马哈拉诺比斯距离自下而上的解释

提问与回答

Apr 13, 2017

1 提问

我正在学习模式识别和统计学，几乎每一本我打开的关于这个主题的书，我都会碰到马哈拉诺比斯距离的概念。这些书给出了一些直观的解释，但仍然不够好，无法让我真正理解到底发生了什么。如果有人问我“什么是马哈拉诺比斯距离?”我只能回答：“这是一个很好的东西，可以测量某种距离”：

定义通常也包含特征向量和特征值，我在将其与马哈兰诺比斯距离联系起来时遇到了一点麻烦。我理解特征向量和特征值的定义，但它们与马哈拉诺比斯距离有什么关系？这是否与线性代数中改变基数等有关?

我也读过这些以前关于这个问题的问题。：

- 什么是马哈兰诺比斯距离，以及它在模式识别中如何使用?
- 高斯分布函数和马哈兰诺比斯距离的直观解释 (Math.SE)。

我也读过这个解释。

答案很好，图片也很好，但我还是不太明白..... 我有一个想法，但还不太清楚。有人能给我一个“你会怎么向你奶奶解释”的解释吗？这样我就能最终把这个问题解决了，并且不再为“什么是马哈拉诺比斯距离?”，“它从哪里来，是什么，为什么?”这些问题烦恼。

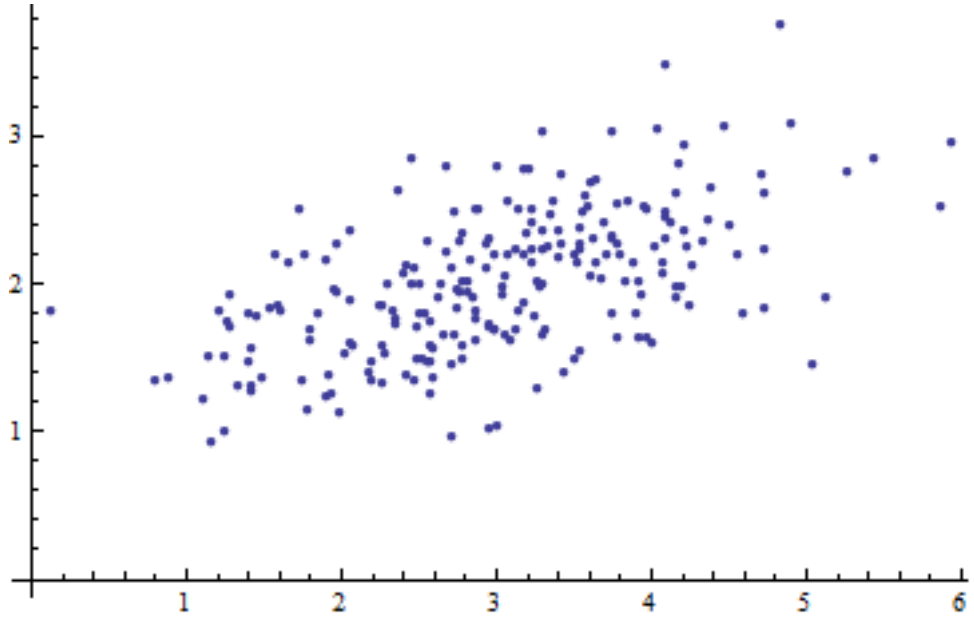
更新：

以下是一些有助于理解马哈拉诺比斯公式的东西：
https://math.stackexchange.com/questions/428064/distance-of-a-test-point-from-the-center-of-an-ellipsoid

