**ECE 645: Estimation Theory**
**Spring 2015**
**Instructor: Prof. Stanley H. Chan**

PURDUE UNIVERSITY

# Lecture Note 1: Bayesian Decision Theory

(LaTeX prepared by Stylianos Chatzidakis)
March 31, 2015

This lecture note is based on ECE 645(Spring 2015) by Prof. Stanley H. Chan in the School of Electrical and Computer Engineering at Purdue University.

## 1 Introduction

Classification appears in many disciplines for pattern recognition and detection methods. In this lecture we introduce the Bayesian decision theory, which is based on the existence of prior distributions of the parameters.

### 1.1 Bayesian Detection Framework

Before we discuss the details of the Bayesian detection, let us take a quick tour about the overall framework to detect (or classify) an object in practice. In the Bayesian setting, we model observations as random samples drawn from some probability distributions. The classification process usually involves extracting features from the observations, and a decision rule that satisfies certain optimality criterion. See Figure 1.

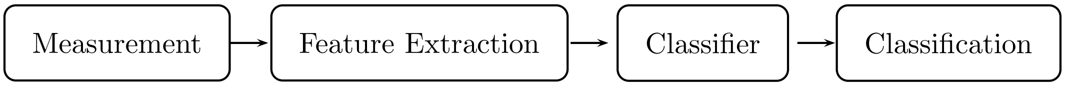


Figure 1: Block diagram of a classifier

When the distributions of the random samples are not known (which is true in most real-world applications), we might need an estimation algorithm to first determine the parameters of the distributions, e.g., mean and standard deviation. A decision rule can then designed based on these estimated parameters. To verify the efficiency of the classifier, testing data are used to calculate the error rate or false alarm rate. In most cases, a classifier with small false alarm rate is sought. This process is shown in Figure 2.

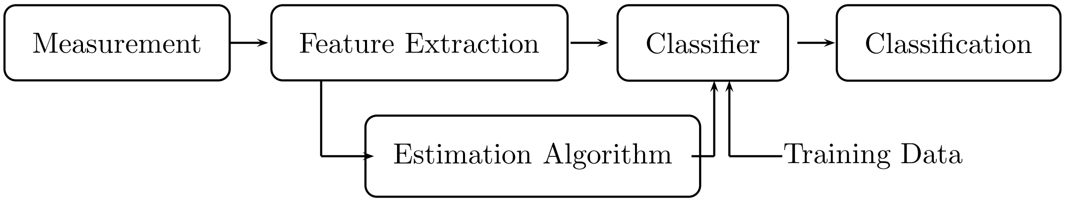


Figure 2: classifier

In practice, of course, all the above building blocks have to be taken into account. However, to help us understand the important ideas of the detection theory, we will focus on the design of the classifiers in this course. Interested readers can consult standard textbooks on pattern recognitions for detailed discussions on these practical issues.

### 1.2 Objectives and Organizations

We begin this lecture note with a brief review of probability. We assume that readers are familiar with introductory probability theory (e.g., ECE 600). After reviewing probability theory, we will discuss the general Bayes' decision rule. Then, we will discuss three special cases of the general Bayes' decision rule: Maximum-a-posteriori (MAP) decision, Binary hypothesis testing, and M-ary hypothesis testing.

## 2 Review of Probability

### 2.1 Probability Space

Any random experiment can be defined using the probability space $(\mathcal{S}, \mathcal{F}, \mathbb{P})$ where $\mathcal{S}$ is the sample space, $\mathcal{F}$ is the event space, and $\mathbb{P}$ is the probability mapping. The sample space $\mathcal{S}$ is a non-empty set containing all outcomes of the experiments. The event space $\mathcal{F}$ is a collection of subsets of $\mathcal{S}$ to which probabilities are assigned. The event space $\mathcal{F}$ must be a non-empty set that satisfies the properties of a $\sigma$-field. The probability mapping is a set function that assigns a real number to every set:

$$\mathbb{P} : \mathcal{F} \rightarrow \mathbb{R} \tag{1}$$

and must satisfy the following three probability axioms:

| | |
|---|---|
| Non-negativity: | $\mathbb{P}(A) \geq 0$, for all $A \in \mathcal{F}$ |
| Normalization: | $\mathbb{P}(\mathcal{S}) = 1$ |
| Additivity: | $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$ if $A \cap B = \emptyset$. |

### 2.2 Conditional Probability

In many situations we would want to know the probability of an event $A$ occurring given that another event $B$ has occurred. In this case, the probability of an event $A$ given that another event $B$ has occured is called conditional probability. The condition probability of $A$ given $B$ is defined as:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \tag{2}$$

assuming $\mathbb{P}(B) > 0$.

---

**Example 1.**
Consider rolling a die. The probability of event $A = 6$ is equal to $1/6$. However, if someone provides additional information, let's say that the event $B$ =roll of a die was bigger than 4, then the probability of $A$ given $B$ is:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} = \frac{1/6}{3/6} = 1/3.$$

A simple calculation of conditional probability allows us to write:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\mathbb{P}(B) \tag{3}$$

and

$$\mathbb{P}(B \cap A) = \mathbb{P}(B|A)\mathbb{P}(A) \tag{4}$$

then equating the left and right hand sides we can derive the Bayes' Theorem:

**Theorem 1.** Bayes' Theorem
For any events $A$ and $B$ such that $\mathbb{P}(B) > 0$,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)P(A)}{\mathbb{P}(B)}. \tag{5}$$

The Bayes' theorem can be generalized to yield the following result.

