什么是马哈拉诺比斯距离?

Rick Wicklin

February 15, 2012

我之前描述过如何利用马哈拉诺比斯距离寻找多元数据中的离群值. 本文将更深入地了解马哈拉诺比斯距离。接下来的文章将描述如何计算马哈拉诺比斯距离。

1 以标准单位表示的距离

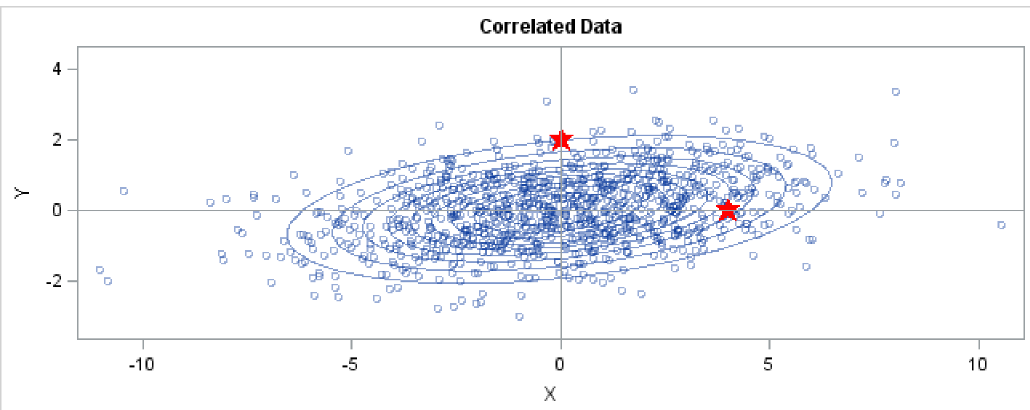
在统计学中，我们有时根据数据的规模来衡量“接近”或“远离”。通常“标度”意味着“标准差”。对于单变量数据，我们说，与均值相差一个标准差的观测值比三个标准差的观测值更接近均值。(你也可以通过指定两个观测值之间的标准差来指定它们之间的距离。)

对于许多分布，如正态分布，这种标度的选择也说明了概率。具体地说，它更可能观测到与均值约为一个标准差的观测值，而不是观测到几个标准差的观测值。为什么? 因为概率密度函数在均值附近更高，当你离开许多标准差时，概率密度函数几乎为零。

对于正态分布的数据，你可以通过计算所谓的z-score来指定与平均值的距离。对于一个值 x，x 的 z-score 为 $z = (x - \mu)/\sigma$，其中 μ 为总体均值，σ 为总体标准差。这是一个无量纲的量，你可以解释为 x 与均值的标准差的数量。

2 距离并不总是看起来的那样

你可以把这些概念推广到多元正态分布。下图显示了模拟的双元正态数据，并与预测椭圆相叠加。图中的椭圆是生成数据的双元正态分布的 10%(最里面)、20%、... 和 90%(最外面) 的预测椭圆。预测椭圆是双元正态密度函数的等值线。靠近原点的椭圆，如 10% 的预测椭圆，其概率密度很高。对于较远的椭圆，如 90% 的预测椭圆，其密度较低。



在图中，用红星作为标记，显示了两个观测结果。第一个观测值在坐标 $(4,0)$，而第二个观测值在 $(0,2)$。问题是：哪个标记更接近原点？(原点是这个分布的多元中心)。

答案是，“这取决于你如何测量距离。”欧几里德距离分别是 4 和 2，所以你可以得出结论，在 (0，2) 处的点更接近原点。然而，对于这种分布，Y 方向的方差小于 X 方向的方差，因此在某种意义上，点 $(0,2)$ 距离原点的“标准差”比 $(4,0)$ 大。

注意这两个观测点相对于椭圆的位置。点 $(0,2)$ 位于 90% 预测椭圆处，而 $(4,0)$ 处的点位于约 75% 预测椭圆处。这是什么意思？这意味着 $(4,0)$ 处的点与原点“更接近”，因为在 $(4,0)$ 附近观测的可能性比在 $(0,2)$ 附近观测的可能性大。在 $(4,0)$ 附近的概率密度高于在 $(0,2)$ 附近的概率密度。

从这个意义上说，预测椭圆是“标准差单位”的多元概括。你可以使用双元概率等值线来比较与双元均值的距离。如果包含 p 的等值线嵌套在包含 q 的等值线内，则点 p 比点 q 更接近。

3 定义马哈拉诺比斯距离

你可以用概率等值线来定义马哈拉诺比斯距离。马哈拉诺比斯距离具有以下特性：

- 它解释了一个事实，即每个方向的方差是不同的。
- 它解释了变量之间的协方差。
- 对于具有单位方差的不相关变量，它可以归结为常见的欧几里德距离。

对于单变量正态数据，单变量 z-score 标准化了分布 (使其具有均值 0 和单位方差)，并给出了一个无量纲的量，用数据的标度来指定从观测值到均值的距离。对于均值 μ 和协方差矩阵 Σ 的多元正态数据，通过应用 Cholesky 变换，你可以对变量进行去相关化并使分布标准化 $z = L^{-1}(x - \mu)$，其中 L 是 Σ 的 Cholesky 因子，$\Sigma = LL^\top$。

对数据进行转换后，你可以计算出点 z 到原点的标准欧几里德距离。为了去掉平方根，我将计算欧几里德距离的平方，即 $\mathrm{dist}^2(z,0) = z^\top z$。它测量一个点离原点有多远，它是 z-score 的多元泛化。

你可以用原来的相关变量来重写 $z^\top z$ 表。平方距离 $\mathrm{Mahal}^2(x,\mu)$ 为

$$
\begin{aligned}
\mathrm{Mahal}^2(x,\mu) &= z^\top z \\
&= \left(L^{-1}(x-\mu)\right)^\top \left(L^{-1}(x-\mu)\right) \\
&= (x-\mu)^\top \left(LL^\top\right)^{-1}(x-\mu) \\
&= (x-\mu)^\top \Sigma^{-1}(x-\mu)
\end{aligned}
$$

最后一个公式是平方马哈拉诺比斯距离的定义。推导过程中使用了几个矩阵特性，如 $(AB)^\top = B^\top A^\top$，$(AB)^{-1} = B^{-1}A^{-1}$，以及 $(A^{-1})^\top = (A^\top)^{-1}$。注意，如果 Σ 是单位矩阵，那么马哈拉诺比斯距离就简化为 x 和 μ 之间的标准的欧几里得距离。

马哈拉诺比斯距离解释了每个变量的方差和变量之间的协方差。在几何上，它通过将数据转换为标准化的不相关数据，并计算转换后数据的普通欧几里德距离来实现这一点。这样，马哈拉诺比斯距离就像一个单变量的 z-score：它提供了一种考虑数据规模的测量距离的方法。