# 机器学习距离度量

| 分类 | 方法 |
| --- | --- |
| 概率分布之间的距离 | 信息熵 (Information Entropy) |
| | KL 散度 (KL-Divergence) |
| | 卡方检验 (Chi-Square) |
| | 互信息（mutual information） |
| 闵氏距离 | 曼哈顿距离 (Manhattan Distance) |
| | 欧氏距离 (Euclidean Distance) |
| | 切比雪夫距离 (Chebyshev Distance) |
| 马氏距离 | 马氏距离 (Mahalanobis Distance) |
| 余弦距离（方向） | 余弦距离 (Cosine Distance) |
| 编辑距离（String Metric） | 汉明距离 (Hamming Distance) |
| | 莱文斯坦距离 (Hamming Distance) |
| | DL 距离 （Damerau–Levenshtein Distance |
| | 最长公共子序列（LCS） |
| 集合之间距离（Set Metric） | 雅卡尔距离 (Jaccard Distance) |
| | 简单匹配系数（SMC) |
| | Dice系数 |
| 相关（Correlation） | 皮尔逊积矩相关系数 （Pearson's r） |
| | 斯皮尔曼等级相关系数(待补充) |

## 度量的基本知识

**集中趋势**： 平均数，中位数，众数　　**离散程度**： 全距，标准差，变异系数，百分位数，四分差，四分位数，方差，标准分数，切比雪夫不等式

**方差**：是标准差的平方，而标准差的意义是数据集中各个点到均值点距离的平均值，**反应的是数据的离散程度。**　　**协方差**：标准差与方差是描述一维数据的，当存在多维数据时，我们通常需要知道每个维数的变量中间是否存在关联。**协方差就是衡量多维数据集中，变量之间相关性的统计量**。比如说，一个人的身高与他的体重的关系，这就需要用协方差来衡量。如果两个变量之间的**协方差为正值，则这两个变量之间存在正相关**，若为**负值，则为负相关**。

# 1 概率分布之间的距离

在统计学里面经常需要测量两组样本分布之间的距离，进而判断出它们是否出自同一个population，常见的方法有卡方检验（Chi-Square）和 KL 散度（ KL-Divergence）　　信息熵、交叉熵、条件熵、自信息、互信息

## 1.1. 信息熵(Information Entropy)

熵（英语：entropy）是接收的每条消息中包含的信息的平均量，又被称为信息熵、信源熵、平均自信息量。信息熵描述的是整个系统内部样本之间的一个距离，或者称之为系统内样本分布的集中程度（一致程度）、分散程度、混乱程度（不一致程度）。**系统内样本分布越分散(或者说分布越平均)，信息熵就越大。分布越有序（或者说分布越集中），信息熵就越小。**

$H(X)=\sum_{i=1}^{n}-p_i\log_2 p_i$

信息熵越大表明样本集S的分布越分散（分布均衡），信息熵越小则表明样本集X的分布越集中（分布不均衡）。当S中n个分类出现的概率一样大时（都是1/n），信息熵取最大值log2(n)。当X只有一个分类时，信息熵取最小值0

## 1.2. KL散度 （相对熵）

信息散度（information divergence），信息增益（information gain）**是衡量两个分布(P、Q)之间的距离；越小越相似。**　　KL散度是两个概率分布P和Q差别的非对称性的度量。 KL散度是用来度量使用基于Q的编码来编码来自P的样本平均所需的额外的位元数。 典型情况下，P表示数据的真实分布，Q表示数据的理论分布，模型分布，或P的近似分布。

$$D_{KL}(P||Q)=-\sum_i P(i)\ln\frac{Q(i)}{P(i)}=\sum_i P(i)\ln\frac{P(i)}{Q(i)}$$

相对熵可以衡量两个随机分布之间的距离，当两个随机分布相同时，它们的相对熵为零，当两个随机分布的差别增大时，它们的相对熵也会增大。所以相对熵（KL散度）可以用于比较文本的相似度，先统计出词的频率，然后计算相对熵。另外，在多指标系统评估中，指标权重分配是一个重点和难点，也通过相对熵可以处理。

## 1.3. 卡方检验

## 1.4. 互信息 （mutual information）

在概率论和信息论中，两个随机变量的互信息（Mutual Information，简称MI）或转移信息（trans information）是**变量间相互依赖性**的量度。不同于相关系数，互信息并不局限于实值随机变量，它更加一般且决定着联合分布 p(X,Y) 和分解的边缘分布的乘积 p(X)p(Y) 的相似程度。互信息是点间互信息（PMI）的期望值。互信息最常用的单位是bit。

$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right)$

# 2 闵可夫斯基距离 (Minkowski Distance)

闵式距离不是一种距离，而是一组距离的定义，是对多个距离度量公式的概括性的表述。闵式距离的定义如下： 两个$n$维变量 $a = (x_1, x_2, \ldots, x_n)$ 与 $b = (y_1, y_2, \ldots, y_n)$ 之间的闵可夫斯基距离定义为 $d_{ab} = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$

