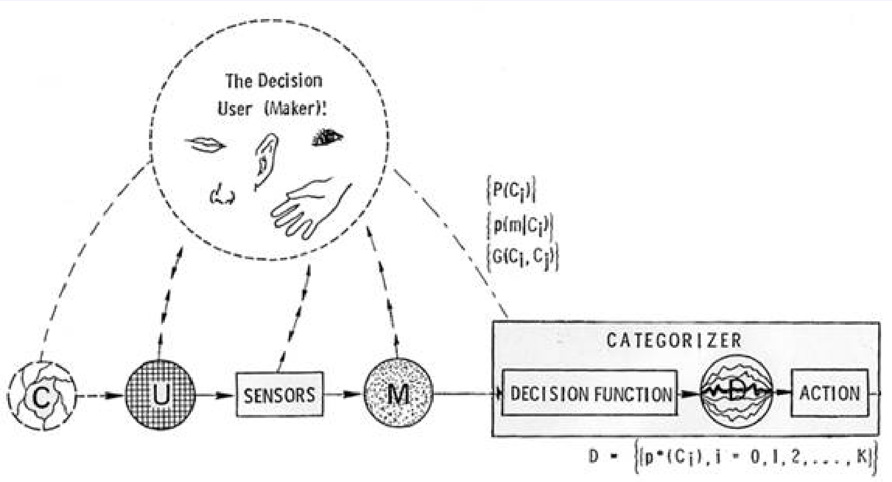
Discrete Bayes

Robert M. Haralick

Computer Science, Graduate Center
City University of New York

Outline

Perspective



Health Example

- Sick patient goes to the physician's assistant
- The assistant examines the patient and orders lab tests
- The lab test results, the measurement vector, is sent by the internet to the assistant's intelligent diagnosis box
- Inside the intelligent diagnosis box is the decision rule
- The decision rule operates on the measurement vector and produces a diagnosis, the assigned illness class

Health Example

- On the basis of the diagnosis, an action, the treatment is given
- There are economic consequences for each possibility
- True class is k and diagnosed class is j
  - For diagnosed class j the treatment is a
  - For diagnosed class k, the treatment is b

The Economic Consequences

- Suppose there is illness j and illness k
- The true, but unknown illness is j
- The smart diagnosis machine diagnoses k, the wrong illness
- So now the treatment is for illness k and it is the wrong treatment
- The treatment costs money and the person stays ill or even might get worse
- The patients suffering can be related to a monetary value

The Economic Consequences

- After suffering from the wrong treatment, the physician assistant is again consulted
- Seeing that the treatment is not working PA decides that the diagnosis the machine produced was wrong.
- The PA gives a treatment for the true illness j
- The patient gets better.
- The PA notes that his smart diagnosis box made an error
- And a cost for the incorrect diagnosis
- The economic consequence is composed of
  - The cost of the PA visit
  - The cost of the lab tests
  - The cost of suffering the wrong treatment
  - The cost of the second PA visit
  - The cost of the second treatment

Economic Consequence

True Class	**Assigned Classs j**	**Assigned Class k**
j	$g - c$	$-3c$
k	$-3c$	$g - c$

Unit of Observation Not Equal to Unit of Classification

Consider the task of recognizing a truck with a machine gun mounted on its back end in the context of guerilla urban warfare. The sensor is a camera. The unit of observation is the pixel, millions of them. No one pixel of the truck carries sufficient information about the class.

Grouping

The units of observation have to be grouped into relevant groups. Classification must be based on features extracted from the relevant groups.