2 回答 1

2.1 几何直觉

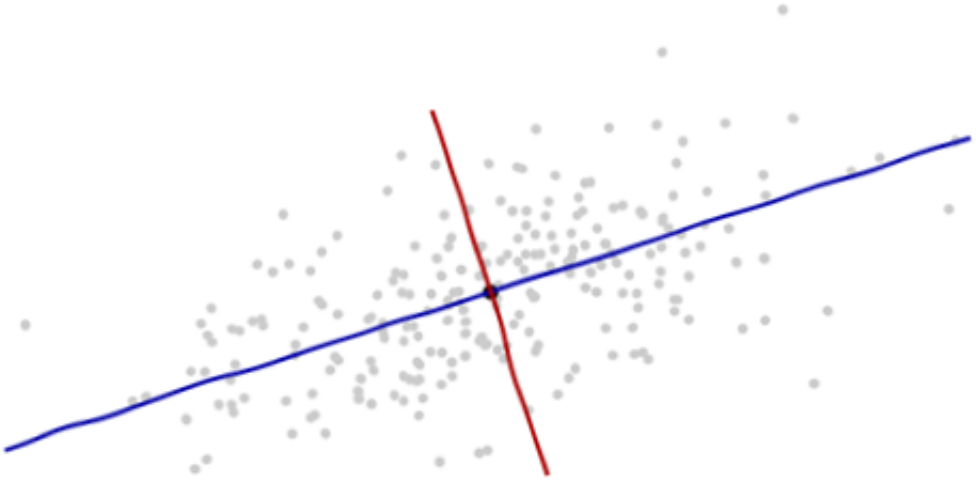
以下是一些多变量数据的散点图 (二维)：



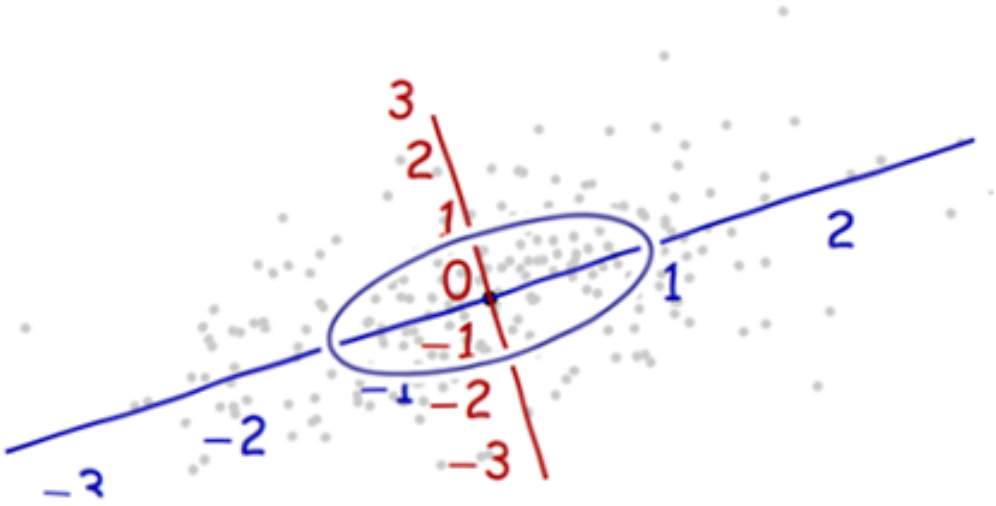
当坐标轴被忽略时，我们能做什么？

介绍数据本身建议的坐标。

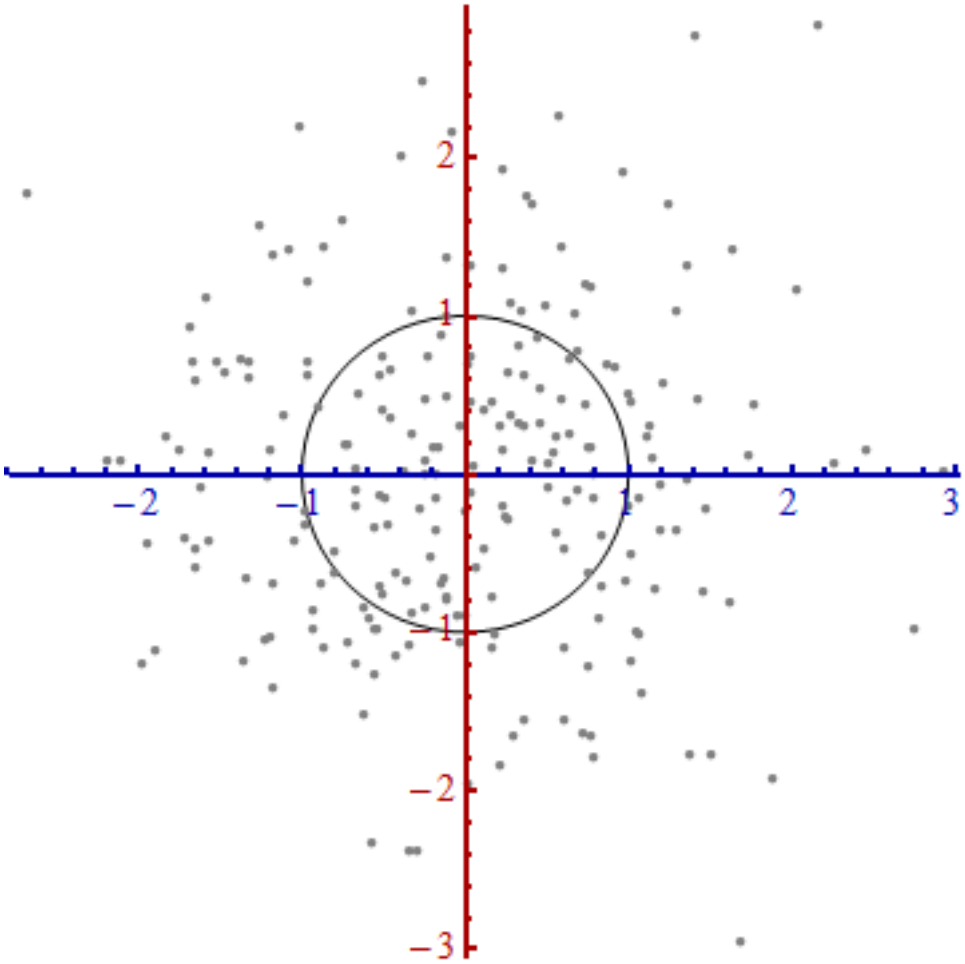
原点将位于各点的中心点 (各点均值的点)。**第一个坐标轴** (下图中的蓝色) 将沿着点的“尖峰”延伸，这 (根据定义) 是方差最大的任意方向。**第二个坐标轴** (图中的红色) 将垂直于第一个坐标轴延伸。(在两个以上的维度中，将选择方差尽可能大的垂直方向，以此类推。)



我们需要一个**刻度**。沿每个轴的标准偏差将很好地确定沿轴的单位。记住 $68-95-99.7$ 规则：大约三分之二 (68%) 的点应该在原点的一个单位内 (沿轴)；大约 95% 应该在两个单位内。这样就很容易找到正确的单位。为便于参考，下图包括以下单位中的单位圆：



那看起来不像一个圆，是吗？这是因为这张图片被扭曲了 (两个轴上的数字之间的间距不同就证明了这一点)。让我们用轴的正确方向 (从左到右，从下到上) 和单位纵横比重新绘制它，这样水平方向上的一个单位实际上等于垂直方向上的一个单位：



