Mahalanobis distance

From Wikipedia, the free encyclopedia

The **Mahalanobis distance** is a measure of the distance between a point P and a distribution D, introduced by P. C. Mahalanobis in 1936.[1] It is a multi-dimensional generalization of the idea of measuring how many standard deviations away P is from the mean of D. This distance is zero if P is at the mean of D, and grows as P moves away from the mean: along each principal component axis, it measures the number of standard deviations from P to the mean of D. If each of these axes is rescaled to have unit variance, then Mahalanobis distance corresponds to standard Euclidean distance in the transformed space. Mahalanobis distance is thus unitless and scale-invariant, and takes into account the correlations of the data set.

Contents

Definition and properties

The Mahalanobis distance of an observation $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^T$ from a set of observations with mean $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^T$ and covariance matrix S is defined as:

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T S^{-1} (\vec{x} - \vec{\mu})}.$$ [2]

Mahalanobis distance (or "generalized squared interpoint distance" for its squared value[3]) can also be defined as a dissimilarity measure between two random vectors $\underline{\boldsymbol{x}}$ and $\underline{\boldsymbol{y}}$ of the same distribution with the covariance matrix S:

$$d(\vec{x},\vec{y}) = \sqrt{(\vec{x}-\vec{y})^T S^{-1}(\vec{x}-\vec{y})}.$$

If the covariance matrix is the identity matrix, the Mahalanobis distance reduces to the Euclidean distance. If the covariance matrix is diagonal, then the resulting distance measure is called a *normalized Euclidean distance*:

$$d(\vec{x},\vec{y}) = \sqrt{\sum_{i=1}^{N} \frac{(x_i - y_i)^2}{s_i^2}},$$

where s_i is the standard deviation of the x_i and y_i over the sample set.

Mahalanobis distance is preserved under full-rank linear transformations of the space spanned by the data. This means that if the data has a nontrivial nullspace, Mahalanobis distance can be computed after projecting the data (non-degenerately) down onto any space of the appropriate dimension for the data.

Intuitive explanation

Consider the problem of estimating the probability that a test point in N-dimensional Euclidean space belongs to a set, where we are given sample points that definitely belong to that set. Our first step would be to find the average or center of mass of the sample points. Intuitively, the closer the point in question is to this center of mass, the more likely it is to belong to the set.

However, we also need to know if the set is spread out over a large range or a small range, so that we can decide whether a given distance from the center is noteworthy or not. The simplistic approach is to estimate the standard deviation of the distances of the sample points from the center of mass. If the distance between the test point and

the center of mass is less than one standard deviation, then we might conclude that it is highly probable that the test point belongs to the set. The further away it is, the more likely that the test point should not be classified as belonging to the set.

This intuitive approach can be made quantitative by defining the normalized distance between the test point and the set to be $\frac{x - \mu}{\sigma}$. By plugging this into the normal distribution we can derive the probability of the test point belonging to the set.

The drawback of the above approach was that we assumed that the sample points are distributed about the center of mass in a spherical manner. Were the distribution to be decidedly non-spherical, for instance ellipsoidal, then we would expect the probability of the test point belonging to the set to depend not only on the distance from the center of mass, but also on the direction. In those directions where the ellipsoid has a short axis the test point must be closer, while in those where the axis is long the test point can be further away from the center.

Putting this on a mathematical basis, the ellipsoid that best represents the set's probability distribution can be estimated by building the covariance matrix of the samples. The Mahalanobis distance is the distance of the test point from the center of mass divided by the width of the ellipsoid in the direction of the test point.

Normal distributions

For a normal distribution in any number of dimensions, the probability of an observation is uniquely determined by the Mahalanobis distance d. Specifically, d^2 is chi-squared distributed. If the number of dimensions is 2, for example, the probability of a particular calculated d being less than some threshold t is $1 - e^{-t^2/2}$. To determine a threshold to achieve a particular probability, p, use $t = \sqrt{-2\ln(1-p)}$,

for 2 dimensions. For number of dimensions other than 2, the cumulative chi-squared distribution should be consulted.

In a normal distribution, the region where the Mahalanobis distance is less than one (i.e. the region inside the ellipsoid at distance one) is exactly the region where the probability distribution is concave.