Cyber Attack Units of Observation

The units of observations are transaction records. For example the records might be records of the secure log. The problem is to decide whether or not an IP address should be banned.

```
Aug 16 04:41:06 cunygrid sshd[12535]:
  Invalid user alexis from 221.130.78.253
Aug 16 04:41:06 cunygrid sshd[12543]:
  input_userauth_request: invalid user alexis
Aug 16 04:41:06 cunygrid sshd[12535]:
  pam_unix(sshd:auth): check pass; user unknown
Aug 16 04:41:06 cunygrid sshd[12535]:
  pam_unix(sshd:auth): authentication failure;
  logname= uid=0 euid=0 tty=ssh
  ruser= rhost=221.130.78.253
```

Cyber Attack Units of Observation

- Each observation is a record
- A record is a set of attribute value pairs

Attributes	**Values**
Date	
Time	
Connection ID	
Login ID	
IP address	
Login Id accepted/rejected	
Password accepted/rejected	
Disconnection	

Grouping Units of Observation

- Thread by connection id
- Thread by login ID
- Thread by IP address
- Thread by Time interval

Extracting Features

For a set of records in the given time interval with same IP address:

- Number of records
- Number of login ID records
- Number of login ID's accepted
- Number of disconnects

Extracting Features

For a set of records in the given time interval with same IP address and login ID:

- Number of records
- Number of passwords attempted
- Number of passwords accepted
- Number of disconnects

Grouping and Feature Extracting

The process of grouping together the raw observation units to form the unit of object in the machine learning process and the process of defining the feature vector for the grouped units is part of the art of Machine Learning.

Be Rational, Use Expected Value

When faced with a number of actions, each of which could give rise to more than one possible outcome with different probabilities, the rational action is to:

- Identify all possible outcomes
- Determine their values, utilities, (positive or negative)
- Determine the probabilities that will result from each course of action
- Multiply the two to give an expected value
- The best action is the one that gives rise to the highest expected value

(Blaise Pascal, Pensées, 1670)

Highest Expected Value

- a_1, a_2 Two possible actions
- e_1, e_2, e_3 Three possible economic gains
- The Probabilities for three gains under action a_1
  - $P(e_1|a_1), P(e_2|a_1), P(e_3|a_1)$
- The Probabilities for three gains under action a_2
  - $P(e_1|a_2), P(e_2|a_2), P(e_3|a_2)$

	Probabilities $P(e\|a)$			**Expected Value**
Action	$e_1 = \mathbf{1}$	$e_2 = \mathbf{3}$	$e_3 = \mathbf{0}$	
a_1	.25	.5	.25	$.25 \times 1 + .5 \times 3 + .25 \times 0 = 1.75$
a_2	.5	.2	.3	$.5 \times 1 + .2 \times 3 + .3 \times 0 = 1.1$

Choose Action a_1

Bayes Theorem

- c Class
- d Observed data
- $P(d \mid c)$ The class conditional probabilities

Current state of knowledge $P(c)$. New evidence, the observed data d. What is the updated probability of c: $P(c \mid d)$?

- $P(c \mid d) = \frac{P(c,d)}{P(d)}$
- $P(d \mid c) = \frac{P(c,d)}{P(c)}$
- $P(c \mid d) = \frac{P(d \mid c)P(c)}{P(d)}$

Thomas Bayes (1701-1761), first showed how to use new evidence to update beliefs.

Units Observed

In what follows, the unit being observed is assumed to be that which arises from the grouping process. The measurement is assumed to be the feature vector.

Basic Process Outline

In the pattern discrimination or pattern identification process, a unit is observed or measured and a category assignment is made on the basis of the measurement. This event can be characterized by its three parts:

- unit has true category identification c^j, a member of the set C
- decision rule assignment is to category c^k, a member of the set C
- measurement d is made from the set D

- unit has true category identification c^j , a member of the set C
- decision rule assignment is to category c^k, a member of the set C
- measurement d is made from the set D

Denote this event by (c^j, c^k, d) It has a probability of occurring:

$$P(True = c^j, Assigned = c^k, Meas = d)$$

Economic Consequences

The act of making the assignment carries consequences, economically or in terms of utility.

Assumption

These consequences depend only on which category is the true category identification for the unit and which category is the assigned category identification for the unit. They do not depend on which particular unit is being assigned or on what measurement values the unit to be assigned has.

