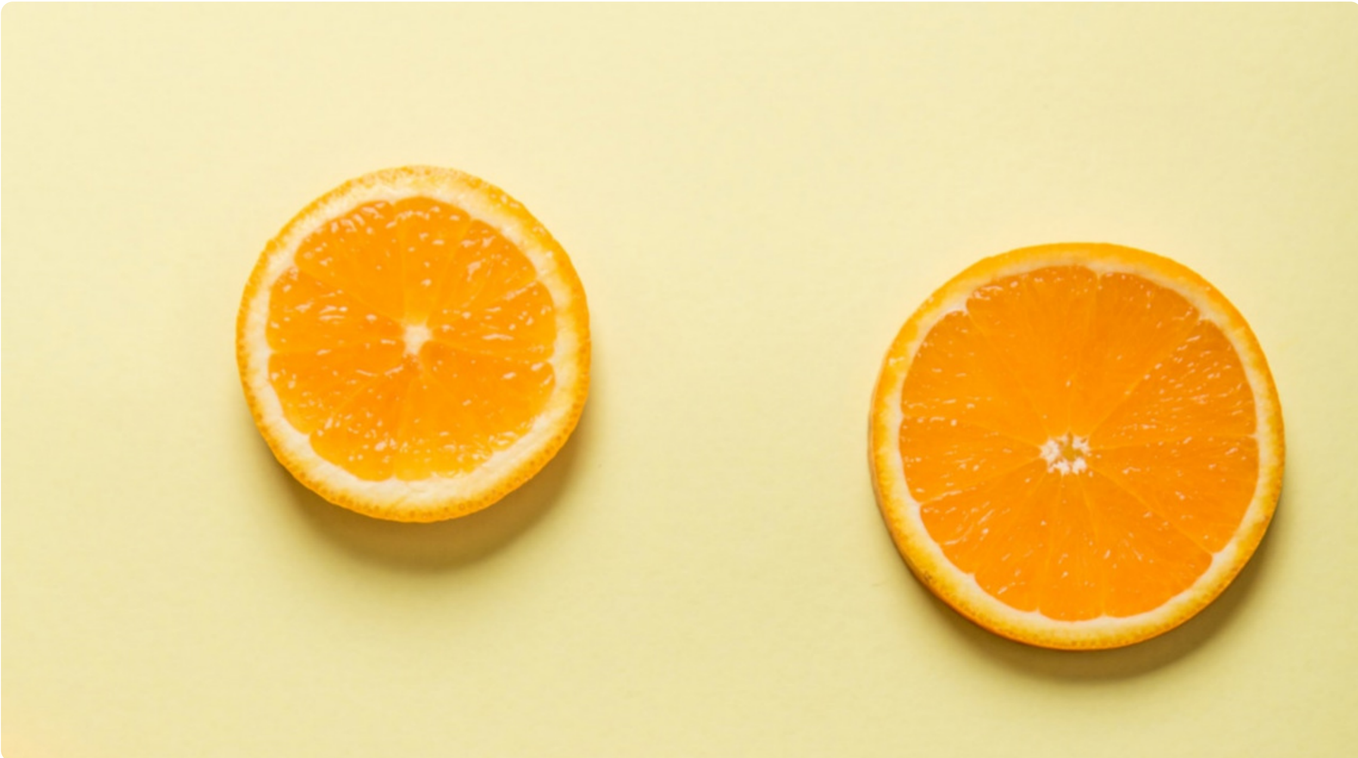
34 | 向量空间模型：如何让计算机理解现实事物之间的关系？

2019-03-04 黄申



讲述：黄申
时长 12:24 大小 11.37M

你好，我是黄申。

之前我们讲过如何让计算机理解现实世界中的事物，方法是把事物的各种特性转为机器所能理解的数据字段。而这些数据字段，在机器学习里通常被称为特征。有了特征，我们不仅可以刻画事物本身，还能刻画不同事物之间的关系。

上一个模块我们只是了解了监督式学习，重点考察了特征和分类标签之间的关系。但是在信息检索和非监督式学习中，我们更关注的是不同事物之间的相似程度。这就需要用到线性代数中的向量空间模型了。