你是在这张图片中测量马哈拉诺比斯距离，而不是在原始图片中。

这里发生了什么？我们让数据告诉我们如何构建一个坐标系，以便在散点图中进行测量。仅此而已。尽管我们在这一过程中有一些选择 (我们总是可以反转其中一个或两个轴；在极少数情况下，沿着“尖峰”的方向 — 主方向 — 不是唯一的)，但它们不会改变最终绘图中的距离。

2.2 技术注释

(这并不是为了奶奶，她很可能在数字再次出现在地块上时就开始失去兴趣，而是为了解决剩下的问题。)

- 沿着新的坐标轴的单位向量是**特征向量** (协方差矩阵或其逆矩阵)。
- 我们注意到，将沿着每个特征向量的距离**除以**标准偏差 (协方差的平方根)，将不失真地把椭圆变成一个圆。让 C 代表协方差函数，两点 x 和 y 之间的新的距离 (马哈拉诺比斯距离) 是 x 到 y 的距离除以 $C(x-y, x-y)$ 的平方根。相应的代数运算，现在认为 C 表示为矩阵，x 和 y 表示为向量，写为 $\sqrt{(x-y)'C^{-1}(x-y)}$。**无论用什么基来表示向量和矩阵**，这都是有效的。特别是，这是**在原始坐标中**马哈拉诺比斯距离的正确公式。
- 在最后一步中，轴被扩展的量是逆协方差矩阵的 (平方根) **特征值**。等价地，轴被协方差矩阵的 (根) 特征值**收缩**。因此，散射越多，将椭圆转换为圆所需的收缩就越多。