**Theorem 2.** Law of Total Probability
If $A_1, A_2, \ldots, A_n$ is a partition of the sample space and $B$ is an event in the event space, then

$$\mathbb{P}(B) = \sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i) \tag{6}$$

The law of total probability suggests that for any event $B$, we can decompose $B$ into a sum of $n$ disjoint subsets $A_i$. Moreover, applying the total probability law to Bayes theorem yields

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)P(A)}{\sum_{i=1}^{n} \mathbb{P}(B|A_i)\mathbb{P}(A_i)} \tag{7}$$

for $A, B \in \mathcal{F}$, $\mathbb{P}(A) > 0$ and $\mathbb{P}(B) > 0$.

### 2.3 Random Variables

A random variable is a real function from the sample space to the real numbers:

$$X : \mathcal{S} \to \mathbb{R} \tag{8}$$

A random variable can be discrete or continuous. For the discrete case, the probability mass function is defined as

$$p_X(x) = \mathbb{P}(\omega \in S : X(\omega) = x) \tag{9}$$

For the continuous case, the cumulative distribution function is defined as

$$F_X(x) = \mathbb{P}(\omega \in S : X(\omega) \leq x) \tag{10}$$

When the cumulative distribution function is differentiable, we can define the probability density function as

$$f_X(x) = \frac{d}{dx}F_X(x). \tag{11}$$

### 2.4 Expectations

The expectation of a random variable described by a probability mass function or a probability density function is

$$\mathbb{E}[X] = \begin{cases} \sum x p_X(x), & \text{if } X \text{ is discrete,} \\ \int_{-\infty}^{\infty} x f_X(x) dx, & \text{if } X \text{ is continuous.} \end{cases} \tag{12}$$

The conditional expectation is

$$\mathbb{E}[X|Y=y] = \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \tag{13}$$

The variance is defined as:

$$\text{Var}[X] = E[(X - E[X])^2] = \mathbb{E}[X^2] - \mathbb{E}[X]^2. \tag{14}$$

A very useful result of the expectation is the total expectation formula, also known as the iterated expectation.

**Theorem 3.** Total Expectation

$$\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y=y]] \tag{15}$$

**Proof.**

$$\begin{aligned}
\mathbb{E}[X] &= \int_{-\infty}^{\infty} x f_X(x) dx = \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{XY}(x,y) dx dy \\
&= \int_{-\infty}^{\infty} x \int_{-\infty}^{\infty} f_{X|Y}(x|y) dy dx \\
&= \int_{-\infty}^{\infty} f_Y(y) \int_{-\infty}^{\infty} x f_{X|Y}(x|y) dx \\
&= \int_{-\infty}^{\infty} \mathbb{E}[X|Y=y] f_Y(y) dy = \mathbb{E}_Y[\mathbb{E}_{X|Y}[X|Y=y]].
\end{aligned}$$

□

### 2.5 Gaussian Distribution

Finally we review the Gaussian distribution. A single variable Gaussian distribution is defined as

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}, \tag{16}$$

where $\mu$ is the mean and $\sigma^2$ is the variance. We write

$$X \sim \mathcal{N}(\mu, \sigma^2) \tag{17}$$

to denote a random variable $X$ drawn from a Gaussian distribution.

For multivariate Gaussian, the distribution is

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right\}, \tag{18}$$

where $\boldsymbol{X} = [X_1, X_2, \cdots, X_d]^T$ is a $d$-dimensional vector, $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_d]^T$ is the mean vector, and

$$\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{X}-\boldsymbol{\mu})(\boldsymbol{X}-\boldsymbol{\mu})^T] = \begin{bmatrix} \text{Var}(X_1) & \cdots & \text{Cov}(X_1, X_d) \\ \vdots & \ddots & \vdots \\ \text{Cov}(X_1, X_d) & \cdots & \text{Var}(X_d) \end{bmatrix} \tag{19}$$

is the covariance matrix.

If $\text{Cov}(X_i, X_j) = 0$ then $X_i$ and $X_j$ are said to be uncorrelated. If $\text{Cov}(X_i, X_j) > 0$ then $X_i$ and $X_j$ are said to be positively correlated. However, it should be clarified that uncorrelated does not imply independent, because $\text{Cov}(X, Y) = 0$ only implies $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ but not $f_{X,Y}(x, y) = f_X(x)f_Y(y)$. The converse is true however. That is, if $X$ and $Y$ are independent, then $\text{Cov}(X, Y) = 0$. Here is a counter example by C. Shalizi.

---

**Example 2.**
Let $X \sim \text{Uniform}(-1, 1)$, and let $Y = |X|$. Then,

$$\mathbb{E}[XY \mid X \geq 0] = \int_0^1 x^2 dx = 1/3$$
$$\mathbb{E}[XY \mid X < 0] = \int_{-1}^0 x^2 dx = -1/3.$$

Thus, by Law of Total Expectation we have $\mathbb{E}[XY] = 0$. However, $X$ and $Y$ are clearly dependent.

