常见距离度量方法优缺点对比!

原创 张峰 Datawhale 5天前

Datawhale干货

译者：张峰，安徽工业大学，Datawhale成员

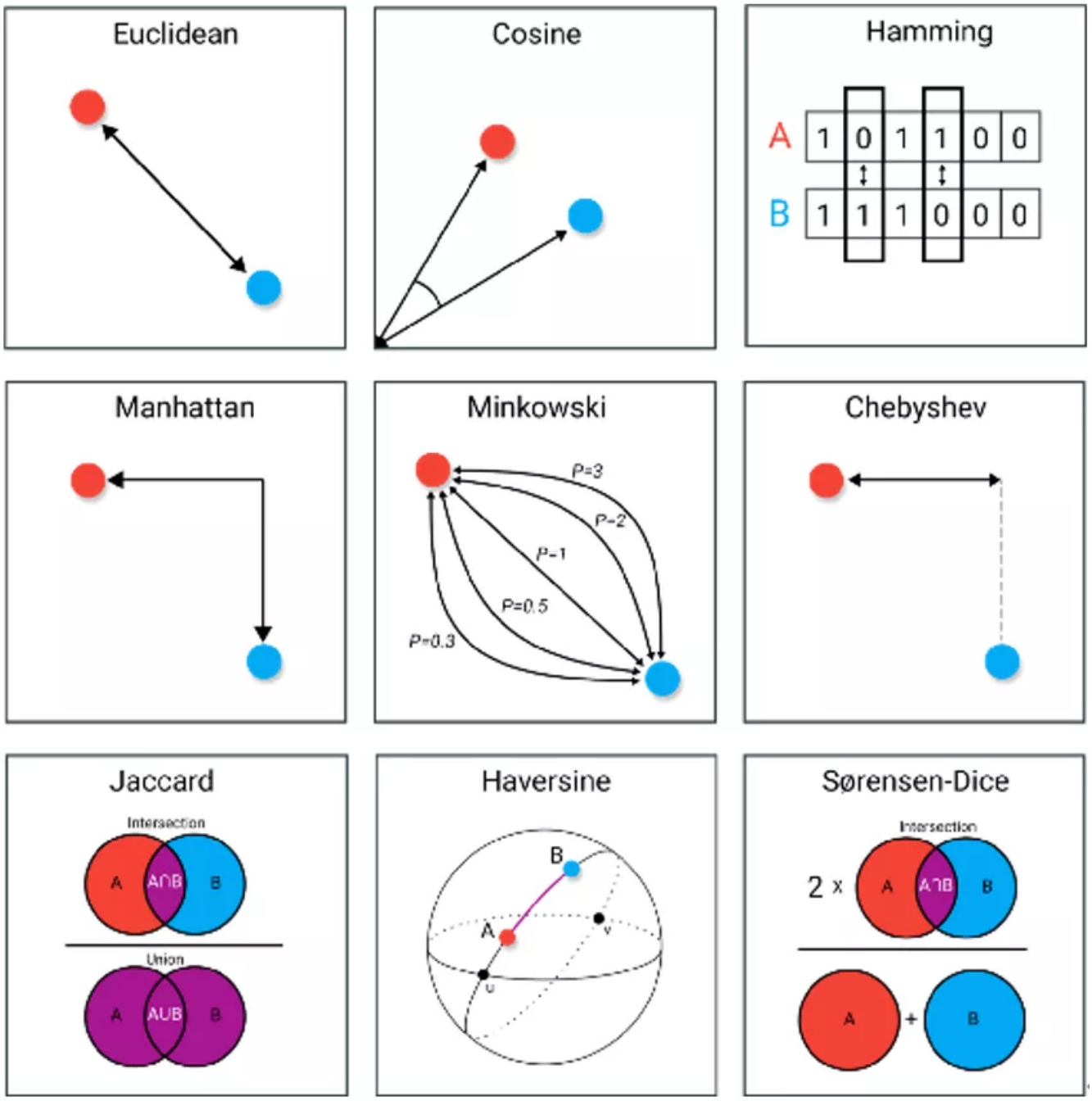
许多算法，不管是有监督的还是无监督的，都会使用距离测量。这些度量方法，如欧氏距离或余弦相似度，经常可以在KNN、UMAP、HDBSCAN等算法中找到。

理解距离测量领域比你可能意识到的更重要。以KNN为例，这是一种常用于监督式学习的技术。作为默认设置，它通常使用欧几里得度量。就其本身而言，是一个很好的距离测量方法。