提到向量空间模型，你可能对其中的概念有点陌生，所以我会从向量空间的基本概念开始说起，讲到向量空间模型的相关知识，最后再讲讲它是如何应用在不同的编程中的。

什么是向量空间？

上一节，我讲到了向量和向量空间的一些基本概念。为了帮助你更好地理解向量空间模型，我这里给出向量和向量空间的严格定义。

首先假设有一个数的集合 F，它满足“F 中任意两个数的加减乘除法（除数不为零）的结果仍然在这个 F 中”，我们就可以称 F 为一个“域”。我们处理的数据通常都是实数，所以这里我只考虑实数域。而如果域 F 里的元素都为实数，那么 F 就是实数域。

如果 x_1，x_2，……，$x_n \in F$，那么 F 上的 n 维向量就是：

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{bmatrix}$$

或者写成转置的形式：

$$\begin{bmatrix} x_1 & x_2 & x_3 & \dots & x_n \end{bmatrix}'$$

向量中第 i 个元素，也称为第 i 个分量。F_n 是由 F 上所有 n 维向量构成的集合。

我们已经介绍过向量之间的加法，以及标量和向量的乘法。这里我们使用这两个操作来定义向量空间。

假设 V 是 F_n 的非零子集，如果对任意的向量 x、向量 $y \in V$，都有 $(x+y) \in V$，我们称为 V 对向量的加法封闭；对任意的标量 $k \in V$，向量 $x \in V$，都有 kx 属于 V，我们称 V 对标量与向量的乘法封闭。

如果 V 满足向量的加法和乘法封闭性，我们就称 V 是 F 上的向量空间。向量空间除了满足这两个封闭性，还满足基本运算法则，比如交换律、结合律、分配律等等。这里介绍的定义和法则有点多，不过你可以不用都死记硬背下来。只要用的时候，知道有这些东西就可以了。

向量空间的几个重要概念

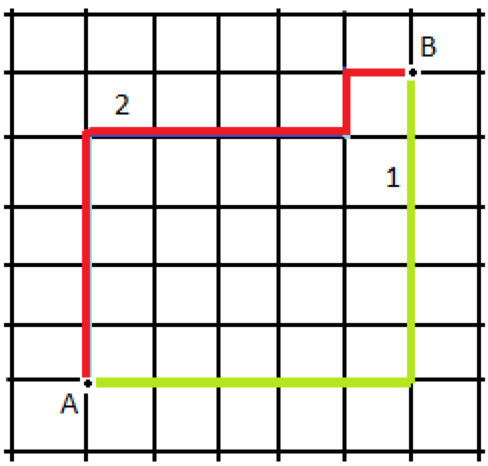
有了刚才的铺垫，接下来我们来看几个重要的概念：向量的长度、向量之间的距离和夹角。

向量之间的距离

有了向量空间，我们就可以定义向量之间的各种距离。我们之前说过，可以把一个向量想象为 n 维空间中的一个点。而向量空间中两个向量的距离，就是这两个向量所对应的点之间的距离。距离通常都是大于 0 的，这里我介绍几种常用的距离，包括曼哈顿距离、欧氏距离、切比雪夫距离和闵可夫斯基距离。

曼哈顿距离（Manhattan Distance）

这个距离度量的名字由来非常有趣。你可以想象一下，在美国人口稠密的曼哈顿地区，从一个十字路口开车到另外一个十字路口，驾驶距离是多少呢？当然不是两点之间的直线距离，因为你无法穿越挡在其中的高楼大厦。你只能驾车绕过这些建筑物，实际的驾驶距离就叫作曼哈顿距离。由于这些建筑物的排列都是规整划一的，形成了一个个的街区，所以我们也可以形象地称把它为“城市街区”距离。我这里画了张图方便你理解这种距离。



从图中可以看出，从 A 点到 B 点有多条路径，但是无论哪条，曼哈顿距离都是一样的。

在二维空间中，两个点（实际上就是二维向量）$x(x_1, x_2)$ 与 $y(y_1, y_2)$ 间的曼哈顿距离是：

$$MD(x,y)=\left|x_1-y_1\right|+\left|x_2-y_2\right|$$

推广到 n 维空间，曼哈顿距离的计算公式为：

$$MD(x,y)=\sum_{i=1}^{n}\left|x_i-y_i\right|$$

其中 n 表示向量维度，x_i 表示第一个向量的第 i 维元素的值，y_i 表示第二个向量的第 i 维元素的值。

欧氏距离（Euclidean Distance）

欧氏距离，其实就是欧几里得距离。欧氏距离是一个常用的距离定义，指在 n 维空间中两个点之间的真实距离，在二维空间中，两个点 $x(x_1, x_2)$ 与 $y(y_1, y_2)$ 间的欧氏距离是：

$$ED(x,y)=\sqrt{\left(x_1-y_1\right)^2+\left(x_2-y_2\right)^2}$$

推广到 n 维空间，欧氏距离的计算公式为：

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

切比雪夫距离（Chebyshev Distance）

切比雪夫其实是在模拟国际象棋里国王的走法。国王可以走临近 8 个格子里的任何一个，那么国王从格子 (x_1, x_2) 走到格子 (y_1, y_2) 最少需要多少步呢？其实就是二维空间里的切比雪夫距离。