---

# 3 Bayesian Decision Theory

In Bayes's detection theory, we are interested in computing the posterior distribution $f_{\Theta|X}(\theta|x)$. Using Bayes' theorem, it is easy to show that the posterior distribution $f_{\Theta|X}(\theta|x)$ can be computed via the conditional distribution $f_{X|\Theta}(x|\theta)$ and the prior distribution $f_\Theta(\theta)$. The prior distribution $f_\Theta(\theta)$ represents the prior knowledge we may have for the distribution of the $\theta$ parameter before we obtain additional information for our dataset. In other words, Bayes' detection theory utilizes prior knowledge in the decision.

Bayes' theorem can be used for discrete or continuous random variables. For discrete random variables it takes the form:

$$p_{\Theta|Y}(\theta|y) = \frac{p_{Y|\Theta}(y|\theta)p_\Theta(\theta)}{p_Y(y)}, \tag{20}$$

where $p$ represents the probability mass function. For continuous random variables:

$$f_{\Theta|Y}(\theta|y) = \frac{f_{Y|\Theta}(y|\theta)f_\Theta(\theta)}{f_Y(y)}, \tag{21}$$

where $f$ is the probability density function.

## 3.1 Notations

To facilitate the subsequent discussion, we introduce the following notations.

1. Parameter $\Theta$. We assume that $\Theta$ is a random variable with realization $\Theta = \theta$. The domain of $\theta$ is defined as $\Lambda$. For detection, we assume that $\Lambda$ is a collection of $M$ states, i.e., $\Lambda \overset{\text{def}}{=} \{0, 1, ...., M-1\}$.

2. Prior distributions $\pi_0, \pi_1, \ldots, \pi_{M-1}$, where $\pi_j = \mathbb{P}(\Theta = j)$. Note that the sum of all $\pi_j$ should be 1:
$$\sum_{j=0}^{M-1} \pi_j = 1. \tag{22}$$

3. Conditional distribution of observing $Y = y$ given that $\Theta = j$:
$$f_j(y) \overset{\text{def}}{=} \mathbb{P}(Y = y | H_j), \tag{23}$$
where $H_j$ denotes the hypothesis that $\Theta = j$.

4. Posterior distributions of having $\Theta = j$ given the observation $Y = y$:
$$\pi_j(y) \overset{\text{def}}{=} \mathbb{P}(H_j | Y = y). \tag{24}$$
By Bayes' theorem, we can show that
$$\mathbb{P}(H_j | Y = y) = \frac{\mathbb{P}(Y = y | H_j) p(H_j)}{\mathbb{P}(Y = y)}, \tag{25}$$
and so
$$\pi_j(y) = \frac{f_j(y)\pi_j}{\sum_j f_j(y)\pi_j}. \tag{26}$$

5. Decision rule $\delta : \Gamma \to \Lambda$. The decision rule is a function that takes an input $y \in \Gamma$ and sends $y$ to a value $\delta(y) \in \Lambda$.

6. Cost function $C(i, j)$ or $C_{ij}$. In detection or classification of objects, every decision is accompanied by a cost. If, for example, there is a flying object or a disease and we are not able to detect, then there is cost with this decision. That is, if we decide that there is no signal but instead there is signal, then we call this a miss. In the case were there is nothing present but we decide that there is, then we have a false alarm. Sometimes the cost is very small or significant depending on the situation. For example, it is preferable to have false alarm than a miss in the case of disease detection. The cost associated with each decision is described by the cost function. Here, we use $C_{ij}$ to describe the cost of choosing $H_i$ when $H_j$ holds. For a binary hypothesis, the cost function can be represented by a table:

| | $H_0$ | $H_1$ |
|---|---|---|
| $\delta(y) = 0$ | $C_{00}$ | $C_{01}$ |
| $\delta(y) = 1$ | $C_{10}$ | $C_{11}$ |

For example, $C_{01}$ is the cost associated with selecting $H_0$ when $H_1$ was the true value, i.e., the cost of having a miss. Similarly, $C_{10}$ is the cost of having a false alarm. $C_{00}$ and $C_{11}$ are the cost of having the correct detection.

### 3.2 Bayesian Risk

The goal of Bayesian detection is to minimize the risk, defined as

$$R(\delta) = \mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)]. \tag{27}$$

In other words, the optimal decision rule is

$$\delta(y) = \underset{\delta}{\operatorname{argmin}} \quad R(\delta). \tag{28}$$

Minimizing the risk defined as the expectation of the cost function is analytically very difficult as it involves the minimization of the double integral. To solve this problem we observe the following result:

**Proposition 1.**

$$\delta(y) = \underset{\delta}{\operatorname{argmin}} \quad \mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \underset{i}{\operatorname{argmin}} \quad \sum_{j=0}^{M-1} C(i,j)\pi_j(y). \tag{29}$$

To prove the above proposition we need to make use of the total expectation, and in particular the following lemma:

**Lemma 1.**

$$\mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \mathbb{E}_Y[\mathbb{E}_{\Theta|Y}[C(\delta(Y), \Theta)|Y = y]]. \tag{30}$$

**Proof.**
By definition of $\mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)]$, we have

$$\mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \iint C(\delta(y), \theta) f_{Y\Theta}(y, \theta) dy d\theta.$$