然而，如果你的数据是高维的呢？那么欧几里得距离还能用吗？或者，如果你的数据由地理空间信息组成呢？也许Haversine距离会是一个更好的选择！

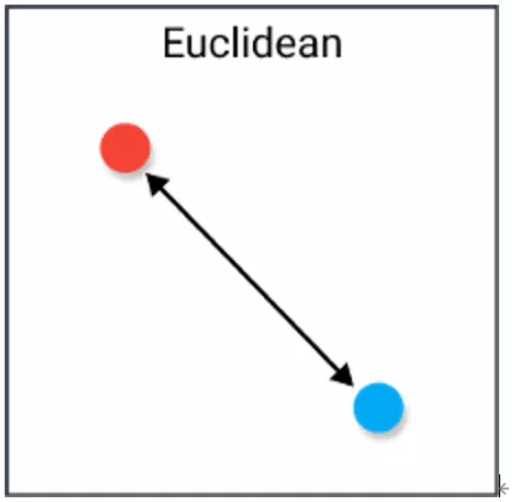
知道何时使用哪种距离测量方法可以帮助你从一个差的分类器变成一个准确的模型。

在本文中，我们将介绍不同的距离测量方法，并探索如何以及何时最好地使用它们。最重要的是，我会谈谈各自的缺点，这样你就能知道何时该避开使用某些距离度量的措施。



1. 欧式距离

我们从最常见的距离测量开始，即欧氏距离。它是一种最好的距离测量方法，可以解释为连接两点的线段长度。



这个公式相当简单，因为距离是根据使用勾股定理的点的笛卡尔坐标计算出来的。

$$D(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

缺点

虽然这是一种常见的距离测量方法，但欧几里得距离并不是尺度不变的，这意味着计算出的距离可能会根据特征的单位而有所偏斜。通常情况下，在使用这种距离测量之前，需要对数据进行归一化。

此外，随着数据维度的增加，欧几里得距离的作用就越小。这与维度的诅咒有关，它涉及到高维空间的概念，并不像我们直观地期望的那样，从二维或三维空间中发挥作用。

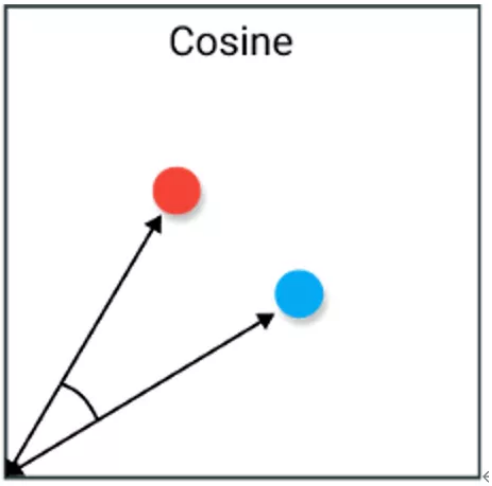
用例

当你有低维数据，并且向量的大小很重要，需要测量时，欧氏距离的效果非常好。如果在低维数据上使用欧氏距离，kNN和HDBSCAN等方法就会显示出很好的效果。

虽然已经开发了许多其他的测量方法来解释欧氏距离的缺点，但它仍然是最常用的距离测量方法之一，这是有充分理由的。它使用起来非常直观，实现起来也很简单，并且在许多用例中都显示出了很好的效果。

2. 余弦相似性

余弦相似性经常被用来抵消欧几里得距离的高维度问题。余弦相似性只是两个向量之间角度的余弦。如果将它们归一化为都有长度为1的向量，它的内积也相同。



两个方向完全相同的向量的余弦相似度为1，而两个方向截然相反的向量的相似度为-1，请注意，它们的大小并不重要，因为这是方向的量度。余弦相似度公式为：

$$D(x,y) = cos(\theta) = \frac{x \cdot y}{\|x\| \|y\|}$$

缺点

余弦相似性的一个主要缺点是不考虑向量的大小，只考虑其方向。在实际应用中，这意味着值的差异没有被完全考虑。以推荐系统为例，那么余弦相似性并没有考虑到不同用户之间的评分等级差异。