一开始，为了走尽量少的步数，国王走的一定是斜线，所以横轴和纵轴方向都会减 1，直到国王的位置和目标位置在某个轴上没有差距，这个时候就改为沿另一个轴每次减 1。所以，国王走的最少格子数是 $|x_1 - y_1|$ 和 $|x_2 - y_2|$ 这两者的较大者。所以，在二维空间中，两个点 $x(x_1, x_2)$ 与 $y(y_1, y_2)$ 间的切比雪夫距离是：

$$CD(x,y)=\max(|x_1-y_1|,|x_2-y_2|)$$

推广到 n 维空间，切比雪夫距离的计算公式为：

$$CD(x,y)=\arg\max_{i=1}^{n}|x_i-y_i|$$

上述三种距离，都可以用一种通用的形式表示，那就是闵可夫斯基距离，也叫闵氏距离。在二维空间中，两个点 $x(x_1, x_2)$ 与 $y(y_1, y_2)$ 间的闵氏距离是：

$$MKD(x,y)=\sqrt[p]{|x_1-y_1|^p+|x_2-y_2|^p}$$

两个 n 维变量 $x(x_1, x_2, \ldots, x_n)$ 与 $y(y_1, y_2, \ldots, y_n)$ 间的闵氏距离的定义为：

$$MKD(x,y)=\sqrt[p]{\sum_{i=1}^{n}|x_i-y_i|^p}$$

其中 p 是一个变参数，尝试不同的 p 取值，你就会发现：

- 当 $p=1$ 时，就是曼哈顿距离；
- 当 $p=2$ 时，就是欧氏距离；
- 当 p 趋近于无穷大的时候，就是切比雪夫距离。这是因为当 p 趋近于无穷大的时候，最大的 $|x_i-y_i|$ 会占到全部的权重。

距离可以描述不同向量在向量空间中的差异，所以可以用于描述向量所代表的事物之差异（或相似）程度。

向量的长度

有了向量距离的定义，向量的长度就很容易理解了。向量的长度，也叫向量的模，是向量所对应的点到空间原点的距离。通常我们使用欧氏距离来表示向量的长度。

当然，我们也可以使用其他类型的距离。说到这里，我也提一下“范数”的概念。范数满足非负性、齐次性、和三角不等式。你可以不用深究这三点的含义，不过你需要知道范数常常被用来衡量某个向量空间中向量的大小或者长度。