Since $f_{Y\Theta}(y, \theta) = f_{\Theta|Y}(\theta|y) f_Y(y)$ by Bayes' theorem, we have

$$\mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \iint C(\delta(y), \theta) f_{\Theta|Y}(\theta|y) f_Y(y) d\theta dy.$$

Switching the order of integration yields

$$\iint C(\delta(y), \theta) f_{\Theta|Y}(\theta|y) f_Y(y) d\theta dy = \int f_Y(y) \int C(\delta(y), \theta) f_{\Theta|Y}(\theta|y) d\theta dy,$$

in which we see that the inner integration is $\mathbb{E}_{\Theta|Y}[C(\delta(Y), \theta)|Y = y]$. Therefore,

$$\begin{aligned} \int f_Y(y) \int C(\delta(y), \theta) f_{\Theta|Y}(\theta|y) d\theta dy &= \int f_Y(y) \mathbb{E}_{\Theta|Y}[C(\delta(Y), \theta)|Y = y] dy \\ &= E_Y[E_{\Theta|Y}[C(\delta(Y), \Theta)|Y = y]]. \end{aligned}$$

□

Using the Lemma we can prove the proposition.

**Proof.**
By Lemma, we have that

$$\underset{\delta}{\operatorname{argmin}} \ \mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \underset{\delta}{\operatorname{argmin}} \ \int \mathbb{E}_{\Theta|Y}[C(\delta(Y), \Theta)|Y = y] f_Y(y) dy.$$

Since $f_Y(y)$ is non-negative, the minimizer of the integral is the same as the minimizer of the inner expectation. Therefore, we have

$$\underset{\delta}{\operatorname{argmin}} \ \mathbb{E}_{Y\Theta}[C(\delta(Y), \Theta)] = \underset{\delta}{\operatorname{argmin}} \ \mathbb{E}_{\Theta|Y}[C(\delta(Y), \Theta)|Y = y].$$

Expressing out the definition of the conditional expectation, we have

$$\delta(y) = \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} C(i,j)\pi_j(y).$$

□

We remark that $\delta(y)$ is a function of $y$. That is, for a different observation $y$, the decision value $\delta(y)$ is different. To denote that this the Bayesian decision rule, we put a subscript $\delta_B(y)$.

### 3.3 Maximum-A-Posteriori rule

We now consider a special case where the cost function is uniform, defined as

$$C(i,j) = \begin{cases} 1, & i \neq j, \\ 0, & i = j. \end{cases} \tag{31}$$

In this case, the decision rule becomes

$$\begin{aligned} \delta_B(y) &= \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} C(i,j)\pi_j(y) \\ &= \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} \pi_j(y) \\ &= \underset{i}{\operatorname{argmin}} \ (1 - \pi_j(y)) \\ &= \underset{i}{\operatorname{argmax}} \ \pi_i(y). \end{aligned}$$

Therefore, for uniform cost, the risk is minimized by maximizing the posterior distribution. Thus we call the resulting decision rule as the Maximum-A-Posteriori (MAP) rule.

An important property of the MAP rule is that is minimizes the probability of error.

**Definition 1.**
The probability of error is defined as

$$\mathbb{P}_{error} = \mathbb{P}(\Theta \neq \delta(Y)). \tag{32}$$

**Proposition 2.**
For any decision rule $\delta$, and for a uniform cost,

$$\mathbb{P}_{error} = R(\delta).$$

**Proof.**
First of all, we note by the law of total probability that

$$\mathbb{P}_{error} = \mathbb{P}(\Theta \neq \delta(Y)) = \int_{-\infty}^{\infty} \mathbb{P}(\Theta \neq \delta(Y)|Y = y) f_Y(y) dy.$$

The conditional probability inside the integral can be written as:

$$\begin{aligned} \mathbb{P}(\Theta \neq \delta(Y)|Y = y) &= 1 - \mathbb{P}(\Theta = \delta(Y)|Y = y) \\ &= \sum_{j=0}^{M-1} \mathbb{P}(\Theta = j|Y = y). \end{aligned}$$

By using the uniform cost, we have

$$\begin{aligned} \sum_{j=0}^{M-1} \mathbb{P}(\Theta = \delta(Y)|Y = y) &= \sum_{j=0}^{M-1} C(\delta(y), j)\mathbb{P}(\Theta = \delta(Y)|Y = y) \\ &= \mathbb{E}_{\Theta|Y}[C(\delta(Y), \Theta), \Theta|Y]. \end{aligned}$$

Therefore the probability of error becomes:

$$\mathbb{P}_{error} = \int_{-\infty}^{\infty} \mathbb{P}(\Theta \neq \delta(Y)|Y = y) f_Y(y) dy = \int_{-\infty}^{\infty} \mathbb{E}_{\Theta|Y}[C(\delta(Y), \Theta), \Theta|Y] f_Y(y) dy,$$

which is equal to the expectation of the cost which is the definition of the risk. Therefore,

$$\mathbb{P}_{error} = \mathbb{E}_Y[C(\delta(Y), \Theta] = R(\delta).$$

□

The result of this proposition says that since the probability of error is equal to the risk for the case of a uniform cost function, and since the Bayes' decision rule minimizes the risk, the Bayes' decision rule should also minimize the probability of error.