Economic Consequences: Gain Matrix

		ASSIGNED					
		c^1	c^2	$\cdots$	c^k	$\cdots$	c^K
	c^1			$\cdots$		$\cdots$	
	$\vdots$						
TRUE	c^j			$\cdots$	$e(c^j, c^k)$	$\cdots$	
	$\vdots$			$\vdots$		$\vdots$	
	c^K			$\cdots$		$\cdots$	

The Economic Gain Matrix is Application Dependent

Confusion Matrix

		ASSIGNED					
		c^1	c^2	$\cdots$	c^k	$\cdots$	c^K
	c^1			$\cdots$		$\cdots$	
	$\vdots$						
TRUE	c^j			$\cdots$	$P_{TA}(c^j, c^k)$	$\cdots$	
	$\vdots$			$\vdots$		$\vdots$	
	c^K			$\cdots$		$\cdots$	

The Confusion Matrix Tells How Correct
Is the Machine Learning

Probability of Correct Identification

Definition

The probability of correct identification, P_c, is the probability that the True Identification equals the Assigned Identification. It is defined by

$$P_c = \sum_{j=1}^{K} P_{TA}(c^j, c^j)$$

the sum of the diagonal entries of the confusion matrix.

Confusion Probabilities

What is the probability of observing a unit whose true class identification is c^j and whose assigned class identification is c^k?

$$P_{TA}(c^j, c^k) = \sum_{d \in D} P_{TA}(c^j, c^k, d)$$

Expected Economic Gain

For each true-assigned category identification pair (c^j, c^k) we know its

economic gain	$e(c^j, c^k)$
and probability	$P_{TA}(c^j, c^k)$

Expected Economic Gain

How do we compute the average or expected value of the consequence?

$$
\begin{aligned}
E[e] &= \sum_{j=1}^{K} \sum_{k=1}^{K} e(c^j, c^k) P_{TA}(c^j, c^k) \\
&= \sum_{j=1}^{K} \sum_{k=1}^{K} e(c^j, c^k) \sum_{d \in D} P_{TA}(c^j, c^k, d)
\end{aligned}
$$

Expected Gain

ASSIGNED

TRUE	c^1	c^2		c^K
c^1				
c^2				
			$P_{TA}(c^j, c^k)$	
c^K				

Confusion Matrix

ASSIGNED

TRUE	c^1	c^2		c^K
c^1				
c^2				
			$e(c^j, c^k)$	
c^K				

Economic Gain Matrix

$$E[e] = \sum_{j=1}^{K} \sum_{k=1}^{K} e(c^j, c^k) P_{TA}(c^j, c^k)$$

Analysis

What are the processes which lead to the event (c^j, c^k, d) occurring?

1. There is a unit having measurement d whose true class identification is c^j.
2. Using only the measurement d, the decision rule makes the assignment to class c^k.

Fair Game Assumption

Nature and the Decision Rule are not in collusion

In making the class assignment, the decision rule must only use the measurement data. The decision rule cannot use the true class identification.

- nature only uses d in deciding true class identification c^j
- decision rule only uses d in deciding the assigned class identification c^k

Fair Game Assumption

c^k assigned class
c^j true class
d measurement data
$P_{AT}(c^k \mid (c^j, d)) = P_A(c^k \mid d)$

Therefore,

$$\begin{aligned} P_{AT}(c^k, c^j \mid d) &= \frac{P_{AT}(c^k, c^j, d)}{P(d)} \\ &= \frac{P_{AT}(c^k \mid c^j, d) P_T(c^j, d)}{P(d)} \\ &= P_A(c^k \mid d) P_T(c^j \mid d) \end{aligned}$$

Fair Game Assumption

Nature and the decision rule are not in collusion. Given that a unit has measurements, d, the assigned class which the decision rule determines and the true class which nature determines are statistically independent.

$$P_{AT}(c^k, c^j \mid d) = P_A(c^k \mid d) P_T(c^j \mid d)$$

The Probabilistic Decision Rule

Let $f_d(c^k)$ be the conditional probability that the decision rule assigns a unit to class c^k given that the unit has measurement d. We call f the decision rule.

