马哈拉诺比斯距离

wikipedia

2 August 2021

马哈拉诺比斯距离 (Mahalanobis distance) 是一种衡量点 P 和概率分布 D 之间距离的方法，由 P. C. Mahalanobis 在 1936 年提出 [1]。它是对衡量 P 距离 D 的均值有多少标准差的概念的多维概括。P 在 D 的均值处的距离为零，并随着 P 沿每个主分量轴远离均值而增长。如果这些轴中的每一个都被重新标定为具有单位方差，那么马哈拉诺比斯距离就对应于变换空间中的标准欧几里德距离。因此，马哈拉诺比斯距离是无单位的、尺度不变的，并考虑到了数据集的相关性。

1 定义和性质

一组观测值为 $\vec{x} = (x_1, x_2, x_3, \ldots, x_N)^\top$ 与一组具有均值 $\vec{\mu} = (\mu_1, \mu_2, \mu_3, \ldots, \mu_N)^\top$ 及其协方差矩阵 $\mathbf{S}$ 的观测值的马哈拉诺比斯距离定义为 [2] 。

$$D_M(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^\top \mathbf{S}^{-1} (\vec{x} - \vec{\mu})}$$

马哈拉诺比斯距离 (或其平方值即为“广义的点间距离的平方”[3]) 也可以定义为两个具有相同的概率分布的随机向量 $\vec{x}$ 和 $\vec{y}$ 之间的不相似度测量，它们的概率分布用协方差矩阵 $\mathbf{S}$ 表示：

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

由于 $\mathbf{S}$ 是协方差矩阵，它是半正定的，并且半正定矩阵的逆也是半正定的，我们得到的 $\mathbf{S}^{-1}$ 也是半正定的。这就解释了为什么可以取平方根，因为所有的值都是正数 [4]。

如果协方差矩阵是单位矩阵，则马哈拉诺比斯距离就会退化为欧几里德距离。如果协方差矩阵是对角的，则产生的距离度量称为**标准化欧几里德距离** (*standardized Euclidean distance*)：

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

其中 s_i 是 x_i 和 y_i 在样本集上的标准偏差。

马哈拉诺比斯距离在数据所跨越的空间的满秩线性变换下保持不变。这意味着，如果数据具有非平凡的零空间，则可以在将数据 (非退化地) 投影到数据的适当维度的任意空间后计算马哈拉诺比斯距离。

我们可以找到有用的马哈拉诺比斯距离的平方分解，这有助于解释多变量观测值离群的一些原因，也为识别离群值提供了一个图形工具 [5]。

2 直观的解释

考虑估计 N 维欧几里德空间中的一个测试点属于集合的概率的问题，其中我们得到的样本点肯定属于该集合。我们的第一步是找到样本点的中心点或质心。直观地说，测试点离这个质心越近，它就越有可能属于这个集合。

但是，我们还需要知道集合是分布在大范围内还是小范围内，这样我们就可以确定距中心的给定距离是否值得注意。最简单的方法是估计样本点到质心距离的标准偏差。如果测试点与质心之间的距离小于一个标准偏差，则我们可以得出结论，测试点极有可能属于集合。距离越远，测试点越有可能不属于集合。

通过将测试点和集合之间的标准化距离定义为 $\frac{\|x-\mu\|_2}{\sigma}$，可以量化这种直观的方法，其可解释为：$\frac{\text{testpoint} - \text{sample mean}}{\text{standard deviation}}$。通过将其插入正态分布，我们可以得出测试点属于集合的概率。

上述方法的缺点是，我们假设样本点以球形方式围绕质心分布。如果分布是明显的非球形的，例如椭球形的，那么我们期望测试点属于集合的概率不仅取决于与质心的距离，还取决于方向。在椭球体具有短轴的方向上，测试点必须更近，而在那些长轴的方向上，测试点可以离中心更远。

将其置于数学基础上，可以通过构建样本的协方差矩阵来估计最能代表集合概率分布的椭球体。马哈拉诺比斯距离是测试点到质心的距离除以测试点方向上椭球的宽度。

3 正态分布

对于任意维数的正态分布，观测的概率密度 $\vec{x}$ 由马哈拉诺比斯距离 d 唯一确定：

$$\Pr[\vec{x}]d\vec{x} = \frac{1}{\sqrt{\det(2\pi\mathbf{S})}} \exp\left(-\frac{(\vec{x}-\vec{\mu})^\top \mathbf{S}^{-1}(\vec{x}-\vec{\mu})}{2}\right) d\vec{x} = \frac{1}{\sqrt{\det(2\pi\mathbf{S})}} \exp\left(-d^2/2\right) d\vec{x}$$

