UNIT 3

The Problem of classification

The Problem of classification arises when an investigator makes a number of measurements on an individuals and wishes to classify the individual into one of several categories on the basis of these measurements. An individual is considered as a random observation from the population, the question is : ~~Given an individual is considered as a random observation from the population~~ Given an individual with certain measurements, from which population did the person arise?

This statistical technique developed for estimation, hypothesis testing and confidence statement based on exact specification of the response variate. In the applied source one kind of multivariate problem frequently occur inwhich on observations must be assign in some optimum fashion to one of several population.

This kind of multivariate technique is called as problem of classification.

For example, a banking officer may ~~which~~ ~~wh~~ wish to classify the loan application has low or high credit risk on the basis of the elements of certain accounting statements.

Standards of good classification

In constructing ~~the~~ a procedure of classification, it is desired to minimize the probability of misclassification. For convenience we shall now consider the case of only two categories.

Suppose an individual is an observation from either population π_1 or population π_2. The classification of an observation depends on the vector of measurements $X = (x_1, x_2, \ldots x_p)'$ on that individual.

we setup a rule that if an individual is characterised by certain sets of values of $x_1, x_2, \dots x_p$ that person will be classified as from Π_1, if other values as from Π_2.

we can think of an observation as a point in a p-dimensional space. we divide this space into two regions. If the observation falls in R_1, we classify it as comming from population Π_1, and if it falls in R_2, we classify it as comming ~~from~~ from population Π_2.

In the classification procedure, the statistician can make two kinds of errors in classification. If the individual is actually from Π_1, the statistician can classify him ~~or~~ as comming from population Π_2. If the individual is actually from Π_2, the statistician can classify him as comming from Π_1.

let the cost of the first type of misclassification be $c(\frac{2}{1})$ (>0) and let the cost of misclassifying ~~misclassification~~ an individual from Π_2 as from Π_1 be $c(1/2)$ (>0). These costs may be measured in any kind of units. There is no rewards for correct classification.

The following table indicates the costs of correct and incorrect classifications. clearly a good classification procedure is one that minimizes the cost of misclassification.

Statistician's Decision

		Π_1	Π_2
Population	Π_1	0	$c(2/1)$
	Π_2	$c(1/2)$	0

Baye's Procedure for minimizing expected loss of misclassification –

Two cases of two populations

Let the probability that the observation comes from population π_1 be q_1 and from π_2 be q_2 such that $q_1+q_2=1$.

Let the density of the population π_1 be $P_1(x)$ and for the population π_2 be $P_2(x)$.

If we have a region R_1 of classification as from π_1, the probability of correctly classifying an observation that actually is ~~drawn~~ drawn from population π_1 is

$$P(1/1, R) = \int_{R_1} P_1(x)\, dx \longrightarrow (1)$$

where $dx = dx_1 \ldots dx_p$

and the probability of misclassification of an observation from π_1 is

$$P(2/1, R) = \int_{R_2} P_1(x)\, dx \longrightarrow (2)$$

|||ly the probability of correctly classifying an observation from π_2 is

$$P(2/2, R) = \int_{R_2} P_2(x)\, dx \longrightarrow (3)$$

and the probability of misclassifying such an observation is

$$P(1/2, R) = \int_{R_1} P_2(x)\, dx \longrightarrow (4)$$

Since the probability of drawing an observation from π_1 is q_1, the probability of drawing an observation from π_1 and correctly classifying ~~as~~ it is

$$q_1 \cdot P(1/1, R)$$

|||ly the probability of drawing an observation from π_2 and correctly classifying it is $q_1\, P(2/1, R)$.

On the other hand, the probability of drawing an observation from Π_2 and correctly classifying it is $q_2 \, P(2/2, R)$

Similarly, the probability of drawing an observation from Π_2 and misclassifying it is $q_2 \cdot P(1/2, R)$.

Therefore, the expected loss of misclassification costs is the sum of the products of costs of misclassifications with their respective probabilities of occurence.

ie $\quad C(2/1) \; P(2/1, R) \, q_1 + C(1/2) \, P(1/2, R) \, q_2 \quad \longrightarrow (5)$

We wish to minimize this average loss, to divide our space into region R_1 and R_2 such that the expected loss is as small as possible. A procedure that minimizes equation (5) for given q_1 and q_2 is called a Baye's procedure.

Scanned by CamScanner

Procedure for classification ~~of~~ into one of two populations with known probability distributions

When a prior probabilities are known, we can define joint probabilities of the population and the observ~~ation~~ed set of variables.

