0、概述

机器学习中的度量包括四种度量：一是度量两个样本之间的距离（或相似度）；二是度量两特征之间的相关性；三是度量两概率分布之间的相似性；四是度量两集合间相似性。

一、样本间距离度量

在机器学习中，很多模型和算法都基于一个前提：相似的输入具有相似的输出，这个“相似”就是指距离较近。因此，样本间的距离的度量是非常重要的。

样本的特征按照其取值是否具有大小的意义，分为“有序特征”和“无序特征”，有序特征指的是其特征的取值具有大小顺序的意义，无序特征指的是其特征的取值没有大小顺序的意义。有序特征可以是连续特征，也可以是离散特征，而无序特征只能是离散特征。

显然，有序特征和无序特征的距离的定义方式不同。

通常，若样本的特征是有序特征，我们使用欧氏距离、曼哈顿距离、闵可夫斯基距离或余弦相似度等作为样本间距离。

（1）欧氏距离：



（2）曼哈顿距离：



（3）闵可夫斯基距离：



可以看出，欧氏距离和曼哈顿距离都是闵可夫斯基距离的特例，欧氏距离是闵可夫斯基距离p=2的情况，曼哈顿距离是p=1的情况。

（4）余弦相似度：

一个样本可以看成是由多维特征构成的特征向量，因此样本间的距离可以用向量间的余弦来衡量，这就是余弦相似度。

设两个样本和，它们的余弦相似度为：



注意，两样本的特征向量的余弦相似度越高，样本的距离越小。

当样本的不同特征的“重要性”不同时，可以使用“加权距离”，比如加权的闵可夫斯基距离：



其中特征的权重衡量了特征的重要性。

若样本的特征为“无序特征”，乍一看似乎不存在距离的意义，但实际上也可以有合适的定义，不过要借助样本的簇（或类别）的划分。常用的无序特征的距离是VDM（Value Difference Metric）距离，其定义如下：

设样本集被划分为k个簇（或类别），样本的任意一个无序特征u上有任意两个离散取值a和b，则a和b的VDM距离为：



其中（）表示特征u上取值为a（b）的总样本数，（）表示第i个簇中特征u上取值为a（b）的样本数。

可以看出，VDM距离的意义是，它认为无序特征的某个取值在不同簇中出现的次数（也就是该取值关于簇的分布），代表了该取值对于样本的簇的划分的作用，两个取值关于簇的分布越一致，认为这两个取值越接近。极端情况下，两个取值关于簇的分布完全一致，则这两个取值的距离为0，也就是对于样本的簇的划分来说，两个取值的作用一样，完全可以用一个取值代替另一个取值。

二、特征间相关性度量

样本的特征本质上是随机变量，两特征间的相关性就相当于两个随机变量的相关性，它一般用皮尔逊相关系数来度量。

皮尔逊相关系数的定义为：



其中是随机变量X和Y的协方差。

在训练集上计算皮尔逊相关系数的公式为：



对比皮尔逊相关系数与余弦相似度的计算公式，可以发现，两个向量的余弦相似度相当于两个向量零均值化后的皮尔逊相关系数。通常我们觉得，“样本之间的相似度”与“特征之间的相关性”应该是风马牛不相及的，但此处我们看到了两者的共性。但是要注意，余弦相似度本身就是针对两个向量进行定义并进行计算的，而皮尔逊相关系数是针对两个随机变量进行定义的，只不过其计算是使用了随机变量的取值构成的向量，这其实是一种估计。

三、概率分布间相似性度量

通常使用KL散度（又叫相对熵）来度量两个概率分布之间的相似度。设对随机变量X，有两个分布p和q，则它们的KL散度为：



其中为X的一个取值，N为X的取值个数。

KL散度越大，则p和q“距离”越大，即p和q越不相似；KL散度越小，则p和q距离越小，即p和q越相似。

四、集合间相似性的度量

通常使用杰卡德距离来度量两集合间的相似度。设两个集合A和B，它们的杰卡德距离为：

