~~Unit - 2~~ Unit-2

Mahalanobis D square

: Mahalanobis distance is the distance between two points in multivariate space In regular Euclidean space variables (x, y, z) are represented by axes drawn at right angles to each other The distance between any two can be measured with a ruler for uncorrelated variables. Euclidean distance equals the MD However, if two or more variables are correlated, the axes are no longer at right angle and the measurement with becomes impossible with a ruler In addition, If you have more than three variables you can't plot them in regular 3D space at all MD solves it

this measurement problem, as it measures distances between points, even correlated points for multiple variables.

Mahalanobis distance

The mahalanobis distance measures distance relative to centroid - a base or central point which can be taught of as an overall mean for multivariate data. The centroid is a point in multivariate space where all means from all variables intersect. The larger the MD, the further away from the centroid the data point is.

Application

MD is widely used in cluster analysis and classification techniques. It is closely related to T- square ~~distance~~ distribution used for multivariate statistical testing and Fisher's linear discriminant analysis that is used for supervised classification.

MD is

In order to use the Mahalanobis distance to classify a test point as belonging to one of N classes, one first estimates the covariance matrix of each class, usually based on samples known to belong to each class. Then, given a test sample, one computes the Mahalanobis distance to each class and classifies the test point as belonging to that class for which the Mahalanobis distance is minimal.

Mahalanobis distance and leverage are often used to detect outliers especially in the development of linear regression models. A point that has a greater Mahalanobis distance from the rest of the sample population of points is said to have higher leverage since it has a ~~higher~~ greater influence on the slope or coefficients of the regression equation

MD is also used to determine multivariate outliers. Regression techniques can be used to determine if a specific case within a sample population is an outlier via the combination of two or more variable scores. Even for normal distributions, a point can be a multivariate outlier even if it is not a univariate outlier for any variables

Formal definition

The MD between two objects is defined as

$$D^2 (\text{Mahalanobis}) = [(x_A - x_B)^T * C^{-1} * (x_A - x_B)]$$

where x_A and x_B are a pair of objects

C is the sample covariance matrix.

Another version of the formula, which uses distances from each observation to the centroid mean

$$d_i = [(x_i - \bar{x})^T (C^{-1}) (x_i - \bar{x})]^{0.5}$$

where

x = an object vector

$\bar{x}$ = A.M vector

C = sample covariance matrix

Disadvantages

Although MD is included with many popular stat packages, some authors question the reliability of results.

A major issue with the MD is that the inverse of the correlation matrix is needed for the calculations. This can't be calculated if the variables are highly correlated.

Generalised T^2 statistic distribution.

~~For univariate distribution~~,

Recalling the univariate theory for determining whether a specific value μ_0 is a plausible value for the population mean μ.

Null hypothesis: $\mu = \mu_0$ Alternative hypothesis $\mu \neq \mu_0$

If $x_1, x_2 \ldots x_n$ denote random sample from normal population then the test statistic is

$$t = \frac{(\bar{x} - \mu_0)}{s/\sqrt{n}} \quad \text{where } \bar{x} = \frac{\Sigma x_i}{n} \quad s^2 = \frac{1}{n-1}\Sigma(x_i - \bar{x})^2$$

$t \sim t$ distribution with $n-1$ d.f

If $|t| > t_{n-1}(\alpha/2)$, reject H_0.

Rejecting H_0 when $|t|$ is large is equivalent to rejecting H_0 if its square $t^2 = \frac{(\bar{x}-\mu_0)^2}{s^2/n} = n(\bar{x}-\mu_0)(s^2)^{-1}(\bar{x}-\mu_0)$ —— (1)

is large. The variable t^2 is a square of the distance from the sample mean $\bar{x}$ to test value μ_0. The units of distance are expressed in terms of $s/\sqrt{n}$.

Reject H_0 at significance level α, if

$$n(\bar{x}-\mu_0)(s^2)^{-1}(\bar{x}-\mu_0) > t_{n-1}(\alpha/2) \quad \text{——} \quad (2)$$

where $t_{n-1}(\alpha/2)$ denotes the upper $100(\alpha/2)$th percentile of the t-distribution with $n-1$ d.f.