The Probabilistic Decision Rule

Since, conditioned on measurement d, the true and assigned class are statistically independent,

$$P_{TA}(c^j, c^k \mid d) = P_T(c^j \mid d) f_d(c^k)$$

The Probabilistic Decision Rule

By definition of conditional probability,

$$\begin{aligned} P_{TA}(c^j, c^k \mid d) &= \frac{P_{TA}(c^j, c^k, d)}{P(d)} \\ P_{TA}(c^j, c^k, d) &= P_{TA}(c^j, c^k \mid d) P(d) \end{aligned}$$

The Probabilistic Decision Rule

$$
\begin{aligned}
P_{TA}(c^j, c^k, d) &= P_{TA}(c^j, c^k \mid d)P(d) \\
&= P_T(c^j \mid d)f_d(c^k)P(d) \\
&= P_T(c^j, d)f_d(c^k)
\end{aligned}
$$

Expected Gain

The Expected Gain is then

$$
\begin{aligned}
E[e;f] &= \sum_{j=1}^{K}\sum_{k=1}^{K} e(c^j,c^k) \sum_{d\in D} P_{TA}(c^j,c^k,d) \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} e(c^j,c^k) \sum_{d\in D} P_T(c^j,d) f_d(c^k) \\
&= \sum_{d\in D}\sum_{k=1}^{K} f_d(c^k)\left[\sum_{j=1}^{K} e(c^j,c^k)\; P_T(c^j,d)\right]
\end{aligned}
$$

Bayes Decision Rule

Definition

A decision rule f is called a Bayes decision rule if and only if $E[e; f] \geq E[e; g]$ for any decision rule g.

The problem is how to compute the Bayes Decision Rule

Maximizing

$a_1, \ldots, a_L$ is a sequence of L numbers
$p_1, \ldots, p_L$ is a sequence of unknown numbers satisfying

$$\begin{aligned} p_i &\geq 0 \\ \sum_{i=1}^{L} p_i &= 1 \end{aligned}$$

Find an upper bound for $\sum_{i=1}^{L} a_i p_i$ over all possible such sequences $p_1, \ldots, p_L$ satisfying the constraints.

Maximizing

$$a_i \leq \max_{\substack{j \\ j=1,\ldots,L}} a_j \text{ for } i = 1, \ldots, L$$

Since $p_i \geq 0$,

$$a_i p_i \leq \left[\max_{\substack{j \\ j=1,\ldots,L}} a_j \right] p_i$$

Maximizing

Since inequality holds for each i, it must also hold for the sum.

$$\begin{aligned}
\sum_{i=1}^{L} a_i p_i &\leq \sum_{i=1}^{L} \left[\max_{\substack{j \\ j=1,\ldots,L}} a_j \right] p_i \\
&\leq \left[\max_{\substack{j \\ j=1,\ldots,L}} a_j \right] \sum_{i=1}^{L} p_i \\
&\leq \left[\max_{\substack{j \\ j=1,\ldots,L}} a_j \right]
\end{aligned}$$

An Upper Bound

If $p_1, \ldots, p_L$ is a sequence of L numbers satisfying

$$p_i \geq 0 \quad \text{and} \quad \sum_{i=1}^{L} p_i = 1$$

then

$$\max_{\substack{j \\ j=1,\ldots L}} \; a_j \quad \geq \quad \sum_{i=1}^{L} a_i p_i$$

Thus,

$$\max_{\substack{j \\ j=1,\ldots,L}} \; a_j \quad \text{is an upper bound for} \quad \sum_{i=1}^{L} a_i p_i$$

Maximizing

Is this upper bound, $\max_{j \atop j=1,\ldots,L} a_j$ for $\sum_{i=1}^{L} a_i p_i$

the lowest possible upper bound?

Yes, since it is achievable by appropriate choice of p_i's

Set $p_i = 0$ if $a_i < \max_{j \atop j=1,\ldots,L} a_j$.