其中，$p$是一个变量：

- $p = 1$, 是曼哈顿距离；
- $p = 2$, 是欧式距离；
- $p \to \infty$, 是切比雪夫距离。

闵式距离主要缺点：

1. 将各个分量的量纲(scale)，也就是“单位”当作相同的看待了。
2. **没有考虑各个分量的分布**（期望，方差等)可能是不同的。

## 2.1. 曼哈顿距离 (Manhattan Distance)

顾名思义，在曼哈顿街区要从一个十字路口开车到另一个十字路口，驾驶距离显然不是两点间的直线距离。这个实际驾驶距离就是“曼哈顿距离”。曼哈顿距离也称为“城市街区距离”(City Block distance)。

- 二维平面两点a(x_1,y_1)与b(x_2,y_2)间的曼哈顿距离：
  $d_{a,b} = |x_1 - x_2| + |y_1 - y_2|$
- n维空间点$a(x_1^1, x_2^1, \ldots, x_n^1)$与 $b(x_1^2, x_2^2, \ldots, x_n^2)$的曼哈顿距离：
  $d_{a,b} = \sum_{i=i}^{n} |x_i^1 - x_i^2|$

## 2.2. 欧氏距离(Euclidean Distance)

在数学中，欧几里得距离或欧几里得度量是欧几里得空间中两点间“普通”（即直线）距离。使用这个距离，欧氏空间成为度量空间。相关联的范数称为欧几里得范数。较早的文献称之为毕达哥拉斯度量。 $d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$

## 2.3. 切比雪夫距离 (Chebyshev Distance)

数学上，切比雪夫距离（Chebyshev distance）或是$L_\infty$度量是向量空间中的一种度量，二个点之间的距离定义为其各座标数值差的最大值。以$(x_1, y_1)$和 $(x_2, y_2)$二点为例，其切比雪夫距离为 $max(|x_2 - x_1|, |y_2 - y_1|)$ 以任一点为准，和此点切比雪夫距离为$r$的点会形成一个正方形，其边长为$2r$，且各边都和坐标轴平行。

国际象棋中，国王可以直行、横行、斜行，所以国王走一步可以移动到相邻8个方格中的任意一个。国王从格子(x1,y1)走到格子(x2,y2)最少需要多少步？这个距离就叫切比雪夫距离。

- 两个点 $a=(x_1,x_2,\ldots,x_n)$ ，$b=(y_1,y_2,\ldots,y_n)$
  $D_{chess}=\max_i|x_i-y_i|$

## 2.4 标准化欧氏距离 (Standardized Euclidean Distance)

标准化欧氏距离是针对欧氏距离的缺点而作的一种改进。标准欧氏距离的思路：既然数据各维分量的分布不一样，那先将各个分量都“标准化”到均值、方差相等。

1. 变量的标准化过程，假设样本集X的均值(mean)为m，标准差(standarddeviation)为s，那么X的“标准化变量”表示为： $X^*=\frac{X-m}{s}$
   - 标准化变量之后的数学期望为$E(x)=0$,方差$D(x)=1$.
2. 样本集标准化过程描述为： $\text{标准化之后的值}=\frac{\text{标准化前的值}-\text{分量的均值}}{\text{分量的标准差}}$
3. 标准化欧式距离
   - 对于标准化之后的向量 $a'=(x_1',x_2',\ldots,x_n')$ 和 $b'=(y_1',y_2',\ldots,y_n')$
     $d_{a'b'}=\sqrt{a'^2+b'^2}$
   - 对于原始变量 $a=(x_1,x_2,\ldots,x_n)$ 和 $b=(y_1,y_2,\ldots,y_n)$
     $d_{ab}=\sqrt{\sum_{i=1}^n(\frac{x_i-y_i}{s_i})^2}$

# 3 马氏距离(Mahalanobis Distance)

马哈拉诺比斯距离是由印度统计学家马哈拉诺比斯提出的，表示数据的协方差距离。它是一种有效的计算两个未知样本集的相似度的方法。与欧氏距离不同的是它考虑到各种特性之间的联系（例如：一条关于身高的信息会带来一条关于体重的信息，因为两者是有关联的）并且是尺度无关的（scale-invariant），即独立于测量尺度。

*定义*：对于一个均值为 $\mu=(\mu_1,\mu_2,\mu_3,\ldots,\mu_p)^T$ 协方差矩阵为 $\Sigma$的多变量向量 $x=(x_1,x_2,x_3,\ldots,x_p)^T$其马氏距离为: $D_m(x)=\sqrt{(x-\mu)^T\Sigma^{-1}(x-\mu)}$ 马哈拉诺比斯距离也可以定义为两个服从同一分布并且其协方差矩阵为 $\Sigma$ 的随机变量 $\vec{x}$与 $\vec{y}$的差异程度：
$d_{\vec{x},\vec{y}}=\sqrt{(\vec{x}-\vec{y})\Sigma^{-1}(\vec{x}-\vec{y})}$