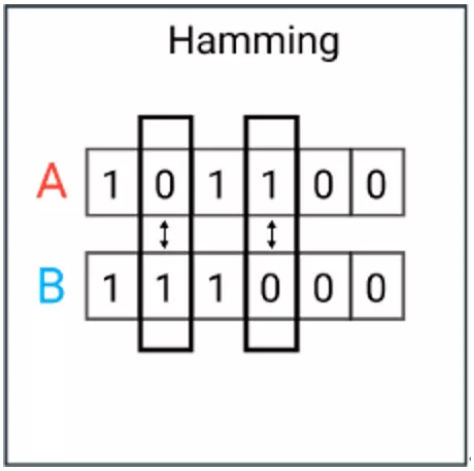
用例

当我们有高维数据且向量的大小并不重要时，我们经常使用余弦相似度。对于文本分析来说，当数据用字数来表示时，这种测量方法是很常用的。

例如，当一个词在一个文档中出现的频率高于另一个文档时，这并不一定意味着一个文档与该词的关系更大。可能是文档的长度不均匀，计数的大小就不那么重要了。那么，我们最好是使用不考虑大小的余弦相似性。

3. 汉明距离

汉明距离是指两个向量之间相差的数值。它通常用于比较两个长度相等的二进制字符串。它也可以用来比较字符串之间的相似度，计算彼此不同的字符数。



缺点

正如你所预料的，当两个向量的长度不相等时，汉明距离很难使用。你会希望将相同长度的向量相互比较，以了解哪些位置不匹配。

而且，只要它们不同或相等，它就不考虑实际值。因此，当幅度是一个重要的衡量标准时，不建议使用这个距离衡量。

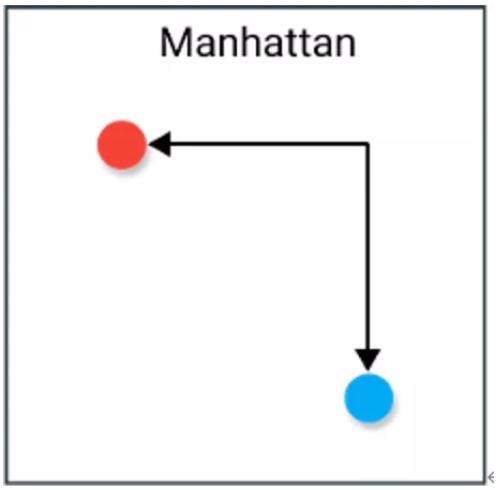
用例

典型的使用情况包括在计算机网络上传输数据时的纠错/检测。它可以用来确定二进制字中的失真位数，以此来估计错误。

此外，你还可以使用汉明距离来测量分类变量之间的距离。

4. 曼哈顿距离

曼哈顿距离，通常被称为出租车距离或城市街区距离，计算实值向量之间的距离。想象一下，在统一的网格上描述物体的向量，如棋盘。



曼哈顿距离则是指两个向量之间的距离，如果它们只能移动直角。计算距离时不涉及对角线的移动。曼哈顿距离公式为：

$$D(x,y) = \sum_{i=1}^{k} |x_i - y_i|$$

缺点

虽然曼哈顿距离对于高维数据似乎还不错，但它是一个比欧几里得距离更不直观的测量方法，尤其是在高维数据中使用时。

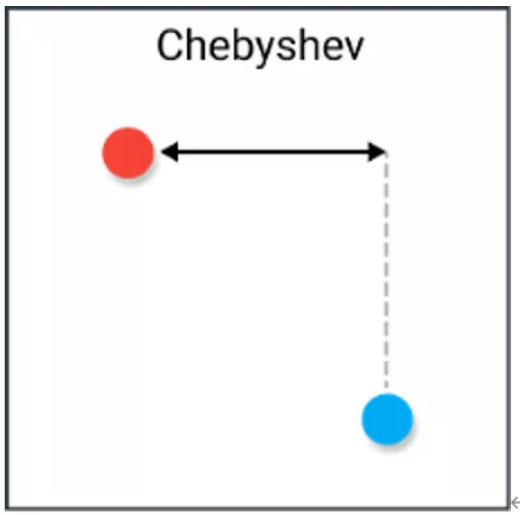
而且，它比欧几里得距离更容易给出一个更高的距离值，因为它不可能是最短路径。这不一定会带来问题，但你应该考虑到这一点。

用例

当你的数据集有离散和/或二进制属性时，曼哈顿似乎很好用，因为它考虑到了现实中在这些属性值内可以采取的路径。以欧氏距离为例，会在两个向量之间创建一条直线，而在现实中这可能实际上是不可能的。

5. 切比雪夫距离

切比雪夫距离被定义为沿任何坐标维度的两个向量之间的最大差异。换句话说，它是沿着一个轴线的最大距离。



由于它的性质，它经常被称为棋盘距离，因为国王从一个方格走到另一个方格所需的最少步数等于切比雪夫距离。切比雪夫距离公式为：

$$D(x,y) = \max_i(|x_i - y_i|)$$

缺点

切比雪夫通常用于非常特殊的使用情况，这使得它很难像欧几里得距离或余弦相似性那样作为一个通用的距离度量。出于这个原因，我们建议只有当你绝对确定它适合你的使用情况时才使用它。