Set remaining p_i's so that $0 \le p_i$ and $\sum_{i=1}^{L} p_i = 1$.

Maximizing

Suppose $a_{ij},\ i = 1, 2, \ldots, L$ and $j = 1, 2, \ldots, J$ is a sequence of numbers and p_{ij} satisfies

$0 \leq p_{ij}$ and $\sum_{j=1}^{J} p_{ij} = 1$ for $i = 1, 2, \ldots, L$. Then the lowest upper bound for $\sum_{j=1}^{J} a_{ij}\, p_{ij}$ is $\underset{k=1,\ldots,K}{\max_k}\ a_{ik},\ i = 1, 2, \ldots, L$.

Maximizing

$a_{ij} \leq \max_k a_{ik}$ for each i and j. Since $p_{ij} \geq 0$.
$p_{ij}\, aij \leq p_{ij} \max_k a_{ik}$ for each i and j. Summing over j,

$$\begin{aligned} \sum_{j=1}^{J} p_{ij}\, a_{ij} &\leq \sum_{j=1}^{J} p_{ij} \max_k a_{ik} = \max_k a_{ik} \sum_{j=1}^{J} p_{ij} \\ &\leq \max_k a_{ik}. \end{aligned}$$

Thus, $\max_k a_{ik}$ is an upper bound for $\sum_{j=1}^{J} p_{ij}\, a_{ij}$.

Maximizing

$\max_k a_{ik}$ is the lowest upper bound for $\sum_{j=1}^{J} p_{ij}\, a_{ij}$ since it is achievable.

Set $p_{ij} = 0$ if $a_{ij} \leq \max_k a_{ik}$
Set remaining p_{ij}'s so that
$0 \leq p_{ij}$ and $\sum_{j=1}^{J} p_{ij} = 1$.

Expected Gain

The Expected Gain is then

$$
\begin{aligned}
E[e;f] &= \sum_{j=1}^{K}\sum_{k=1}^{K} e(c^j, c^k) \sum_{d\in D} P_T(c^j, c^k, d) \\
&= \sum_{j=1}^{K}\sum_{k=1}^{K} e(c^j, c^k) \sum_{d\in D} P_T(c^j, d) f_d(c^k) \\
&= \sum_{d\in D}\sum_{k=1}^{K} f_d(c^k) \left[\sum_{j=1}^{K} e(c^j, c^k)\, P_T(c^j, d) \right]
\end{aligned}
$$

Maximizing Expected Gain

But for any conditional probability $f_d(c)$,

$$\sum_{d \in D} \left\{ \sum_{k=1}^{K} f_d(c^k) \left[\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d) \right] \right\} \leq$$
$$\sum_{d \in D} \left\{ \max_n \left[\sum_{j=1}^{K} e(c^j, c^n) P_T(c^j, d) \right] \right\}$$

Maximizing Expected Gain

If for each d, there exists a unique $m = m(d)$ such that

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d) \leq \sum_{j=1}^{K} e(c^j, c^m) P_T(c^j, d), \text{ for each } k,$$

$$\text{Then setting} \quad \left. \begin{array}{l} f_d(c^m) = 1 \\ f_d(c^k) = 0, \ k \neq m \end{array} \right\}$$

achieves the upper bound.

Maximizing Expected Gain

In general, the upper bound is achievable by using a decision rule f defined by $f_d(c^k) = 0$ for each c^k and d such that

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d) < \max_n \sum_{j=1}^{K} e(c^j, c^n) P_T(c^j, d)$$

and the remaining $f_d(c^k)$ are set in any way so that

$$\begin{aligned} f_d(c^k) &\geq 0 \\ \sum_{k=1}^{K} f_d(c^k) &= 1 \end{aligned}$$

Bayes Decision Rule

Definition

A decision rule f is called a Bayes decision rule if and only if $E[e; f] \geq E[e; g]$ for any decision rule g.