- 尽管此过程始终适用于任意数据集，但对于近似多元正态分布的数据，它看起来非常漂亮 (经典足球形状的云)。在其他情况下，平均点可能无法很好地表示数据中心，或者无法使用方差作为传播的度量以准确识别“尖峰”(数据的一般趋势)。
- 坐标原点的移动、轴的旋转和扩展共同形成**仿射变换**。除了初始偏移外，这是一个基的改变，从原来的基 (使用指向正坐标方向的单位向量) 到新的基 (使用单位特征向量的选择)。
- 这与主成分分析 (PCA) 密切相关。仅此一点就足以解释“它从何而来”和“为什么”问题 — 如果你还不坚信让数据决定你用来描述它们和测量它们的差异的坐标的优雅和实用性的话。
- 对于多元正态分布 (我们可以使用概率密度的特性而不是点云的类似特性进行相同的构造)，马哈拉诺比斯距离 (到新原点) 出现在描述标准正态分布的概率密度的表达式 $\exp(-\frac{1}{2}x^2)$ 中，以取代 “x”。因此，在新坐标系中，当投影到通过原点的任意直线上时，多元正态分布看起来像是**标准正态分布**。特别是，它在每个新坐标中都是标准正态分布。从这个角度来看，多元正态分布之间唯一本质上的区别在于它们使用了多少维度。(注意，该维度数量可能是，而且有时是，小于维度尺寸数量。)
- 注意我在提到仿射变换之后，立即对其进行了描述：先是平移，然后是基的改变。我选择这种语言是因为它与问题中使用的语言相同。(我们必须对 "基的改变 "做一些自由的理解，以包括不可逆转的线性变换：这是一个对 PCA 很重要的问题，它有效地放弃了一些基础元素。)

马哈拉诺比斯距离定义为 $d(x,y) = \sqrt{(x-y)^T\Sigma^{-1}(x-y)}$，其中 Σ 是某些数据的协方差矩阵的估计；这意味着它是对称的。如果用于估计 Σ 的列不是线性相关的，则 Σ 是正定的。对称矩阵是可对角化的，其特征值和特征向量是实数。PD 矩阵的特征值均为正。特征向量可以选择为具有单位长度，并且是正交的，因此我们可以写为 $\Sigma = Q^TDQ$ 和 $\Sigma^{-1} = QD^{-\frac{1}{2}}D^{-\frac{1}{2}}Q^T$。将其插入到距离定义中，$d(x,y) = \sqrt{[(x-y)^TQ]\,D^{-\frac{1}{2}}D^{-\frac{1}{2}}\,[Q^T(x-y)]} = \sqrt{z^Tz}$。显然，方括号中的乘积是转置，乘以 Q 的效果是将向量 $(x-y)$ 旋转到一个正交基。最后，$D^{-\frac{1}{2}}$ 是对角线，由对角线上的每个元素倒置，然后取平方根来生成，是对每个向量的每个元素进行重新缩放。事实上，$D^{-\frac{1}{2}}$ 是正交空间中每个特征的逆标准偏差 (即 D^{-1} 是一个精度矩阵，由于数据是在正交基上，所以矩阵是对角线的)。其效果是通过“扁平化”其轴线，将旋转椭圆的物体变换成一个圆。很明显，z^Tz 是以平方单位测量的，所以取平方根会将距离返回到原始单位。

3 回答 2

马哈拉诺比斯与其说是“配料量”的距离，不如说是“最佳口味”的距离。真正“有效”的成分，那些对变化非常敏感的成分，是你必须最仔细控制的成分。

如果你考虑一下任意高斯分布和标准正态分布的区别是什么？基于中心趋势 (均值) 和变化趋势 (标准差) 的中心和标度。一个是另一个的坐标变换。马哈拉诺比斯就是这种转变。它向你展示了如果你的兴趣分布被重新转换为标准正态分布而不是高斯分布，世界会是什么样子。

4 回答 3

作为起点，我认为马哈拉诺比斯距离是在 $\mathbb{R}^n$ 中的向量 x 和 y 之间的通常的欧几里德距离 $d(x,y)=\sqrt{\langle x,y\rangle}$ 的适当变形。这里的额外信息是，x 和 y 实际上是**随机**向量，即随机变量向量 X 不同的两维实现，位于我们讨论的背景中。马哈拉诺比斯试图解决的问题如下：

知道 x 和 y 是同一个多元随机变量的实现，我如何衡量它们之间的“不相似性”?

