数据挖掘原理与应用

第二章-认识数据

叶志鹏

南京理工大学泰州科技学院

1st Jan, 2023

1. 数据基本概念
2. 数据的统计特征
3. Iris 可视化案例
4. 数据的距离与相似性
5. 参考文献

1 数据基本概念
- 数据对象
- 数据属性
- 属性的类型
- 属性类型的对比
- 离散属性与连续属性

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

1 数据基本概念
数据对象
数据属性
属性的类型
属性类型的对比
离散属性与连续属性
2 数据的统计特征
3 Iris 可视化案例
4 数据的距离与相似性
5 参考文献

数据对象

数据对象也称样品、示例、实例、数据点、对象和元组。例如，一个人或一部车都可以被认为是数据对象，在某种意义上它们可以用一组属性来定义。乳腺癌数据集

1 数据基本概念
- 数据对象
- 数据属性
- 属性的类型
- 属性类型的对比
- 离散属性与连续属性

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

数据属性

数据属性是一个数据字段，代表一个数据对象的特征或功能，属性（attribute）、维度（dimension）、特征（feature）、变量（variance）可以互换使用。

1 数据基本概念
- 数据对象
- 数据属性
- 属性的类型
- 属性类型的对比
- 离散属性与连续属性

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

属性的类型

数据属性有不同类型，包括标称属性 (nominal attribute)、二元属性 (binary attribute)、序数属性 (ordinal attribute) 和数值属性 (numerical attribute)

- 标称属性: 变成属性与“名称”有关。标称属性的值是一些符号或实物的名称，每个值代表某种类别、编码或状态。不必具有顺序性，并且不定量。尽管标称属性的值是一些符号或“事物名称”，但也可以用数学表示这些符号或名称。例如，对于 hair_color, 可以用 0 表示黑色，1 表示黄色。标称属性中最常出现的值为众数，可以作为中心化趋势度量。

属性的类型

- 二元属性：一种特殊的标称属性，只有两个类别或状态：0 和 1，其中 0 表示不出现，1 表示出现。如果将 0 和 1 对应于布尔值 (false, true)，则称为布尔属性。例如，属性 smoker 表示患者对象，1 表示患者抽烟，0 表示患者不抽烟。如果它的两个状态同等数据价值，则称为对称的二元属性如 gender 的两种状态（男，女）。非对称的二元属性如流感病毒化验结果（阳性，阴性），通常在数据挖掘中用 1 代表最重要的结果（稀有）编码，0 表示通常的状态。

属性的类型

- 序数属性：序数属性可能的取值之间具有有意义的序或秩评定，但相继值得差是未知得。例如，学生得成绩包括优，良，中，差四个等级。快餐店的饮料杯具有大，中，小三个可能值，而“大”比“中”大多少是未知的。通常数据预处理中的数据规约，序数属性可以通过将数据的值域划分成有限个有序类别，通过将数值属性离散化得到。可以用众数和中位数表示序数属性的中性趋势，但不能定义均值。
- 标注、二元和序数属性都是定性的，只描述样本的特征，而不给出实际大小或数量。

数值属性

- 数值属性是可度量的量，用整数或实数值表示，分为区间标度属性和比率标度属性两种类型。
  1. 区间标度属性（interval-scaled）：用相等的单位尺度度量。区间属性的值有序。所以，除了秩评定之外，这种属性允许比较和定理评估值之间的差。例如，身高是区间标度属性。假设我们有一个班学生的身高统计值，将每一个人视为一个样本，学生的身高属性，可以量化不同值之间的差。而对于温度（摄氏温度与华氏温度）是一种区间标度属性，但不能比较倍数。比如，我们不会说 10°C 比 5°C 温暖 2 倍，因为摄氏温度与华氏温度，没有绝对零度。
  2. 比率标度属性：可以用比率来描述两个值，即一个值是另一个值的倍数，也可以计算两者之差。例如，开氏温度，具有绝对零点。在零点，构成物质的粒子具有零动能。比率标度属性的例子还包括字数和工龄等计数属性。

1 数据基本概念
- 数据对象
- 数据属性
- 属性的类型
- 属性类型的对比
- 离散属性与连续属性

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

属性类型的对比

属性当型		描 述	例 子	操 作
分类的 (完生傷)	标称	标称属性的值仅仅只是不同的名字，即标称值只提供足够的信息以区分对象 $(=, \neq)$	邮政编码、雇员ID号、眼球颜性别	众数、熵、列联相关、χ^2 检验
	序数	序数属性的值提供足够的信息确对象的序 $(<, >)$	矿石硬度、好，较好，最好}、成绩、街道号码	中值、百分位、秩相关、游程检验、符号检验
数值的 (定彙的)	区间	对于区间属性，值之间的差是有意义的，即存在测量坦位 $(++, -)$	日历日期、摄氏或华氏温度	均值、标准差、皮尔逊相关、 t 和F检验
	比率	对于比率变量，差和比率都是有意义的 $(+, -, ^*, I)$	绝对温度、货币量、计数、年龄、质量、长度、电流	几何平均、调和平均、百分比变差