Bayes Decision Rule

A Bayes rule can always be implemented as a deterministic decision rule

ASSIGNED

			c^1	c^2		c^K
T	c^1	$P_T(c^1, d)$	$e(c^1, c^1)$	$e(c^1, c^2)$		$e(c^1, c^K)$
R	c^2	$P_T(c^2, d)$	$e(c^2, c^1)$			
U					$\dots$	
E					$\vdots$	
	c^K	$P_T(c^K, d)$	$e(c^K, c^1)$	$e(c^K, c^2)$		$e(c^K, c^K)$

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d)$$

Assign any c^k to d such that $\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d)$ is maximized

Bayes Decision Rule

Given measurement d, assign to any class c^k satisfying

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d) \geq \sum_{j=1}^{K} e(c^j, c^n) P_T(c^j, d), n = 1, \ldots, K$$

If the economic gain matrix e is the identity, then assign to any class c^k satisfying

$$P_T(c^k, d) \geq P_T(c^n, d), n = 1, \ldots, K$$

$$E[e] = \sum_{d \in D} \max_{\substack{n \\ n=1,\ldots,K}} P_T(c^n, d) = P_{correct}$$

Bayes Decision Rule: d^1

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04

e	Assigned	
True	c^1	c^2
c^1	1	0
c^2	0	2

$f_d(c)$	Measurement		
Assigned	d^1	d^2	d^3
c^1	0	0	1
c^2	1	1	0

For d^1 : bring $\begin{pmatrix} .12 \\ .2 \end{pmatrix}$ to the first column (assign c^1) of the economic gain matrix, take the sum of products to get answer $.12 \times 1 = .12$ Then bring it to the second column (assign c^2) of the economic gain matrix, take the sum of products to get answer $.2 \times 2 = .4$
Since $.4 > .12$ assign to class c^2

Bayes Decision Rule: d^2

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04

e	Assigned	
True	c^1	c^2
c^1	1	0
c^2	0	2

$f_d(c)$	Measurement		
Assigned	d^1	d^2	d^3
c^1	0	0	1
c^2	1	1	0

For d^2 : bring $\begin{pmatrix} .18 \\ .16 \end{pmatrix}$ to the first column (assign c^1) of the economic gain matrix, take the sum of products to get answer $.18 \times 1 = .18$ Then bring it to the second column (assign c^2) of the economic gain matrix, take the sum of products to get answer $.16 \times 2 = .32$
Since $.32 > .18$ assign to class c^2

Bayes Decision Rule: d^3

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04

e	Assigned	
True	c^1	c^2
c^1	1	0
c^2	0	2

$f_d(c)$	Measurement		
Assigned	d^1	d^2	d^3
c^1	0	0	1
c^2	1	1	0

For d^3 : bring $\begin{pmatrix} .3 \\ .04 \end{pmatrix}$ to the first column (assign c^1) of the economic gain matrix, take the sum of products to get answer $.3 \times 1 = .3$ Then bring it to the second column (assign c^2) of the economic gain matrix, take the sum of products to get answer $.04 \times 2 = .08$
Since $.3 > .08$ assign to class c^1

Bayes Decision Rule

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04

e	Assigned	
True	c^1	c^2
c^1	1	0
c^2	0	2

$f_d(c)$	Measurement		
Assigned	d^1	d^2	d^3
c^1	0	0	1
c^2	1	1	0

$$E[e] = .2 \times 2 + .16 \times 2 + .3 \times 1 = 1.02$$

Non-Bayes Decision Rule

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04

e	Assigned	
True	c^1	c^2
c^1	1	0
c^2	0	2

A Non-Bayes Rule

$f_d(c)$	Measurement		
Assigned	d^1	d^2	d^3
c^1	0	1	1
c^2	1	0	0

$$E[e] = .2 \times 2 + .18 \times 1 + .3 \times 1 = .88$$

The Confusion Matrix

$$
\begin{aligned}
P_{TA}(c^1,c^1) &= \sum_{d\in D} f_d(c^1)P_T(c^1,d) = .18+.3 \\
P_{TA}(c^1,c^2) &= \sum_{d\in D} f_d(c^2)P_T(c^1,d) = .12 \\
P_{TA}(c^2,c^1) &= \sum_{d\in D} f_d(c^1)P_T(c^2,d) = .16+.04 \\
P_{TA}(c^2,c^2) &= \sum_{d\in D} f_d(c^2)P_T(c^2,d) = .2
\end{aligned}
$$