具体而言，d^2 遵循具有 n 维自由度的卡方分布，其中 n 是正态分布的维数。例如，如果维数为 2，则特定计算的 d 小于某个阈值 t 的概率为 $1 - e^{-t^2/2}$。要确定实现特定概率的阈值，p，对于 2 维来说，使用 $t = \sqrt{-2\ln(1-p)}$。对于 2 以外的维度数，应参考累积卡方分布。

在正态分布中，马哈拉诺比斯距离小于 1 的区域 (即距离为 1 的椭球体内部区域) 正是概率分布为凹形的区域。

对于正态分布，马哈拉诺比斯距离与负的对数似然的平方根成正比 (在添加一个常数后，最小值为零)。

4 与正态随机变量的关系

一般来说，给定一个正态 (高斯) 随机变量 X，其方差为 $S = 1$ 且均值为 $\mu = 0$。任意其他正态随机变量 R(其均值为 μ_1 且方差为 S_1) 可以通过 X 的各项用方程 $R = \mu_1 + \sqrt{S_1}X$ 定义。相反，为了从任意正态随机变量中恢复一个正态随机变量，通常可以求解 $X = (R - \mu_1)/\sqrt{S_1}$。如果我们两边都平方，并取平方根，我们将得到一个与马哈拉诺比斯距离非常相似的度量方程：

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

由此产生的幅度总是非负的，并随着数据与平均值的距离而变化，这些属性在试图为数据定义一个模型时非常方便。

5 与杠杆统计的关系

马哈拉诺比斯距离与杠杆统计 h 密切相关，但有一个不同的尺度 [6]：

$$D^2 = (N-1)\left(h - \frac{1}{N}\right)$$

6 应用

马哈拉诺比斯的定义是由 1927 年根据测量结果确定头骨相似性的问题引起的 [7]。

马哈拉诺比斯距离广泛应用于聚类分析和分类技术中。它与用于多变量统计测试的霍特林的 T 平方分布和用于监督分类的费舍尔的线性判别分析密切相关 [8]。

为了使用马哈拉诺比斯距离将测试点分类为属于 N 个类别中的一个类别，首先为每个类别估计协方差矩阵，通常基于已知属于每个类别的样本。然后，给定一个测试样本，计算到每个类别的马哈拉诺比斯距离，并将测试点分类为属于马哈拉诺比斯距离最小的类别。

马哈拉诺比斯距离和杠杆作用通常用于检测离群点，尤其是在线性回归模型的开发中。与其他样本点群有较大的马哈拉诺比斯距离的点被认为具有较高的杠杆作用，因为它对回归方程的斜率或系数有较大的影响。马哈拉诺比斯距离也用于确定多变量的离群点。回归技术可用于通过两个或多个变量得分的组合确定样本总体中的一个特定案例是否是离群点。即使对于正态分布来说，一个点也可以是多变量离群点，即使它对任意变量来说都不是一个单变量的离群点 (例如，考虑沿着线 $x_1 = x_2$ 集中的概率密度)，使得马哈拉诺比斯距离成为比单独检查维度更敏感的度量。

7 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, R.; Jouan-Rimbaud, D.; Massart, D. L. “The Mahalanobis distance”. Chemometrics and Intelligent Laboratory Systems. 50 (1): 1–18. doi:10.1016/s0169-7439(99)00047-7.
3. Gnanadesikan, R.; Kettenring, J. R. (1972). “Robust Estimates, Residuals, and Outlier Detection with Multiresponse Data”. Biometrics. 28 (1): 81 – 124. doi:10.2307/2528963. JSTOR 2528963.
4. “Inverse matrix’s eigenvalue?”.
5. Kim, M. G. (2000). “Multivariate outliers and decompositions of Mahalanobis distance”. Communications in Statistics –Theory and Methods. 29 (7): 1511–1526. doi:10.1080/03610920008832559.
6. Weiner, Irving B.; Schinka, John A.; Velicer, Wayne F. (23 October 2012). Handbook of Psychology, Research Methods in Psychology. John Wiley & Sons. ISBN 978-1-118-28203-8.
7. Mahalanobis, Prasanta Chandra (1927); Analysis of race mixture in Bengal, Journal and Proceedings of the Asiatic Society of Bengal, 23:301–333.
8. McLachlan, Geoffrey (4 August 2004). Discriminant Analysis and Statistical Pattern Recognition. John Wiley & Sons. pp. 13–. ISBN 978-0-471-69115-0.