## 3.4 Binary Hypothesis Testing

We now discuss the binary hypothesis testing problem. To begin with, let us consider the general cost function. Denoting $C_{00}, C_{01}, C_{10}, C_{11}$ as the cost and $\pi_0, \pi_1$ as the prior, we can write Bayesian decision rule as

$$\delta_B(y) = \underset{i}{\operatorname{argmin}} \sum_{j=0}^{M-1} C_{ij} \pi_j(y).$$

Since there are only two possible choices of decisions (because it is a binary decision problem), we have

$$C_{00}\pi_0(y) + C_{01}\pi_1(y) \lessgtr_{H_1}^{H_0} C_{10}\pi_0(y) + C_{11}\pi_1(y).$$

With some simple calculations we can show that

$$\begin{aligned} & C_{00}\pi_0(y) + C_{01}\pi_1(y) && \lessgtr_{H_1}^{H_0} C_{10}\pi_0(y) + C_{11}\pi_1(y) \\ \Rightarrow\quad & \pi_1(y)(C_{01} - C_{11}) && \lessgtr_{H_1}^{H_0} \pi_0(y)(C_{10} - C_{00}) \\ \Rightarrow\quad & \tfrac{\pi_1(y)}{\pi_0(y)} && \lessgtr_{H_1}^{H_0} \tfrac{C_{10}-C_{00}}{C_{01}-C_{11}}, \end{aligned}$$

where the last inequality follows because $C_{00} < C_{10}$ and $C_{11} < C_{01}$. Since $\pi_j(y) = \frac{f_j(y)\pi_j}{\sum_j f_j(y)\pi_j}$, we have

$$\frac{f_1(y)}{f_0(y)} \lessgtr_{H_1}^{H_0} \frac{(C_{00} - C_{10})\pi_0}{(C_{11} - C_{01})\pi_1}.$$

If we define

$$L(y) \overset{\text{def}}{=} \frac{f_1(y)}{f_0(y)},$$

and

$$\eta \overset{\text{def}}{=} \frac{(C_{00} - C_{10})\pi_0}{(C_{11} - C_{01})\pi_1},$$

then the decision rule becomes

$$L(y) \lessgtr_{H_1}^{H_0} \eta.$$

The function $L(y)$ is called the likelihood ratio and the above decision rule the likelihood ration test (LRT).

---

**Example 3.**
Let two sample data drawn from two classes. The classes are described by two Gaussian distributions having equal variance but different means:

$$\begin{aligned} H_0 :& Y \sim N(0, \sigma^2) \\ H_1 :& Y \sim N(\mu, \sigma^2) \end{aligned}$$

To determine the Bayes' decision rule, we first compute the likelihood ratio

$$L(y) = \frac{e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2}}{e^{-\frac{1}{2}(\frac{y}{\sigma})^2}} = \exp\left(-\frac{2y\mu + \mu^2}{2\sigma^2}\right).$$

By taking log on both sides, we have

$$\ln L(y) \lessgtr_{H_1}^{H_0} \ln \frac{(C_{00} - C_{10})\pi_0}{(C_{11} - C_{01})\pi_1} \overset{\text{def}}{=} \ln \eta.$$

With some calculations we can show that this is equivalent to

$$y \lessgtr_{H_1}^{H_0} \frac{\sigma^2 \ln \eta}{\mu} + \frac{\mu}{2}.$$

Thus, if the observed value $y$ is larger that the right hand side of the above equation then must choose class $i = 1$. If not, we must choose class $i = 0$. Figure 3 illustrates an numerical example for the 1D and the 2D case.

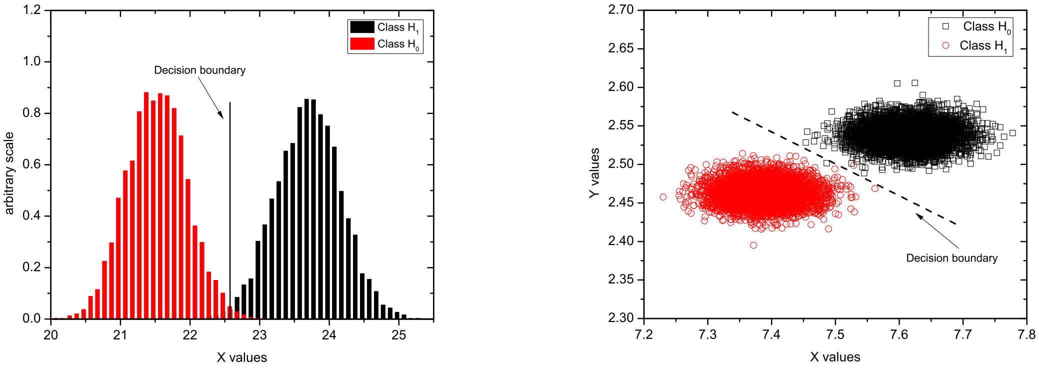


Figure 3: Decision boundaries for a binary hypothesis testing problem of 1D and 2D Gaussian.