The probability that an observation comes from the population π_1 and that each variate is less than the corresponding component in Y is

$$\int_{-\infty}^{y_p} \cdots \int_{-\infty}^{y_1} q_1\, p_1(x)\, dx_1 \cdots dx_p$$

The conditional probability ~~of~~ for the population comming from π_1 given an observation x is

$$\frac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)}.$$

The expected loss of misclassification is

$$C(2/1)\cdot P(2/1, R)\, q_1 + C(1/2)\, P(1/2, R)\, q_2 \longrightarrow (1)$$

Suppose $C(2/1) = C(1/2) = 1$ then the expected loss is

$$q_1\, P(2/1, R) + q_2\, P(1/2, R)$$

ie $$q_1 \int_{R_2} p_1(x)\, dx + q_2 \int_{R_1} p_2(x)\, dx \longrightarrow (2)$$

This is also the probability of misclassification.

For a given ~~observation~~ observed point x, we minimize the probability of misclassification by assigning the population that has the higher conditional probability.

If $\dfrac{q_1 p_1(x)}{q_1 p_1(x) + q_2 p_2(x)} \geq \dfrac{q_2 p_2(x)}{q_1 p_1(x) + q_2 p_2(x)}$ we choose population π_1

Otherwise we choose population π_2.

Since we minimize the probability of misclassification at each point, we minimize it over the whole space.

Thus the rule is,

$$\left.\begin{aligned} R_1 &: q_1 p_1(x) \geq q_2 p_2(x) \\ R_2 &: q_1 p_1(x) < q_2 p_2(x) \end{aligned}\right\} \longrightarrow (3)$$

If $P_r \left\{ \frac{p_1(x) = q_2}{p_2(x) = q_1} \Big/ \pi_i \right\} = 0, \quad i = 1, 2$

Then the Baye's procedure is unique except for sets of probability zero.

$\therefore$ choose R_1 and R_2 so as to minimize equation (2) the solution is equation (3).

We wish to minimize equation (1), which can be written as

$$c(2/1)\, q_1 \int_{R_2} p_1(x)\,dx + c(1/2)\, q_2 \int_{R_1} p_2(x)\,dx$$

We choose R_1 and R_2 according to

$$\left.\begin{aligned} R_1 &: c(2/1)\, q_1\, p_1(x) \geq c(1/2)\, q_2\, p_2(x) \\ R_2 &: c(2/1)\, q_1\, p_1(x) < c(1/2)\, q_2\, p_2(x) \end{aligned}\right\} \longrightarrow (4)$$

~~Since~~ Since, $c(2/1)\, q_1$ and $c(1/2)\, q_2$ are non negative constants.

Another way of writting equation (4) is

$$R_1 : \frac{p_1(x)}{p_2(x)} \geq \frac{c(1/2)\, q_2}{c(2/1)\, q_1}$$

$$R_2 : \frac{p_1(x)}{p_2(x)} < \frac{c(1/2)\, q_2}{c(2/1)\, q_1}$$

Classification into one of two known multivariate normal populations.

Let us consider in the case of two multivariate normal populations with equal covariance matrices ie $(\Sigma_1 = \Sigma_2 = \Sigma)$ namely $N(\mu^{(1)}, \Sigma)$ and $N(\mu^{(2)}, \Sigma)$ where $\mu^{(i)} = (\mu_1^{(i)}, \mu_2^{(i)} \dots \mu_p^{(i)})'$ is the vector of means of the i^{th} population. $i = 1,2$ and Σ is the matrix of variances and covariances of the each population. Then the i^{th} density is

$$P_i(x) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x-\mu^{(i)})' \Sigma^{-1} (x-\mu^{(i)})\right\} \rightarrow ①$$

The ratio of the densities ~~are~~ is

$$\frac{P_1(x)}{P_2(x)} = \frac{\exp\left\{-\frac{1}{2}(x-\mu^{(1)})' \Sigma^{-1} (x-\mu^{(1)})\right\}}{\exp\left\{-\frac{1}{2}(x-\mu^{(2)})' \Sigma^{-1} (x-\mu^{(2)})\right\}}$$

$$= \exp\left\{-\frac{1}{2}\left[(x-\mu^{(1)})' \Sigma^{-1} (x-\mu^{(1)}) - (x-\mu^{(2)})' \Sigma^{-1} (x-\mu^{(2)})\right]\right\}$$

$\rightarrow$ ②

Let R_1 is the region of classification into π_1 is the set of values x for which equation ② is greater than or equal to K (~~K~~ choose K arbitrarily) then the inequality can be written as (using log.)

$$-\frac{1}{2}\left[(x-\mu^{(1)})' \Sigma^{-1} (x-\mu^{(1)}) - (x-\mu^{(2)})' \Sigma^{-1} (x-\mu^{(2)})\right] \geq \log K$$