{Accept $H_0: \mu = \mu_0$ at level α} (or) $\left|\frac{\bar{x}-\mu_0}{s/\sqrt{n}}\right| \le t_{n-1}(\alpha/2)$

is equivalent to

{μ_0 lies in the $100(1-\alpha)$ % confidence interval $\bar{x} \pm t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}}$}

(or)

$$\bar{x} - t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \le \mu_0 \le \bar{x} + t_{n-1}(\alpha/2)\frac{s}{\sqrt{n}} \quad \text{——} \quad (3)$$

Confidence interval in (3) is a random interval because the end points depends upon the random variable $\bar{x}$ and s.

A natural generalization of the squared distance in (1) is its multivariate analog

$$T^2 = (\bar{x} - \mu_0)' \left(\frac{S}{n}\right)^{-1} (\bar{x} - \mu_0) = n(\bar{x} - \mu_0)' S^{-1} (\bar{x} - \mu_0) \quad (4)$$

$$\underset{p\times 1}{\bar{x}} = \frac{\sum_{j=1}^{n} x_j}{n} \qquad \underset{p\times p}{S} = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \bar{x})(x_j - \bar{x})' \qquad \underset{p\times 1}{\mu_0} = \begin{bmatrix} \mu_{10} \\ \mu_{20} \\ \vdots \\ \mu_{p0} \end{bmatrix}$$

The statistic T^2 is called Hotelling T^2.

$(1/n)S$ is the estimated covariance matrix of $\bar{x}$.

If T^2 is too large, H_0 is rejected

T^2 is distributed as $\frac{(n-1)p}{n-p} F_{p,n-p}$

$$\bar{x} = \frac{\sum x_j}{n} \qquad S = \frac{1}{n-1} \sum_{j} (x_j - \bar{x})(x_j - \bar{x})'$$

$$\alpha = P\left[T^2 > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right]$$

$$\alpha = P\left[n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)\right]$$

Reject H_0 if $n(\bar{x} - \mu)' S^{-1} (\bar{x} - \mu) > \frac{(n-1)p}{n-p} F_{p,n-p}(\alpha)$

$$T^2 = \sqrt{n}(\bar{x} - \mu_0)' \left(\frac{\sum_j (x_j - \bar{x})(x_j - \bar{x})'}{n-1} \right) \sqrt{n}(\bar{x} - \mu_0)$$

which contains normal $N(0, \Sigma)$ random vector and

Wishard $W_{p, n-1}(\Sigma)$ random matrix in the form

$$T^2_{p,n-1} = \left(\begin{matrix} \text{multivariate normal} \\ \text{random vector} \end{matrix} \right) \left(\frac{\text{Wishard random matrix}}{df} \right) \left(\begin{matrix} \text{multivariate normal} \\ \text{random vector} \end{matrix} \right)$$

Then a analogous is

$$t^2 = \sqrt{n}(\bar{x} - \mu_0)(s^2)^{-1} \sqrt{n}(\bar{x} - \mu_0)$$

$$t^2_{n-1} = \left(\begin{matrix} \text{normal} \\ \text{random vector} \end{matrix} \right) \left(\frac{\text{chi-squared}}{df} \right)^{-1} \left(\begin{matrix} \text{normal} \\ \text{random variable} \end{matrix} \right)$$

for univariate are

Unit 2:- Sampling distribution of Covariance matrix

The sample covariance matrix, $S = \frac{1}{N-1} \sum_{\alpha=1}^{n} (x_\alpha - \bar{x})(x_\alpha - \bar{x})'$,

is an unbiased estimator of the popln covariance matrix Σ. This result will be generalised to the case of a matrix A of any order when $\Sigma = I$, this distribution is in a sense a generalisation of the χ^2-distribution. The distribution of A (or S), often called the wishard distribution, is fundamental to multivariate statistical analysis

The tentative assumption that $x_1, x_2 \ldots x_n$ constitute a random sample from a normal popln with mean μ and covariance Σ completely determines the sampling distributions of $\bar{x}$ and S. Here we present the results on the sampling distn of $\bar{x}$ and S by drawing a parallel with the familiar univariate conclusions

In the univariate case ($p=1$), W.K.T $\bar{x}$ is normal with mean μ = popln mean and variance

$\frac{1}{n}\sigma^2$ = $\frac{\text{Population variance}}{\text{Sample size}}$

The result for the multivariate case ($p \geq 2$) is analogous in that $\bar{x}$ has a normal distn with mean μ and covariance matrix $\left(\frac{1}{n}\right)\Sigma$

For the sample variance, recall that ~~(n-1)~~

$(n-1)s^2 = \sum_{j=1}^{n} (x_j - \bar{x})^2$ is distributed as σ^2 times a chi-square variable having $(n-1)$ degrees of freedom. In turn, this chi-square is the distribution of a sum of squares of independent standard normal random variables.