---

The above general Bayes decision rule can be simplified when we assume a uniform cost. In this case, we have

$$\frac{f_1(y)}{f_0(y)} \lessgtr^{H_0}_{H_1} \frac{\pi_0}{\pi_1},$$

which is equivalent to

$$\pi_1(y) \lessgtr^{H_0}_{H_1} \pi_0(y).$$

Therefore, we will claim $H_0$ if $\pi_0(y) > \pi_1(y)$ and vice versa. Or equivalently, we have

$$\delta(y) = \operatorname*{argmax}_i \pi_i(y).$$

Since $\pi_i(y)$ is the posterior probability, we can the resulting decision rule as the maximum-a-posteriori decision.

---

**Example 4.**
Consider Example 3 with uniform cost. Then, the MAP decision rule is (with $\eta = 1$)

$$y < \frac{\mu}{2}.$$

---

### 3.5 M-ary Hypothesis Testing

We can generalize the above binary hypothesis testing problem to a $M$-ary hypothesis testing problem. In M-ary hypothesis testing, there are $M$ hypotheses or classes which we wish to assign our observations. The Bayesian decision rule is based again on minimizing the risk similarly to the binary case:

$$\delta_B(y) = \operatorname*{argmin}_i \sum_{j=0}^{M-1} C(i,j)\pi_j(y). \tag{33}$$

By Bayes' theorem, we have

$$\delta_B(y) = \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} C(i,j) \frac{\pi_j f_j(y)}{f(y)}.$$

Now, we can divide the posterior distribution by the posterior of $H_0$ without affecting the minimizer of the optimizaiton:

$$\begin{aligned}\delta_B(y) &= \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} C(i,j) \frac{\frac{\pi_j f_j(y)}{f(y)}}{\frac{\pi_0 f_0(y)}{f(y)}}, \\ &= \underset{i}{\operatorname{argmin}} \ \sum_{j=0}^{M-1} C(i,j) \frac{\pi_j f_j(y)}{\pi_0 f_0(y)}.\end{aligned}$$

By defining

$$L_j(y) \overset{\text{def}}{=} \frac{f_j(y)}{f_0(y)}, \quad \text{and} \quad h_i(y) = \sum_{j=0}^{M-1} C(i,j) \frac{\pi_j}{\pi_0} L_j(y),$$

we can show that

$$\delta_B(y) = \underset{i}{\operatorname{argmin}} \ \ h_i(y).$$

Therefore, our goal is to select hypothesis $H_0$ if $h_0 < h_1$ and $h_0 < h_2$ and $h_0 < h_3$ up to $h_0 < h_{M-1}$. To visualize this, let's consider three hypotheses described by three 1-d Gaussian distributions (Figure 4). In this case, there can be no single boundary as in the binary hypothesis testing, instead the observed value has to be compared with all individual hypotheses to reach a conclusion.

TO DO: Add an example to example M-ary hypothesis testing.

## 4 Numerical Example

In this example, we generate data from a random number generator. Our goal is to use Bayes' decision rule to classify the data into 2 classes. The m-file is provided in the appendix. Two sets of experiments were performed. The first one was based on equal priors for class 1 and class 2, $\pi_1 = \pi_2 = 0.5$. 1000 samples were drawn. The probability density functions were assumed to be Gaussian to represent the two classes but they differed in mean value. Two typical probability density distributions that were used and the sampled data are shown in Figure 5a. For different priors $\pi_1 = 0.9$ and $\pi_2 = 0.1$ the probability density distributions that were used and the sampled data are shown in Figure 5b. It can be seen that the sampled data for the second pdf are more scarce. However, the decision boundary can be easily drawn. Finally, two pdfs with different standard deviations is shown in Figure 6. In this case, there is significant overlap between the two distributions and the decision boundary is more complicated than before.

For simplicity, let's start the classification example with equal standard deviations for both Gaussians distributions, i.e., $\sigma_1 = \sigma_2 = 2.0$. For class 1, the average value was selected to be $m_1 = 1.0$ and was kept constant. For class 2, the average value $m_2$ varied from 4.0 to 20.0 with a step of 2. That is $m_2 = 4, 6, 8, 10, 12, 14, 16, 18, 20$ and therefore $m_2 - m_1 = 3, 5, 7, 9, 8, 11, 13, 15, 17, 19$. This experiment was performed 10 times and the number of misclassifications was measured as shown in the following Figure 7a. The priors were different in this case, $\pi_1 = 0.2$ and $\pi_2 = 0.8$.

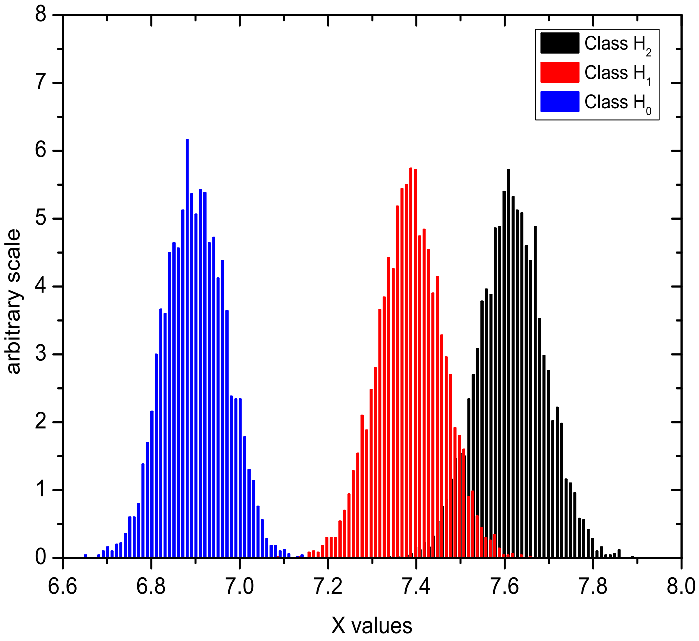


Figure 4: M=3 hypotheses

It appears that increasing the distance between the distributions, the error is reduced and from approximately 90 decreases to 0 for $m_2 - m_1 > 12$. Therefore, for completely separated classes the misclassifications are almost zero, as expected.