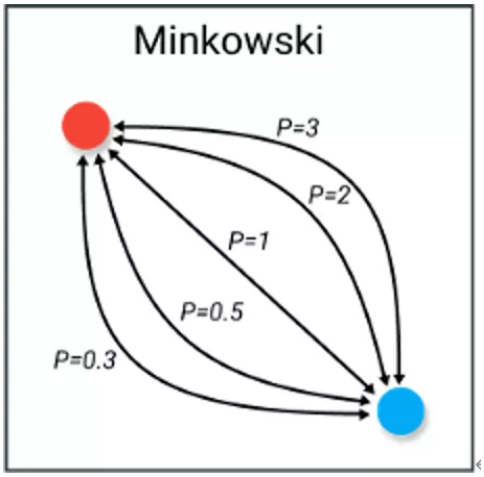
用例

如前所述，切比雪夫距离可以用来提取从一个方格到另一个方格所需的最少步数。此外，在允许无限制的8向移动的棋局中，它也是一个有用的测量方法。

在实践中，切比雪夫距离经常被用于仓库物流，因为它很像天车移动一个物体所需的时间。

6. 闵可夫斯基距离

闵可夫斯基距离是一个相对复杂的度量方法。它是在规范向量空间（n维实空间）中使用的一种度量方法，这意味着它可以在表示为一个有长度的向量空间中使用。



这个度量有三个要求：

零向量：零向量的长度为零，而其它向量的长度为正。例如，如果我们从一个地方到另一个地方，那么这个距离总是正数。但是，如果我们从一个地方到它本身，那么这个距离就是零；

标量因子：当你用正数乘以向量时，它的长度会改变，但方向不变。例如，如果我们在一个方向上走了一定的距离，再加上同样的距离，方向不会改变；

三角形不等式：两点之间的最短距离是一条直线。

闵可夫斯基距离的公式如下:

$$D(x,y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}}$$

这个距离度量最有趣的是使用参数p。我们可以用这个参数来操作距离度量，使之与其它度量方法非常相似。

p的常见值有：

$p = 1$ - 曼哈顿距离；

$p = 2$ - 欧氏距离；

$p = \infty$ -切比雪夫距离。

缺点

闵可夫斯基距离的缺点与它们所代表的距离度量一样，所以对曼哈顿、欧几里得和切比雪夫距离等度量的了解是极其重要的。

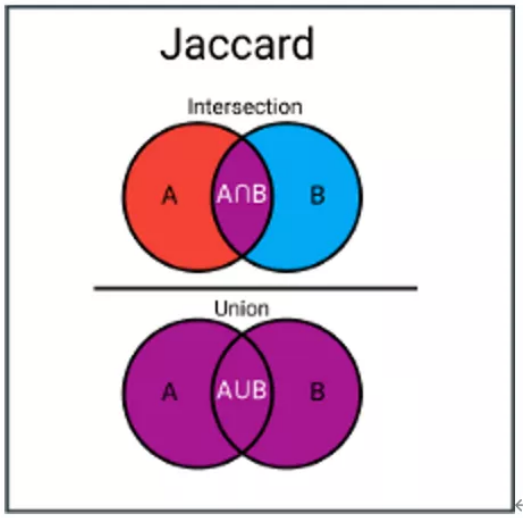
此外，参数 p 实际上在工作中可能会很麻烦，因为根据你的用例，找到正确的值可能会在计算上相当低效。

用例

p的优点是可以对它进行迭代，找到最适合你的使用情况的距离度量。它允许你在距离度量上有很大的灵活性，如果你对p和许多距离度量非常熟悉，这将是一个巨大的好处。

7. Jaccard指数

Jaccard指数（或称交集比联合）是一种用于计算样本集相似性和多样性的度量。它是交集的大小除以样本集的联合大小。



在实践中，它是集合之间相似实体的总数除以实体的总数。例如，如果两个集合有1个共同的实体，而总共有5个不同的实体，那么Jaccard指数将是$1/5 = 0.2$。