**马氏距离特点：**

1. 尺度不变性:（a）两点之间的马氏距离与原始数据的测量单位无关，(b) 标准化数据和中心化数据(即原始数据与均值之差）计算出的二点之间的马氏距离相同。
2. 可以排除变量之间的相关性的干扰
3. 考虑到数据的分布
4. 满足距离的四个基本公理：非负性、自反性、对称性和三角不等式。
5. 马氏距离的计算是建立在总体样本的基础上的，这一点可以从上述协方差矩阵的解释中可以得出，也就是说，如果拿同样的两个样本，放入两个不同的总体中，最后计算得出的两个样本间的马氏距离通常是不相同的，除非这两个总体的协方差矩阵碰巧相同；
6. 在计算马氏距离过程中，要求总体样本数大于样本的维数，否则得到的总体样本协方差矩阵逆矩阵不存在，这种情况下，用欧式距离计算即可。

| 马氏距离 | 欧式距离 |
| --- | --- |
| 建立在总样本基础上 | 建立在有单位的坐标系上 |
| 必须求得协方差矩阵 | 有坐标即可计算 |
| 更加明显的反应样本元素的距离 | 反应一般距离 |
| 排除变量之间的相关性的干扰，考虑分布 | 不考虑分布 |

# 4 余弦距离（Cosine Distance）

余弦距离也称余弦相似性：通过测量两个向量的夹角的余弦值来度量它们之间的相似性。0度角的余弦值是1，而其他任何角度的余弦值都不大于1；并且其最小值是-1。从而两个向量之间的角度的余弦值确定两个向量是否大致指向相同的方向。**这结果是与向量的长度无关的，仅仅与向量的指向方向相关**。

- 几何中，夹角余弦可用来衡量两个向量方向的差异；
- 机器学习中，借用这一概念来衡量样本向量之间的差异；
- 它通常用于文本挖掘中的文件比较；
- 在数据挖掘领域中，会用到它来度量集群内部的凝聚力。

*定义：* 给定两个向量 $a=(x_1,x_2,\ldots,x_n)$ 与 $b=(y_1,y_2,\ldots,y_n)$ 余弦距离表示为：

$$\begin{aligned}\cos(\theta) =& \frac{a\cdot b}{||a||*||b||}\\ =& \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}\end{aligned}$$

夹角余弦取值范围为$[-1,1]$。余弦越大表示两个向量的夹角越小，余弦越小表示两向量的夹角越大。当两个向量的方向重合时余弦取最大值1，当两个向量的方向完全相反余弦取最小值-1。

# 5 编辑距离

## 5.1. 汉明距离 (Hamming Distance)

*定义:* 在信息论中，两个等长字符串之间的汉明距离（Hamming distance）是两个字符串对应位置的不同字符的个数。

**汉明重量**：是字符串相对于同样长度的零字符串的汉明距离，也就是说，它是字符串中非零的元素个数：对于二进制字符串来说，就是 1 的个数，所以 11101 的汉明重量是 4。因此，如果向量空间中的元素a和b之间的汉明距离等于它们汉明重量的差a-b。

**应用**：汉明重量分析在包括信息论、编码理论、密码学等领域都有应用。比如在信息编码过程中，为了增强容错性，应使得编码间的最小汉明距离尽可能大。但是，如果要比较两个不同长度的字符串，不仅要进行替换，而且要进行插入与删除的运算，在这种场合下，通常使用更加复杂的编辑距离等算法。

## 5.2. 莱文斯坦距离（Levenshtein Distance）

莱文斯坦距离，又称Levenshtein距离，是编辑距离的一种。指两个字串之间，由一个转成另一个所需的最少编辑操作次数。允许的编辑操作包括将一个字符**替换**成另一个字符，**插入**一个字符，**删除**一个字符。

例如将kitten $\rightarrow$ 字转成sitting：

1. sitten （k→s）
2. sittin （e→i）
3. sitting （→g）

**应用：** DNA分析，拼写检查，语音辨识，抄袭侦测

## 5.3 DL 距离 （Damerau–Levenshtein Distance）

Damerau–Levenshtein Distance用来测量两个字符序列之间的编辑距离的字符串度量标准。两个词的Damerau–Levenshtein Distance是从一个词转换为另一个词的最少操作数，与Levenshtein Distance不同的是，除了单个字符的插入、删除和变更之外，还包括两个**相邻字符的转换**。