The second experiment considers equal priors, $\pi_1 = 0.5$ and $\pi_2 = 0.5$. The results are shown graphically in Figure 7b. In this case, the number of misclassifications is larger and approaches approximately 250 for mean different less than 4.

Overall, it was shown that for well separated classes the error of classification is very small and tends asymptotically to zero as the separation increases. On the contrary, for poor separation between the two classes the error was large and was approaching the priors. Finally, it is observed that when the priors and the probability density distributions are known, Bayes rule is an efficient and simple tool to provide classification decisions in an optimal way.

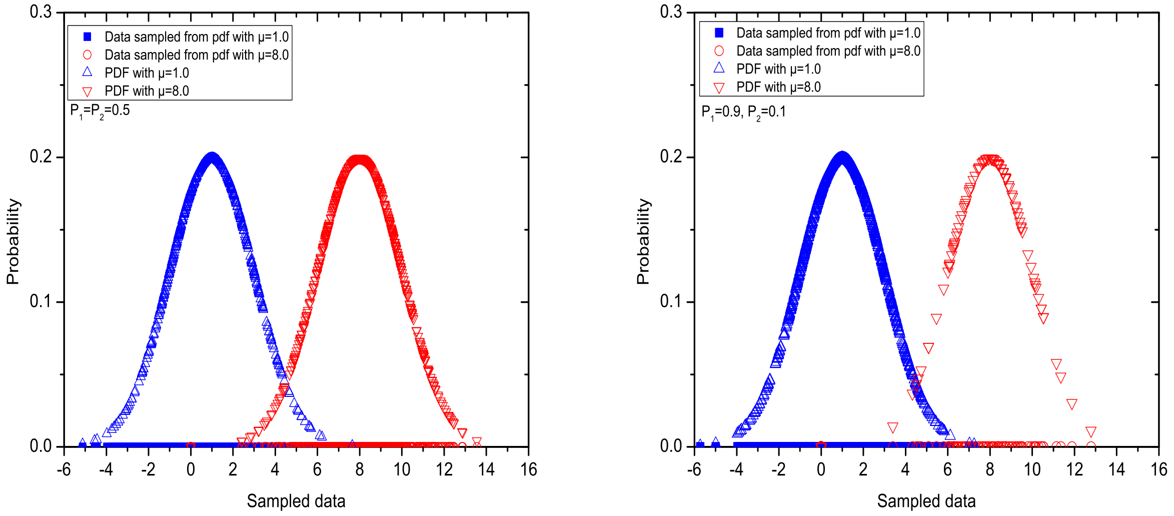
## 5 References

C.M. Bishop, 2006. Pattern Recognition and Machine Learning, Springer

R. O. Duda, P. E. Hart, and D. G. Stork. 2000, Pattern Classification (2nd Edition). Wiley-Interscience.

K. Fukunaga, 1990. Introduction to Statistical Pattern Recognition (2nd Edition). Academic Press Prof., Inc., San Diego, CA, USA.

A. Papoulis and S. U. Pillai, Probability, 2001. Random Variables and Stochastic Processes, McGraw-Hill.



(a) Equal priors

(b) Different priors

Figure 5: Gaussian distributions and sampled data

## 6 Appendix

Matlab code:

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
%
% This m-file calculates the number of misclassifications of a
% classification algorithm using Bayes rule. There are 100 data
% that belong to 2 different classes drawn from a random number generator.
% Bayes rule is used to classify the data. The experiment is repeated
% 10 times for each m2-m1=0,2,4,6,8 and priors P1=P2=0.5 (i.e., 50 experiments total)
% and 10 times for each m2-m1=0,2,4,6,8 and priors P1=0.2 and P2=0.8. m1
% and m2 are the mean values of the Gaussian distribution. Sigma1 and
% sigma2 are the standard deviations of the Gaussian distribution.
%
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
clc;
clear all;
P1=0.5;     % prior for class 1
P2=0.5;     % prior for class 2
m1=1.0;     % mean for normal distribution class 1
mu111=1.0071;
su111=3.9004;
sigma1=2.0; % standard deviation for normal distribution class 1
m2=16.0;     % mean for normal distribution class 2
sigma2=2.0; % standard deviation for normal distribution class 2
mu222=2.983;
su222=3.953;
N(1)=1000;
for n=1:1
for k=1:1    % increase number of data
for j=1:10    % perform experiment 10 times
```

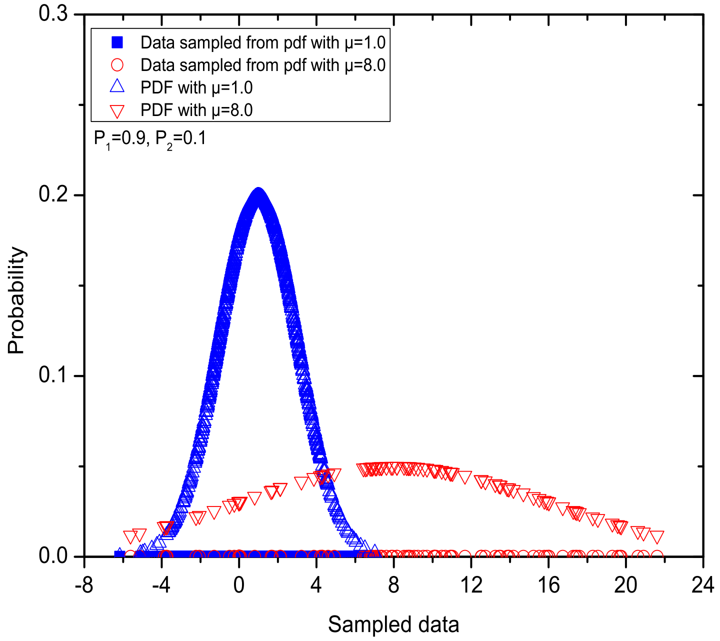
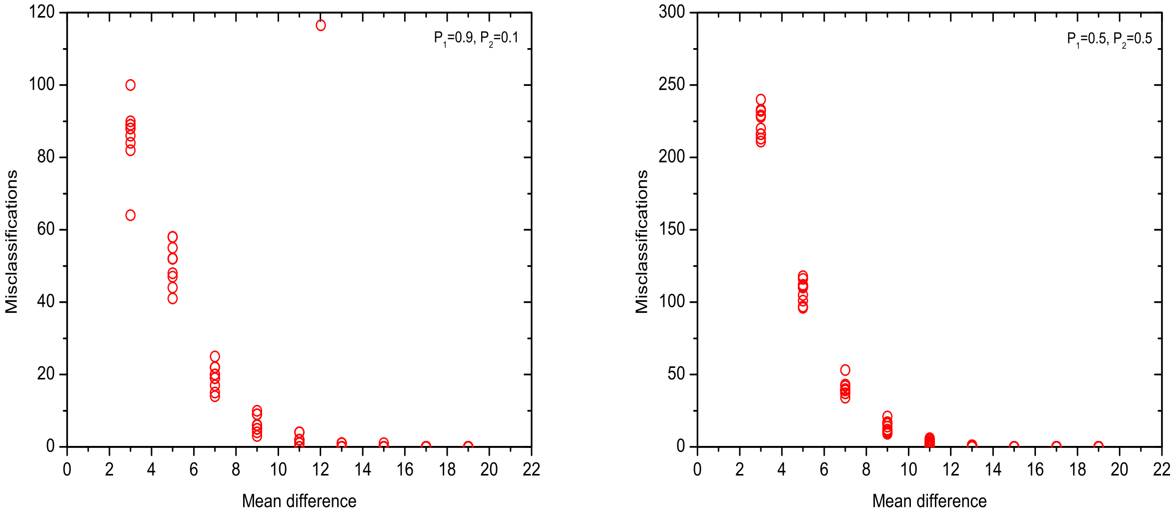


Figure 6: Gaussian distributions with different standard deviations and priors

```
for i=1:N(k)   % 100 randomly selected data
r=rand;
if r>=P2
    c(i)=1;             % class 1
    r1(i)=m1+sigma1*randn;
    r2(i)=r1(i);
    r3(i)=0;
else
    c(i)=2;             % class 2
    r1(i)=m2+sigma2*randn;
    r2(i)=0;
    r3(i)=r1(i);
end
g1(i)=pdf('Normal',r1(i),m1,sigma1)*P1;   %calculate probability
g2(i)=pdf('Normal',r1(i),m2,sigma2)*P2;   % calculate probability
g(i)=g1(i)-g2(i);                         % take the difference
if g(i)>=0                                % if g>0 then class 1
    y(i)=1;
else
    y(i)=2;
end
m(i)=abs(c(i)-y(i));                     % if c-y=0 then correct classifications
end
A=[c',y',m'];
mi(k,j)=nnz(m');  % returns the number nonzero elements of matrix k
end
mu(k)=mean(mi(k,j));     %calculate average of misclassifications for 10 experiments
sd(k)=std(mi(k,j));      % standard deviation of misclassifications for 10 experiments
mi(k,j);
```



(a) Different priors

(b) Equal priors

Figure 7: Misclassifications for equal and different priors

```
N(k+1)=N(k)+100;
end
Dm=m2-m1;   % difference of means
figure(3)
hold on
plot(Dm,mu(k),'o')
figure(4)
hold on
plot(Dm,mi(k,:),'o')
m2=m2+0;
end
pdf2=pdf('Normal',r2,m1,sigma1);
pdf3=pdf('Normal',r3,m2,sigma2);
figure(1)
plot(N(1:k),mi(1:k,:),'o')
%figure(2)
%plot(r2,zeros(N(1)),'ro',r3,zeros(N(1)),'bo',r2,pdf2,'go',r3,pdf3,'ko')
```