Mahalanobis distance is proportional, for a normal distribution, to the square root of the negative log likelihood (after adding a constant so the minimum is at zero).

Relationship to normal random variables

In general, given a normal (Gaussian) random variable X with variance $S = 1$ and mean $\mu = 0$, any other normal random variable R (with mean μ_1 and variance S_1) can be defined in terms of X by the equation $R = \mu_1 + \sqrt{S_1} X$. Conversely, to recover a normalized random variable from any normal random variable, one can typically solve for $X = (R - \mu_1)/\sqrt{S_1}$. If we square both sides, and take the square-root, we will get an equation for a metric that looks a lot like the Mahalanobis distance:

$$D = \sqrt{X^2} = \sqrt{(R - \mu_1)^2 / S_1} = \sqrt{(R - \mu_1) S_1^{-1} (R - \mu_1)}.$$

The resulting magnitude is always non-negative and varies with the distance of the data from the mean, attributes that are convenient when trying to define a model for the data.

Relationship to leverage

Mahalanobis distance is closely related to the leverage statistic, h, but has a different scale:[4]

$$D^2 = (N - 1)(h - \tfrac{1}{N}).$$

Applications

Mahalanobis's definition was prompted by the problem of identifying the similarities of skulls based on measurements in 1927.[5]

Mahalanobis distance is widely used in cluster analysis and classification techniques. It is closely related to Hotelling's T-square distribution used for multivariate statistical testing and Fisher's Linear Discriminant Analysis that is used for supervised classification.[6]

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 greater influence on the slope or coefficients of the regression equation. Mahalanobis distance 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 variable (consider a probability density concentrated along the line $x_1 = x_2$, for example), making Mahalanobis distance a more sensitive measure than checking dimensions individually.

See also

- Bregman divergence (the Mahalanobis distance is an example of a Bregman divergence)
- Bhattacharyya distance related, for measuring similarity between data sets (and not between a point and a data set)

- Hamming distance identifies the difference bit by bit of two strings
- Hellinger distance, also a measure of distance between data sets
- Similarity learning, for other approaches to learn a distance metric from examples.

References

1. Mahalanobis, Prasanta Chandra (1936). "On the generalised distance in statistics" (PDF). *Proceedings of the National Institute of Sciences of India*. **2** (1): 49–55. Retrieved 2016-09-27.
2. De Maesschalck, Roy; Jouan-Rimbaud, Delphine; and Massart, Désiré L. (2000); *The Mahalanobis distance*, Chemometrics and Intelligent Laboratory Systems 50:1–18
3. Gnanadesikan, Ramanathan; and Kettenring, John R. (1972); *Robust estimates, residuals, and outlier detection with multiresponse data*, Biometrics 28:81–124
4. Schinka, John A.; Velicer, Wayne F.; and Weiner, Irving B. (2003); *Handbook of psychology: Research methods in psychology*, John Wiley and Sons
5. Mahalanobis, Prasanta Chandra (1927); *Analysis of race mixture in Bengal*, Journal and Proceedings of the Asiatic Society of Bengal, 23:301–333
6. McLachlan, Geoffrey J. (1992); *Discriminant Analysis and Statistical Pattern Recognition*, Wiley Interscience, p. 12. ISBN 0-471-69115-1

External links

- Hazewinkel, Michiel, ed. (2001), "Mahalanobis distance", *Encyclopedia of Mathematics*, Springer, ISBN 978-1-55608-010-4
- Mahalanobis distance tutorial (http://people.revoledu.com/kardi/tutorial/Similarity/MahalanobisDistance.html) – interactive online program and spreadsheet computation
- Mahalanobis distance (Nov-17-2006) (http://matlabdatamining.blogspot.com/2006/11/mahalanobis-distance.html) – overview of Mahalanobis distance, including MATLAB code
- What is Mahalanobis distance? (http://blogs.sas.com/content/iml/2012/02/15/what-is-mahalanobis-distance/) – intuitive, illustrated explanation, from Rick Wicklin on blogs.sas.com

Categories: Statistical distance | Multivariate statistics