要计算Jaccard距离，我们只需将Jaccard指数从1中减去。Jaccard距离公式为：

$$D(x,y) = 1 - \frac{|x \cap y|}{|y \cup x|}$$

缺点

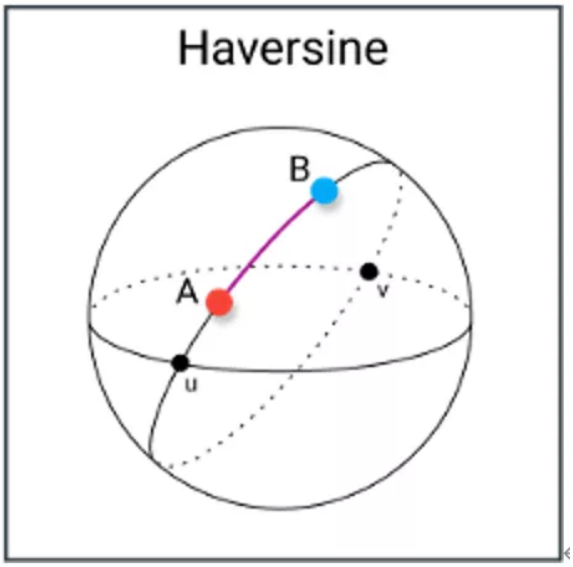
Jaccard指数的一个主要缺点是，它受数据大小的影响很大。大的数据集会对指数产生很大的影响，因为它可以在保持相似的交叉点的同时显著增加联合。

用例

Jaccard指数经常用于使用二进制或二值化数据的应用中。当你有一个深度学习模型预测图像的片段时，例如，一辆汽车，Jaccard指数就可以用来计算给定真实标签的预测片段的准确度。同样，它也可以用于文本相似性分析，以衡量文档之间的选词重叠程度。因此，它可以用来比较模式的集合。

8. Haversine距离

Haversine距离是指球面上两点之间的经度和纬度距离。



它与欧几里得距离非常相似，因为它计算的是两点之间的最短线。主要的区别是不可能有直线，因为这里的假设是两点在一个球体上。两点间的Haversine距离公式为：

$$d = 2\arcsin\left(\sqrt{\sin^2(\frac{\varphi_2 - \varphi_1}{2}) + \cos(\varphi_1)\cos(\varphi_2)\sin^2(\frac{\lambda_2 - \lambda_1}{2})}\right)$$

缺点

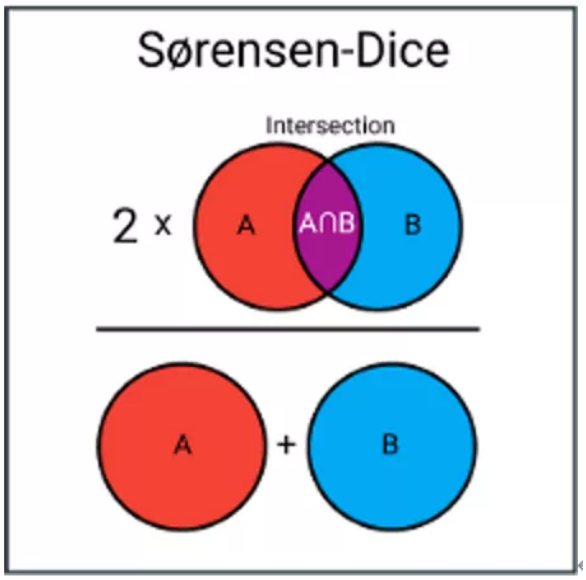
这种距离测量方法的一个缺点是，它假定各点位于一个球体上。在实践中，这种情况很少发生，例如，地球并不是完全的圆形，这可能会使计算在某些情况下变得困难。相反，如果能采用Vincenty距离，则会很有趣，因为它假设的是一个椭圆体。

用例

正如你所期望的那样，Haversine距离经常用于导航。例如，当你在两个国家之间飞行时，你可以用它来计算它们之间的距离。需要注意的是，如果本身距离已经不大，它就不太适合了。曲率不会有那么大的影响。

9. Sørensen-Dice指数

Sørensen-Dice指数与Jaccard指数非常相似，因为它衡量样本集的相似性和多样性。



虽然它们的计算方法相似，但Sørensen-Dice指数更直观一些，因为它可以被看作是两组之间的重叠百分比，这个数值在0和1之间。Sørensen–Dice指数公式为：

$$D(x,y) = \frac{2|x \cap y|}{|x| + |y|}$$

缺点

与Jaccard指数一样，它们都高估了集合的重要性，只有很少或没有TP（Truth Positive）值的正集合。因此，它可以求得多盘的平均分数。它将每个项目与相关集合的大小成反比加权，而不是平等对待它们。

用例

与Jaccard指数相似，通常用于图像分割任务或文本相似性分析。

注意：除了这里提到的9种距离度量，还有更多的度量。如果你正在寻找更多有趣的度量，我建议你研究以下其中一个：Mahalanobis, Canberra, Braycurtis, 和 KL-散度 !