P_{TA}	Assigned	
True	c^1	c^2
c^1	$P_{TA}(c^1,c^1)$	$P_{TA}(c^1,c^2)$
c^2	$P_{TA}(c^2,c^1)$	$P_{TA}(c^2,c^2)$

The Confusion Matrix

$P_T(c,d)$	Measurement		
True	d^1	d^2	d^3
c^1	.12	.18	.3
c^2	.2	.16	.04
f_d	c^2	c^1	c^1

P_{TA}	Assigned		P_T
True	c^1	c^2	
c^1	.48	.12	.6
c^2	.2	.2	.4

$$\begin{aligned} P_{correct\_assign} &= .48 + .4 \\ &= .68 \end{aligned}$$

Bayes Decision Rule

Given measurement d, assign to any class c^k satisfying

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j, d) \geq \sum_{j=1}^{K} e(c^j, c^n) P_T(c^j, d), n = 1, \ldots, K$$

Or equivalently, assign to any class c^k satisfying

$$\sum_{j=1}^{K} e(c^j, c^k) \frac{P_T(c^j, d)}{P(d)} \geq \sum_{j=1}^{K} e(c^j, c^n) \frac{P_T(c^j, d)}{P(d)}, n = 1, \ldots, K$$

Bayes Decision Rule

Given measurement d, assign to any class c^k satisfying

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j|d) \geq \sum_{j=1}^{K} e(c^j, c^n) P_T(c^j|d), n = 1, \ldots, K$$

Therefore, the only probability information we need for a Bayes rule are the conditional probabilities

$$P_T(c^j|d), \; j = 1, \ldots, K$$

Two Class Bayes Rule

Assume the economic gain for a correct assignment is greater than the gain for an incorrect assignment.

$$\begin{aligned} e(c^1,c^1) &> e(c^1,c^2) \\ e(c^2,c^2) &> e(c^2,c^1) \end{aligned}$$

Assume all probabilities are positive
Assign to class c^1 if

$$P(c^1|d)e(c^1,c^1) + P(c^2|d)e(c^2,c^1) \geq P(c^1|d)e(c^1,c^2) + P(c^2|d)e(c^2,c^2)$$

Otherwise assign to class c^2.

This inequality can be simplified.

$$\begin{aligned} P(c^1|d)e(c^1,c^1) - P(c^1|d)e(c^1,c^2) &\geq P(c^2|d)e(c^2,c^2) - P(c^2|d)e(c^2,c^1) \\ P(c^1|d)(e(c^1,c^1) - e(c^1,c^2)) &\geq P(c^2|d)(e(c^2,c^2) - e(c^2,c^1)) \\ \frac{P(c^1|d)}{P(c^2|d)} &\geq \frac{e(c^2,c^2) - e(c^2,c^1)}{e(c^1,c^1) - e(c^1,c^2)} \end{aligned}$$

Two Class Bayes Rule: Odd's Ratio

Assign to class c^1 if

$$\frac{P(c^1|d)}{P(c^2|d)} \geq \frac{e(c^2,c^2) - e(c^2,c^1)}{e(c^1,c^1) - e(c^1,c^2)}$$

Otherwise assign to class c^2.
The odd's ratio $\mathcal{R}$ in favor of class c^1 is defined by

$$\mathcal{R}(d) = \frac{P(c^1|d)}{P(c^2|d)}$$

Assign to class c^1 if

$$\begin{aligned} \mathcal{R}(d) &\geq \frac{e(c^2,c^2) - e(c^2,c^1)}{e(c^1,c^1) - e(c^1,c^2)} \\ &\geq \Theta \end{aligned}$$