## 5.4 最长公共子序列 （Longest Common Subsequences）

最长公共子序列（LCS）是一个在一个序列集合中（通常为两个序列）用来查找所有序列中最长子序列的问题。这与查找最长公共子串的问题不同的地方是：**子序列不需要在原序列中占用连续的位置** 。最长公共子序列问题是一个经典的计算机科学问题，也是数据比较程序，比如Diff工具，和生物信息学应用的基础。它也被广泛地应用在版本控制，比如Git用来调和文件之间的改动。

*定义:* 一个数列 $S$，如果分别是两个或多个已知数列的子序列，且是所有匹配此条件序列中最长的，则$S$称为已知序列的最长公共子序列。

**复杂度**:对于一般性的LCS问题（即任意数量的序列）是属于NP-hard。但当序列的数量确定时，问题可以使用动态规划（Dynamic Programming）在多项式时间内解决。

# 6 集合距离 （Set Metric）

## 6.1. 雅卡尔距离 (Jaccard Distance)

**雅卡尔指数**（Jaccard index），又称为并交比（Intersection over Union）、雅卡尔相似系数（Jaccard similarity coefficient），是用于比较样本集的相似性与多样性的统计量。雅卡尔系数能够量度有限样本集合的相似度，其定义为两个集合交集大小与并集大小之间的比例：
$J(A,B)=\frac{|A\bigcap B|}{|A\bigcup B|}$ 如果A与B完全重合，则定义$J(A,B)=1$, 于是有$0\leq J(A,B)\leq 1$。 **雅卡尔距离**（Jaccard distance）则用于量度样本集之间的不相似度，其定义为1减去雅卡尔系数，即 $d_J=1-J(A,B)$

其他类似距离：简单匹配系数，汉明距离，Dice系数

## 6.2. 简单匹配系数 (Simple Matching Coefficient)

简单匹配系数（英语：simple matching coefficient，缩写SMC），又称为兰德相似系数（Rand similarity coefficient），是用于比较样本信合之间相似性与多样性的统计量。 $SMC=\frac{匹配属性数量}{属性总数}$

类似地，可以定义简单匹配距离（simple matching distance，缩写SMD）为 $1-SMC$，用于量度样本集合间的不相似度。

SMC与**汉明相似度**间呈线性关系 $SMC=(Hamann+1)/2$。

## 6.3. Dice系数

根据 Lee Raymond Dice命名，是一种集合相似度度量函数，通常用于计算两个样本的相似度：$s=\frac{2|A\bigcap B|}{|A|+|B|}$ 它在形式上和Jaccard指数没多大区别，但是有些不同的性质。和Jaccard类似，它的范围为0到1。 与Jaccard不同的是，相应的差异函数 $1-s$ 不是一个合适的距离度量措施，因为它没有三角形不等性的性质。

**应用**： 在信息检索中, 给定关键词集合X 和Y ， 相似度定义为两倍的共同信息(重叠部分)除以基数的总和.

# 7 相关 （Correlation）

在概率论和统计学中，相关（Correlation），显示两个随机变量之间**线性关系的强度和方向**。在统计学中，相关的意义是用来衡量两个变量相对于其相互独立的距离。在这个广义的定义下，有许多根据数据特点而定义的用来衡量数据相关的系数。

## 7.1 皮尔逊积矩相关系数 （Pearson's r）

在统计学中，皮尔逊积矩相关系数（Pearson product-moment correlation coefficient，又称作PPMCC或PCCs, 文章中常用r或Pearson's r表示）用于度量两个变量X和Y之间的相关（线性相关，其值介于-1与1之间。在自然科学领域中，该系数广泛用于度量两个变量之间的相关程度。它是由卡尔·皮尔逊从弗朗西斯·高尔顿在19世纪80年代提出的一个相似却又稍有不同的想法演变而来。这个相关系数也称作"皮尔森相关系数r"。

*定义*:两个变量之间的皮尔逊相关系数定义为两个变量之间的**协方差**和**标准差**的商：

- 总体相关系数 $\rho$
  $\rho_{X,Y}=\frac{cov(X,Y)}{\sigma_X,\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}$
- 样本相关系数 $r$
  $r=\frac{\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i-\bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i-\bar{Y})^2}}$

相关距离： $D_{X,Y}=1-\rho_{XY}$　　皮尔逊相关系数具有平移不变性和尺度不变性，计算出了两个向量（维度）的相关性。不过，一般我们在谈论相关系数的时候，将 x 与 y 对应位置的两个数值看作一个样本点，皮尔逊系数用来表示这些样本点分布的相关性。　　由于皮尔逊系数具有的良好性质，在各个领域都应用广泛，例如，在推荐系统根据为某一用户查找喜好相似的用户,进而提供推荐，优点是可以不受每个用户评分标准不同和观看影片数量不一样的影响。

在数据标准化后，Pearson相关性系数、Cosine 相似度、欧式距离的平方可以认为是等价的。