L_1 范数 $||x||$ ，它是为 x 向量各个元素绝对值之和，对应于向量 x 和原点之间的曼哈顿距离。

L_2 范数 $||x||_2$ ，它是 x 向量各个元素平方和的 $\frac{1}{2}$ 次方，对应于向量 x 和原点之间的欧氏距离。

L_p 范数 $||x||_p$ ，为 x 向量各个元素绝对值 p 次方和的 1/p 次方，对应于向量 x 和原点之间的闵氏距离。

L_∞ 范数 $||x||_\infty$ ，为 x 向量各个元素绝对值最大那个元素的绝对值，对应于向量 x 和原点之间的切比雪夫距离。

所以，在讨论向量的长度时，我们需要弄清楚是 L 几范数。

向量之间的夹角

在理解了向量间的距离和向量的长度之后，我们就可以引出向量夹角的余弦，它计算了空间中两个向量所形成夹角的余弦值，具体的计算公式我列在了下面：

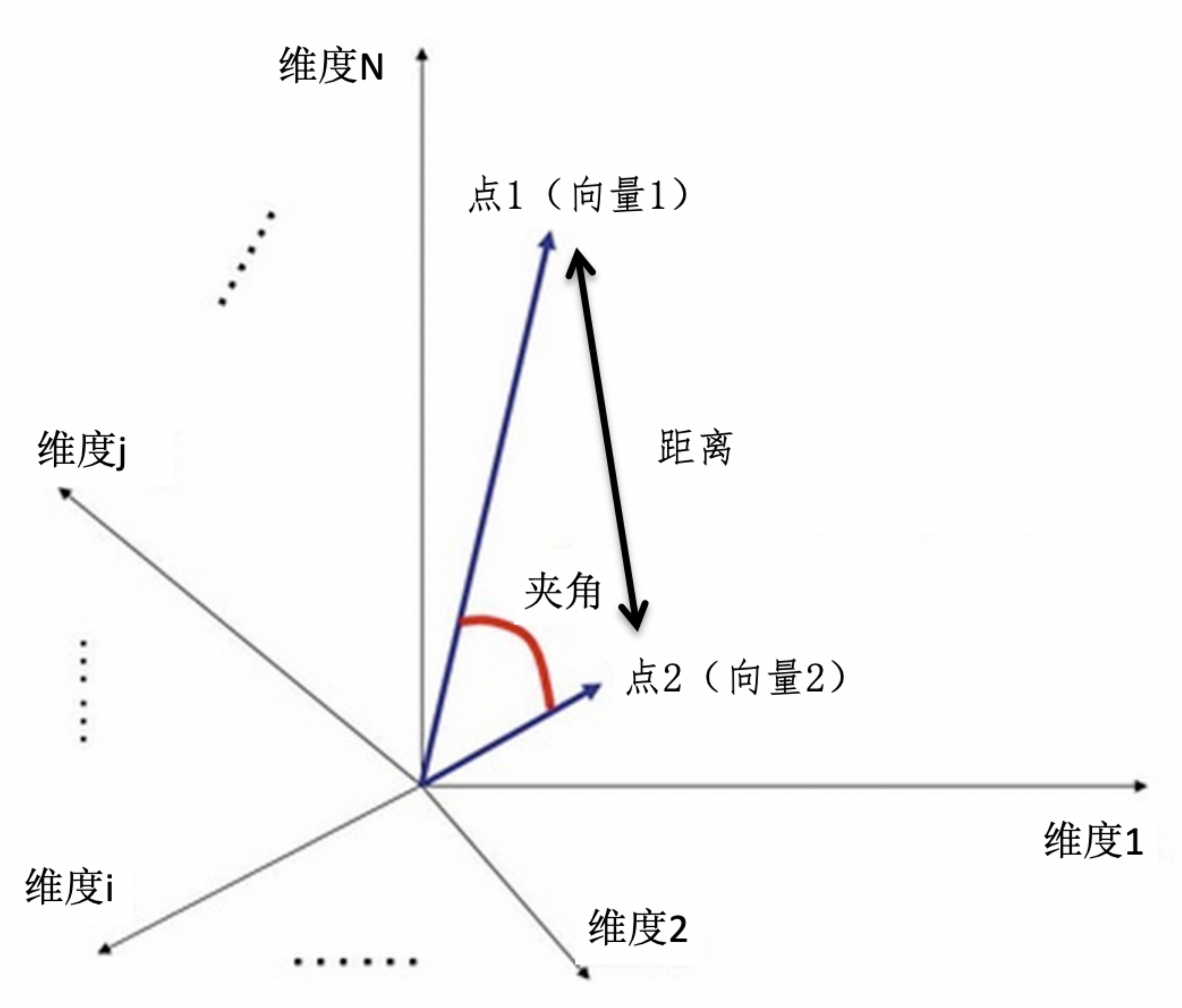
$$\mathrm{Cos}ine(X,Y)=\frac{\sum_{i=1}^{n}(x_i \times y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2 \times \sum_{i=1}^{n}y_i^2}}$$

从公式可以看出，分子是两个向量的点乘，而分母是两者长度（或 L2 范数）的乘积，而 L2 范数可以使用向量点乘自身的转置来实现。夹角余弦的取值范围在 [-1,1]，当两个向量的方向重合时夹角余弦取最大值 1，当两个向量的方向完全相反夹角余弦取最小值 -1。值越大，说明夹角越小，两点相距就越近；值越小，说明夹角越大，两点相距就越远。

向量空间模型

理解了向量间距离和夹角余弦这两个概念，你再来看**向量空间模型**（Vector Space Model）就不难了。

向量空间模型假设所有的对象都可以转化为向量，然后使用向量间的距离（通常是欧氏距离）或者是向量间的夹角余弦来表示两个对象之间的相似程度。我使用下图来展示空间中向量之间的距离和夹角。



由于夹角余弦的取值范围已经在 -1 到 1 之间，而且越大表示越相似，所以可以直接作为相似度的取值。相对于夹角余弦，欧氏距离 ED 的取值范围可能很大，而且和相似度呈现反比关系，所以通常要进行 1/(1-ED) 这种归一化。