That is $(n-1)s^2$ is distributed as

$\sigma^2(z_1^2 + \dots + z_{n-1}^2) = (\sigma z_1)^2 + (\sigma z_2)^2 + \dots + (\sigma z_{n-1})^2$

The individual term σz_i are independently distributed as $N(0, \sigma^2)$

Wishard distribution

The sampling distribution of the sample covariance matrix is called the wishard distribution, after its discoverer, It is defined as the sum of independent products of multivariate normal random vectors. specifically,

$W_m(\cdot|\Sigma)$ = wishard distribution with m d.f

= distribution of $\sum_{j=1}^{m} z_j z_j'$

where z_j are each independently distributed as $N_p(0, \Sigma)$

We summarise the sampling distn results as follows:-

Let $x_1, x_2, \dots x_n$ be a random sample of size n from a p-variate normal distn with mean μ and covariance matrix Σ.

Then

1) $\bar{x}$ is distributed $N_p(\mu, (1/n)\Sigma)$

2) $(n-1)S$ is distributed as a wishard random matrix with $(n-1)$ df

3) $\bar{x}$ and S are independent

* Because Σ is unknown, the dists of $\bar{x}$ cannot be used directly to make inferences about μ. However, S provides independent information about Σ, and the dists of S does not depend on μ. This allows us to construct a statistic for making inferences about μ

Properties of the wishard distribution

1) If A_1 is distributed as $W_{m_1}(A_1/\Sigma)$ independently of A_2, which is distributed as $W_{m_2}(A_2/\Sigma)$, then A_1+A_2 is distributed as $W_{m_1+m_2}(A_1+A_2/\Sigma)$ That is, the degrees of freedom add.

2) If A is distributed as $W_m(A/\Sigma)$, then CAC' is distributed as $W_m(CAC'/C\Sigma C')$

The probability density function of the wishard distribution, the sample size n is greater than the no. of variables p, then

$$|A|^{(n-p-2)/2} \quad e^{-tr[A\Sigma^{-1}]/2}$$

$$W_{p-1}(A/\Sigma) = \frac{}{2^{p(n-1)/2} \quad \pi^{p(p-1)/4} \quad |\Sigma|^{(n-1)/2} \prod_{i=1}^{p} \Gamma\left(\frac{1}{2}(n-i)\right)}$$

A positive definite where $\Gamma(\cdot)$ is the gamma function.

Uses and Application of T^2

1) Hotelling T-square statistic allows for the testing of hypotheses on multiple (often correlated) measures within the same sample.
2) A generalization of student's t-statistic, called Hotelling T^2 statistic, allows for the testing of hypotheses on multiple (often correlated) measures within the same sample.
3) Hotelling's T-squared has several advantages over the t-test
4) The type I error rate is well controlled,
5) The relationship b/w multiple variables is taken into account.
6) It can generate an overall conclusion even if multiple (single) t-tests are inconsistent. While a t-test will test you which variable differ b/w groups, Hotelling's summarizes the b/w-group differences.
7) From the Hotelling's T-squared test it was concluded that there is significant difference in the average performance of boarding and day students. On the other hand, the 2-way ANOVA indicates a significant difference in the performance of students based on subjects and then gives no evidence to conclude that there is significant difference b/w the years under study

Case I : Inference problems regarding 2 means from a paired popln

Case II : Inference problems regarding a single multivariate popln

Case III : Inference problems regarding 2 means from 2 independent popln

All three cases make use of the Hotelling T-square statistic

8) In statistics, particularly in hypothesis testing, the Hotelling T-squared distribution (T^2), proposed by Harold Hotelling, is a multivariate prob distn that is tightly related to the F-distn and is most notable for arising as the distn of a set of sample statistics that are natural generalizations of the statistics underlying the student's t-distn.

a) The Hotelling's T-square test plays the same role as the t-test when there is more than one dependent variable It is a special case of MANOVA but not of ANOVA

One sample hypothesis testing of the mean

2 sample hypothesis testing of the mean with paired samples

2 - sample hypothesis testing of the mean with independent samples

2 - sample hypothesis testing of means with unequal covariance matrices

~~2 sample hypothesis~~

Real statistics capabilities

statistical power and sample size requirements

Hotelling's T-squared is based on Hotelling's T^2 distn and forms the basis for various multivariate control charts