$\rightarrow$ ③

The LHS of ③ can be expressed as

$$-\frac{1}{2}\left[x'\Sigma^{-1}x - x'\Sigma^{-1}\mu^{(1)} - \mu^{(1)'}\Sigma^{-1}x + \mu^{(1)'}\Sigma^{-1}\mu^{(1)} - x'\Sigma^{-1}x + x'\Sigma^{-1}\mu^{(2)} + \mu^{(2)'}\Sigma^{-1}x - \mu^{(2)'}\Sigma^{-1}\mu^{(2)}\right]$$

$$= -\frac{1}{2}\left[-2x'\Sigma^{-1}\mu^{(1)} + 2x'\Sigma^{-1}\mu^{(2)}\right] - \frac{1}{2}\left[\mu^{(1)'}\Sigma^{-1}\mu^{(1)} - \mu^{(2)'}\Sigma^{-1}\mu^{(2)}\right]$$

$$= x'\Sigma^{-1}\mu^{(1)} - x'\Sigma^{-1}\mu^{(2)} - \frac{1}{2}\left[\mu^{(1)'}\Sigma^{-1}\mu^{(1)} - \mu^{(1)'}\Sigma^{-1}\mu^{(2)} + \mu^{(2)'}\Sigma^{-1}\mu^{(1)} - \mu^{(2)'}\Sigma^{-1}\mu^{(2)}\right]$$

$$= x'\Sigma^{-1}\mu^{(1)} - x'\Sigma^{-1}\mu^{(2)} - \frac{1}{2}\left[\mu^{(1)} + \mu^{(2)'}\Sigma^{-1}\mu^{(1)} - \mu^{(2)}\right]$$

$$= x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \quad \rightarrow (4)$$

The first term of equation (4) is well known discriminant function.

It is a linear function of the components of the observation vector.

If π_i has the density (1) $i = 1, 2$ the best region of classification are

$$R_1 : x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) \geq \log k$$

$$R_2 : x'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) - \frac{1}{2}(\mu^{(1)} + \mu^{(2)})'\Sigma^{-1}(\mu^{(1)} - \mu^{(2)}) < \log k$$

If a prior probabilities q_1 and q_2 are known then k is given by

$$k = \frac{q_2\, C(1/2)}{q_1\, C(2/1)}$$

In a particular case of the two population being equally likely and the cost being equal, $k=1$ ~~and k=0~~ $\log k = 0$. Then the region of classification are

$$R_1 : \quad x' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) \geq \frac{1}{2} (\mu^{(1)} + \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})$$

$$R_2 : \quad x' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)}) < \frac{1}{2} (\mu^{(1)} + \mu^{(2)})' \Sigma^{-1} (\mu^{(1)} - \mu^{(2)})$$

If we donot have a prior probabilities we may select $\log k = c$ (say) on the basis of the expected loss due to misclassification equal.

Classification ~~into~~ with Several Populations

Let $f_i(x)$ be the density associated with population π_i

$i = 1, 2 \ldots g$ $(g \geq 2)$.

Let p_i is the prior probability of population π_i

$c(k/i)$ is the cost of allocating an item to π_k

In fact it belongs to π_i for $k, i = 1, 2 \ldots g$

For $k = i$ $c(i/i) = 0$

Finally let R_k be the set of x's classified as π_k

and $P(k/i) = P(\text{classify item as } \pi_k/\pi_i) = \int_{R_k} f_i(x)\, dx$

for $k = i = 1, 2, \ldots g$ with $P(i/i) = 1 - \sum_{\substack{k=1 \\ k \neq i}}^{g} P(k/i)$

The conditional expected cost of misclassifying an x from π_1 into π_2 or $\pi_3 \ldots \pi_g$ is

$$ECM(1) = P(2/1)\, c(2/1) + P(3/1)\, c(3/1) + \cdots + P(g/1)\, c(g/1)$$

$$= \sum_{k=2}^{g} P(k/1) \cdot c(k/1)$$

This conditional expected cost occurs with prior probability p_1, the population of π_1.

In a similar manner, we can obtain the conditional expected costs of misclassification,

$ECM(2), \ldots ECM(g)$.