当 ED 为 0 的时候，变化后的值就是 1，表示相似度为 1，完全相同。当 ED 趋向于无穷大的时候，变化后的值就是 0，表示相似度为 0，完全不同。所以，这个变化后的值，取值范围是 0 到 1 之间，而且和相似度呈现正比关系。

早在上世纪的 70 年代，人们把向量空间模型运用于信息检索领域。由于向量空间可以很形象地表示数据点之间的相似程度，因此现在我们也常常把这个模型运用在基于相似度的一些

机器学习算法中，例如 K 近邻（KNN）分类、K 均值（K-Means) 聚类等等。

总结

为了让计算机理解现实世界中的事物，我们会把事物的特点转换成为数据，并使用多维度的特征来表示某个具体的对象。多个维度的特征很容易构成向量，因此我们就可以充分利用向量和向量空间，来刻画事物以及它们之间的关系。

我们可以在向量空间中定义多种类型的向量长度和向量间距离，用于衡量向量之间的差异或者说相似程度。此外，夹角余弦也是常用的相似度衡量指标。和距离相比，夹角余弦的取值已经控制在 [-1, 1] 的范围内，不会因为异常点所产生的过大距离而受到干扰。

向量空间模型充分利用了空间中向量的距离和夹角特性，来描述文档和查询之间的相似程度，或者说相关性。虽然向量空间模型来自信息检索领域，但是也被用广泛运用在机器学习领域中。在接下来的文章里，我会结合具体的案例，分别来说说如何在这些领域使用向量空间模型。

思考题

假设在三维空间中有两个点，它们的坐标分别是 (3, -1, 8) 和 (-2, 3, -6)，请计算这两个点之间的欧氏距离和夹角余弦。

欢迎留言和我分享，也欢迎你在留言区写下今天的学习笔记。你可以点击“请朋友读”，把今天的内容分享给你的好友，和他一起精进。

上一篇 33 | 线性代数：线性代数到底都讲了些什么？

下一篇 35 | 文本检索：如何让计算机处理自然语言？

精选留言 (8)

写留言

mickey
2019-03-05
1

欧氏距离：√237
夹角余弦：-57/√(74*49)

作者回复: 正确👌

Wing·三金
2019-03-04
1

欧式距离的平方=25+16+196=237

欧式距离为根号 237≈15.4

cos=（-6-3-48）/（√（9+1+64）*√（4+9+36））=（-57）/（7*√74）≈-0.95

...

展开

作者回复: 对 MinMax 也是可以的。不过这里是1除以(1-ED)，所以不会出现负无穷大。而最大的值也不会超过1。

cugphoenix

2019-05-01

欧式距离1/(1-ED)是如何归一到0-1区间的？ED<=1怎么办？

Monica

2019-04-17

欧式距离=√（x1-y1）²+（x2-y2）²+（x3-y3）²=√（3+2）²+（-1-3）²+（8+6）²=√237

夹角余弦 = 向量点积/ (向量长度的叉积) =（x1x2 + y1y2 + z1z2) / (√(x1²+y1²+z1²) x √(x2²+y2²+z2²))=(-6-3-48)/（√（9+1+64）*√（4+9+36））=（-57）/（7*√74）

展开

qinggeouy...

2019-03-18

向量 a = [3, -1, 8] ，b = [-2, 3, -6]

欧氏距离 ED(a, b) = 开根号(25+16+196)

夹角余弦 Cosine(a, b) = (-6 -3 -48)/开根号(74*49)

展开

作者回复: 对👌

YiFān.W

2019-03-13

那这个向量应当包括字典中所有的词条吧？实际情况中岂不是非常非常大

作者回复: 是的，这个向量维数很多，在实际应用中，我们可以使用降维、倒排索引等措施来提高效率。后面也会介绍

Joe
2019-03-12

切比雪夫的公式写错了吧，不是x y之间相减，应该是x和x，y和y之间差的最大值。

作者回复: 这里x和y分别表示两个向量，例如[x1, x2, x3, x4...xn], [y1, y2, y3, y4..yn]，和二维坐标的表示(x1,y1) (x2,y2)有所不同。

李皮皮皮皮...
2019-03-04

V是Fn的子集，Fn是F中的n维向量。那怎么会有标量属于V呢？不太明白😥

作者回复: 好问题，V是针对向量，标量不受限