Two Class Bayes Rule: Likelihood Ratio

The Likelihood ratio in favor of class c^1 is defined by

$$\mathcal{L}(d) = \frac{P(d|c^1)}{P(d|c^2)}$$

Assign to class c^1 if

$$\begin{aligned}
\frac{P(c^1|d)}{P(c^2|d)} &\geq \frac{e(c^2, c^2) - e(c^2, c^1)}{e(c^1, c^1) - e(c^1, c^2)} \\
\frac{P(c^1|d)P(d)/P(c^1)}{P(c^2|d)P(d)/P(c^2)} &\geq \frac{e(c^2, c^2) - e(c^2, c^1)}{e(c^1, c^1) - e(c^1, c^2)} \frac{P(c^2)}{P(c^1)} \\
\mathcal{L}(d) = \frac{P(d|c^1)}{P(d|c^2)} &\geq \frac{e(c^2, c^2) - e(c^2, c^1)}{e(c^1, c^1) - e(c^1, c^2)} \frac{P(c^2)}{P(c^1)} \\
\mathcal{L}(d) &\geq \Theta
\end{aligned}$$

Two Class Bayes Rule: Class Conditional Probability

Assign to class c^1 if

$$P(c^1|d)e(c^1,c^1) + P(c^2|d)e(c^2,c^1) \geq P(c^1|d)e(c^1,c^2) + P(c^2|d)e(c^2,c^2)$$

Otherwise assign to class c^2.

$$\begin{aligned} P(c^1|d)(e(c^1,c^1) - e(c^1,c^2)) &\geq P(c^2|d)(e(c^2,c^2) - e(c^2,c^1)) \\ P(c^1|d)(e(c^1,c^1) - e(c^1,c^2)) &\geq (1 - P(c^1|d))(e(c^2,c^2) - e(c^2,c^1)) \end{aligned}$$

$$P(c^1|d)(e(c^1,c^1) - e(c^1,c^2) + e(c^2,c^2) - e(c^2,c^1)) \geq e(c^2,c^2) - e(c^2,c^1)$$

$$\begin{aligned} P(c^1|d) &\geq \frac{e(c^2,c^2) - e(c^2,c^1)}{e(c^1,c^1) - e(c^1,c^2) + e(c^2,c^2) - e(c^2,c^1)} \\ &\geq \Theta \end{aligned}$$

Summary

Given measurement d, assign to any class c^k satisfying

$$\sum_{j=1}^{K} e(c^j, c^k) P_T(c^j|d) \geq \sum_{j=1}^{K} e(c^j, c^n) P_T(c^j|d), n = 1, \ldots, K$$

Therefore, the only probability information we need for a Bayes rule are the conditional probabilities

$$P_T(c^j|d),\ j = 1, \ldots, K$$

Summary

$$
\begin{aligned}
P_T(c \mid d) &= \frac{P_T(d \mid c)P(c)}{P(d)} \\
&= \frac{P_T(d \mid c)P(c)}{\sum_{\gamma \in C} P_T(d \mid \gamma)P(\gamma)}
\end{aligned}
$$

- For a given c, $P_T(d \mid c)$ is the class conditional distribution
- $P_T(c)$ is the prior probability of class c

Estimating $P_T(d \mid c)$ from data is a major problem in Machine Learning

Same Decision Rules Different Economic Gains

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \begin{bmatrix} 0 & -1 & -1 \\ -1 & 0 & -1 \\ -1 & -1 & 0 \end{bmatrix} \quad \begin{bmatrix} 1 & -1 & -1 \\ -1 & 1 & -1 \\ -1 & -1 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 5 & -1 & -1 \\ -1 & 5 & -1 \\ -1 & -1 & 5 \end{bmatrix} \quad \begin{bmatrix} a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}$$

Same Decision Rule Different Economic Gain

$$(a > b)$$

$$(a-b)\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} + \begin{bmatrix} b & b & b \\ b & b & b \\ b & b & b \end{bmatrix} = \begin{bmatrix} a & b & b \\ b & a & b \\ b & b & a \end{bmatrix}$$