Multiplying each conditional ECM by its prior probability and ~~Sum~~ summing gives the overall ECM.

$$\therefore ECM = p_1\, ECM(1) + p_2\, ECM(2) + \cdots + p_g\, ECM(g)$$

$$= p_1\left[\sum_{k=2}^{g} P(k/1)\, c(k/1)\right] + p_2\left[\sum_{\substack{k=1\\k\neq 2}}^{g} P(k/2)\, c(k/2)\right] + \cdots$$

$$+\, p_g\left[\sum_{k=1}^{g-1} P(k/g)\, c(k/g)\right]$$

$$\therefore\ ECM = \sum_{i=1}^{g} p_i\left[\sum_{\substack{k=1\\k\neq i}}^{g} P(k/i)\, c(k/i)\right] \longrightarrow (1)$$

Determining an optimal classification procedure amounts to choosing the mutually exclusive and exhaustive classification region $R_1, R_2, \ldots R_g$ such that equation (1) is a minimum.

The classification regions that minimize the ECM of equation (1) are defined ~~as follows~~ by allocating x to that population π_k : $k = 1, 2 \ldots g$ for which

$$\sum_{\substack{i=1\\k\neq i}}^{g} p_i\, f_i(x)\, c(k/i) \longrightarrow (2)$$

is smallest.

Suppose all the misclassification cost are equal, in which case the minimum expected cost of misclassification rule is the minimum total probability of misclassification.

Using the argument leading to equation (2) we would allocate x to that population π_k ; where $k = 1, 2 \ldots g$

for which $\sum_{i=1}^{g} p_i f_i(x)$ ⟶ (3) is minimum.

when the misclassification cost are same, the minimum ECM rule has the following form

Allocate x to π_k if

$$p_k f_k(x) > p_i f_i(x) \quad \text{for all } i \neq k$$

or ~~equa~~ equivalently,

Allocate x to π_k if

$$\ln\left[p_k f_k(x)\right] > \ln\left[p_i f_i(x)\right] \quad \text{for all } i \neq k$$

Classification into one of Several Multivariate Normal Population.

we Know that

$$f_i(x) = \frac{1}{(2\pi)^{P/2} |\Sigma_i|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu_i)' \Sigma_i^{-1} (x-\mu_i)\right\} \quad i=1,2\dots g \quad \longrightarrow ①$$

are multivariate normal densities with mean vector μ_i and Covariance matrices Σ_i.

If $c(i/i) = 0$ and $c(k/i) = 1$ for $k \neq i$

the minimum expected cost misclassification (ECM) rule becomes,

allocate x to π_k if

$$P_k f_x(x) = P_k (2\pi)^{-P/2} |\Sigma_k|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}(x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)\right\}$$

Taking log on both sides

$$\log\left[P_k f_x(x)\right] = \log P_k - \frac{P}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(x-\mu_k)' \Sigma_k^{-1}(x-\mu_k)$$

$$= \max_i \log P_i f_i(x) \quad \longrightarrow ②$$

Here the Constant term $\frac{P}{2}\log(2\pi)$ Can be ignored in ② Since it is fixed for all the ~~terms~~ populations

$\therefore$ we define the Quadratic discrimination Score for the i^{th} population

$$d_i^Q(x) = -\frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)' \Sigma_i^{-1}(x-\mu_i) + \log P_i \quad i=1,2\dots g \quad \longrightarrow ③$$

The quadratic Score $d_i^Q(x)$ is Composed of Contributions from the generalised variance $|\Sigma_i|$, the Prior Probability P_i and the squared distance from the X to the population μ_i.

using discriminant Score to the classification rule in the equation (2), ~~it becomes~~

Allocate X to π_K if

$$dQ_K(x) = \text{maximum of } dQ_1(x),\ dQ_2(x) \ldots dQ_g(x) \longrightarrow (4)$$

where $dQ_K(x)$ is given in equation (3),

In Particular, μ_i and Σ_i are unknown we use the relevant estimates $\bar{x}_i$ and S_i then (3) becomes

$$dQ_i(x) = -\frac{1}{2}\log|S_i| - \frac{1}{2}(x-\bar{x}_i)' S_i^{-1}(x-\bar{x}_i) + \log P_i \longrightarrow (5)$$

If $\Sigma_i = \Sigma$

$$dQ_i(x) = -\frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_i)'\Sigma^{-1}(x-\mu_i) + \log P_i$$

$$= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x + \frac{1}{2}x'\Sigma^{-1}\mu_i + \frac{1}{2}\mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \log P_i$$

$$= -\frac{1}{2}\log|\Sigma| - \frac{1}{2}x'\Sigma^{-1}x + \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \log P_i$$

The first two terms are ignored, since they are common for $dQ_1(x), dQ_2(x) \ldots dQ_g(x)$.