很明显，任意实现 x 与自身的不相似性应该等于 0；此外，不相似性应该是实现的对称函数，并且应该反映背景中随机过程的存在。最后一个方面是通过引入多元随机变量的协方差矩阵 C 来考虑的。

综合以上想法，我们自然会得出结论

$$D(x,y)=\sqrt{(x-y)\,C^{-1}(x-y)}$$

如果多元随机变量 $X=(X_1,\dots,X_n)$ 的成分 X_i 是不相关的，例如，$C_{ij}=\delta_{ij}$(我们对 X_i 进行了 "归一化"，以便有 $Var(X_i)=1$)，那么马哈拉诺比斯距离 $D(x,y)$ 是 x 和 y 之间的欧几里德距离。在存在非平凡相关性的情况下，(估计的) 相关性矩阵 $C(x,y)$“变形”欧几里德距离。

5 回答 4

为了补充上面的优秀解释，马哈拉诺比斯距离在（多元）线性回归中自然产生。这是在其他答案中讨论的马哈拉诺比斯距离和高斯分布之间的一些联系的一个简单结果，但我认为无论如何都值得解释一下。

假设我们有一些数据 $(x_1,y_1),\dots,(x_N,y_N)$，其中 $x_i\in\mathbb{R}^n$ 且 $y_i\in\mathbb{R}^m$。我们假设存在一个参数向量 $\beta_0\in\mathbb{R}^m$ 和一个参数矩阵 $\beta_1\in\mathbb{R}^{m\times n}$，使得 $y_i=\beta_0+\beta_1x_i+\epsilon_i$，其中 $\epsilon_1,\dots,\epsilon_N$ 是具有均值为 0 且协方差为 C 的 m 维高斯随机向量（并且它们独立于 x_i）。那么给定 x_i 的 y_i 是高斯的，均值为 $\beta_0+\beta_1x_i$，且协方差为 C。

由此可见，给定 x_i 的 y_i 的负的对数似然（作为 $\beta=(\beta_0,\beta_1)$ 的函数）由以下公式给出

$$-\log p(y_i\mid x_i;\beta)=\frac{m}{2}\log(2\pi\det C)+\frac{1}{2}(y_i-(\beta_0+\beta_1x_i))^\top C^{-1}(y_i-(\beta_0+\beta x_i)).$$

我们把协方差 C 视为常数，所以

$$\operatorname{argmin}_\beta[-\log p(y_i\mid x_i;\beta)]=\operatorname{argmin}_\beta D_C(\beta_0+\beta_1x_i,y_i),$$

其中

$$D_C(\hat{y},y)=\sqrt{(y-\hat{y})^\top C^{-1}(y-\hat{y})}$$

是 $\hat{y},y\in\mathbb{R}^m$ 之间的马哈拉诺比斯距离。

根据独立性，$\mathbf{x}=(x_1,\dots,x_N)$ 给定的 $\mathbf{y}=(y_1,\dots,y_N)$ 的对数似然 $\log p(\mathbf{y}\mid\mathbf{x};\beta)$，由总和给出

$$\log p(\mathbf{y}\mid\mathbf{x};\beta)=\sum_{i=1}^N\log p(y_i\mid x_i;\beta)$$

因此

$$\operatorname{argmin}_\beta[-\log p(\mathbf{y}\mid\mathbf{x};\beta)]=\operatorname{argmin}_\beta\frac{1}{N}\sum_{i=1}^N D_C(\beta_0+\beta_1x_i,y_i),$$

其中因子 $1/N$ 不会影响 argmin。

总之，使观测数据的负对数似然（即最大似然）最小化的系数 β_0, β_1 也使数据的经验风险最小化，损失函数由马哈拉诺比斯距离给出。