$\therefore$ The linear discriminant score

$$dQ_i(x) = \mu_i'\Sigma^{-1}x - \frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \log P_i \longrightarrow (6)$$

$\therefore$ The classification rule becomes, allocate x to π_K if

$$dQ_K(x) = \text{Max}\{dQ_1(x), dQ_2(x) \ldots dQ_g(x)\}$$

where $dQ_i(x)$ is given in (6) for $i = 1, 2, \ldots g$

Discriminant Analysis

Discriminant analysis and classification are the multivariate techniques concerned with separating distinct set of objects or observations and with allocating new objects to previously defined groups.

Discriminant analysis is rather exploratory in nature. As a separatory procedure, it is often exployed on a one-time basis inorder to investigate observed differences when caused relationships are not well understood.

The basic idea of discriminant analysis consist of assigning an individual or a group of individual to one or several known or ~~unknown~~ unknown distinct population on the basis of the observations on several characteristics of the individual or the group.

In scientific literature, discriminant analysis has many synonyms such as classification, identification, prediction and selection depending on the types of scientific area inwhich is used.

Thus the immediate goal of discriminant analysis is to describe either graphically or algebraically, the differential features of objectives from several known collections. We try to find 'discriminants' whose numerical values are such that the collections are separated as much as possible.

Fisher Discriminant function – Separation of Population

Fisher idea was to transform the multivariate observations X to univariate observation Y such that the Y's derived from population Π_1 and Π_2 were separated as much as possible.

If we let μ_{1Y} be the mean of Y obtained from X belonging to Π_1 and μ_{2Y} be the mean of Y obtained from X belonging to Π_2.

Fisher selected the linear combination $Y = l'X$, to maximize the distance between μ_{1Y} and μ_{2Y}. ~~by~~ ~~defining~~ by defining

$$\mu_1 = E(X/\Pi_1) = \text{Expected value of multivariate observation from } \Pi_1$$

$$\mu_2 = E(X/\Pi_2) = \text{Expected value of multivariate observation from } \Pi_2$$

and the Covariance matrix

$$\Sigma = E\left[(X-\mu_i)(X-\mu_i)'\right] \; ; \; i = 1, 2$$

is same for the populations Π_1 and Π_2

Consider the linear Combination

$$Y = l'X$$
$$= (\mu_1 - \mu_2)' \Sigma^{-1} X$$
$$= (\bar{X}_1 - \bar{X}_2)' S^{-1} X$$

using this transformation, Y has a mean

$$\mu_{1Y} = E(Y/\pi_1) = E(l'x/\pi_1) = l'\mu_1$$

$$\mu_{2Y} = E(Y/\pi_2) = E(l'x/\pi_2) = l'\mu_2$$

and its variances are same,

ie, $$\sigma^2_Y = var(l'x)$$

$$= l' \, covar(x) \, l \qquad \{\because x \text{ is a vector}$$

$$= l' \Sigma l$$

The best linear combination $Y = l'x$, maximizes the ratio

$$\frac{\text{square distance between mean of } Y}{\text{variance of } Y} = \frac{(\mu_{1Y} - \mu_{2Y})^2}{\sigma_Y^2}$$

$$= \frac{(l'\mu_1 - l'\mu_2)^2}{l'\Sigma l}$$

$$= \frac{l'.(\mu_1 - \mu_2)(\mu_1 - \mu_2)' \, l}{l'\Sigma l}$$

$$= \cancel{\frac{[l'(\mu_1-\mu_2)][l'(\mu_1-\mu_2)]'}{l'\Sigma l}}$$

$$= \frac{(l' s)^2}{l'\Sigma l}$$

where $S = \mu_1 - \mu_2$ and $l' = (l_1 \; l_2 \dots l_p)$ is the fisher linear combinations of co-efficients.

Maximizing the ratio, fisher introduce the linear Combination $Y = (\mu_1 - \mu_2)' \Sigma^{-1} x$ which is known as fisher linear discriminant function.

let $Y_0 = (\mu_1 - \mu_2)' \Sigma^{-1} x$ be the value of the discriminant function for a new observation x_0 and let

$m = \frac{1}{2} (\mu_1 - \mu_2)' \Sigma^{-1} (\mu_1 - \mu_2)$ be the mid point between the two population mean. Therefore the classification rule is,

allocate x_0 to π_1 if $Y_0 = (\mu_1 - \mu_2)' \Sigma^{-1} x_0 \geq m$

allocate x_0 to π_2, if $Y_0 = (\mu_1 - \mu_2)' \Sigma^{-1} x_0 < m$

alternatively, subtract m from Y_0 and compare the results with zero, in such case, the rule becomes,

Allocate x_0 to π_1 if $Y_0 - m \geq 0$

allocate x_0 to π_2 if $Y_0 - m < 0$