图 1: 属性类型的对比

1 数据基本概念
- 数据对象
- 数据属性
- 属性的类型
- 属性类型的对比
- 离散属性与连续属性

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

离散属性与连续属性

前面介绍的四种属性之间不是互斥的，还可以用许多其他方法来组织属性类型，使类型间不互斥。机器学习领域的分类算法常把属性分为离散属性和连续属性。不同类型的属性有不同的处理方法。

1. 离散属性具有有限或无限可数个值。离散属性一般用整数类型变量（int）表示。
2. 连续属性的属性值为实数。在实践中，实数只能用有限位数字的数度量和表示。一般用浮点数（float）表示。如果一个属性不是离散，则必定连续。在部分文献中，“数值属性”与“连续属性”概念等价。

1 数据基本概念

2 数据的统计特征
- 数据的中心化趋势统计量
- 离散度度量
- 分布形状度量

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

数据的统计特征

对于数据分析来说，使用统计量来检查数据特征的操作必不可少。统计量可以衡量数据的集中程度，离散程度和分布形状，通过这些统计量可以识别数据集整体上的一些重要性质。

1 数据基本概念

2 数据的统计特征
- 数据的中心化趋势统计量
- 离散度度量
- 分布形状度量

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

数据的中心化趋势统计量

中心化趋势统计量是指表示一个属性的值大部分落在何处。

- 均值又称算数平均数，数学表达式：

$$mean = \mu = \bar{x} = \frac{1}{m}\sum_{i=1}^{m} x_i$$

为了解决极端值（离群点）问题，可以使用截尾均值，即丢弃极端值的均值。

数据的中心化趋势统计量

- 中位数对于倾斜（非对称）的数据，能够更好地描述数据中心地统计量是中位数（medium）。中位数是有序数据值的中间值，可避免极端数据，代表着数据总体的中等情况。例如，从小到大排序，总数是奇数，取中间数，总数是偶数，取中间两个数的平均数。数学表达式如下：

$$meadian(x) = \begin{cases} x_{r+1} & if\ \ m\%2 == 1 \\ \frac{1}{2}(x_r + x_{r+1}) & if\ \ m\%2 == 0 \end{cases}$$

- 众数（mode）是变量中出现频率最高的值，通常对定性数据确定众数。

1 数据基本概念

2 数据的统计特征
- 数据的中心化趋势统计量
- 离散度度量
- 分布形状度量

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

离散度度量

度量数据离散程度的统计量主要是标准差和四分位极差。

- 标准差与方差用于度量数据分布的离散程度。标准差是方差的算数平方根，低标准差意味着数据更集中。方差公式：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

$$\sigma = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- 差异系数标准差与平均数的比率称为差异系数，又称为相对标准差，符号为 CV。

$$CV = \frac{\sigma}{mean} * 100\%$$

箱线图与四分位极差

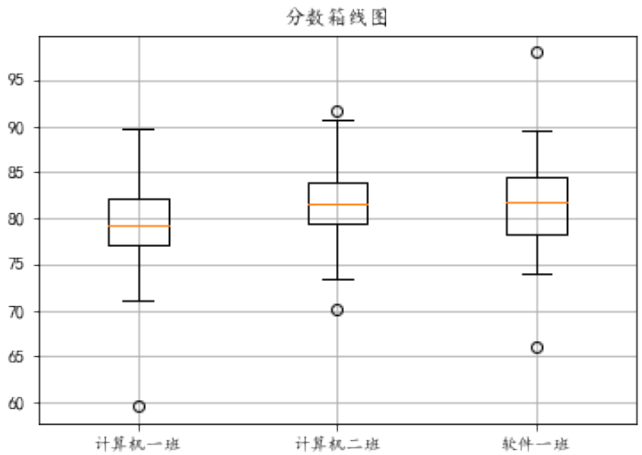

图 2: 箱线图示例

百分位数（quantile）是将一组数据从小到大排序，并计算累计百分位。第一个四位数数记作 Q1，即第 25% 个百分位上的数据，第三个四分位数记作 Q3，即第 75% 个百分位上的数。四分位极差（值的间距）定义为 IQR = Q3 - Q1。数值越大，数据越分散。下边缘定义为 Q1 - 1.5*(IQR)，上边缘 Q3 + 1.5*(IQR)，异常点超出上下边缘的部分。

1 数据基本概念

2 数据的统计特征
- 数据的中心化趋势统计量
- 离散度度量
- 分布形状度量

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

偏度系数

偏度是用于衡量数据分布对称性的统计量。通过对偏度系数的测量，能够判断数据分布的不对称程度即方向。公式定义如下：

$$SK = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}$$

- 对于正太分布（对称分布），偏度等于 0
- 若偏度为负，则 x 均值左侧的离散度比右侧强，左偏
- 若偏度为正，则 x 均值左侧离散度比右侧弱，右偏

偏度系数

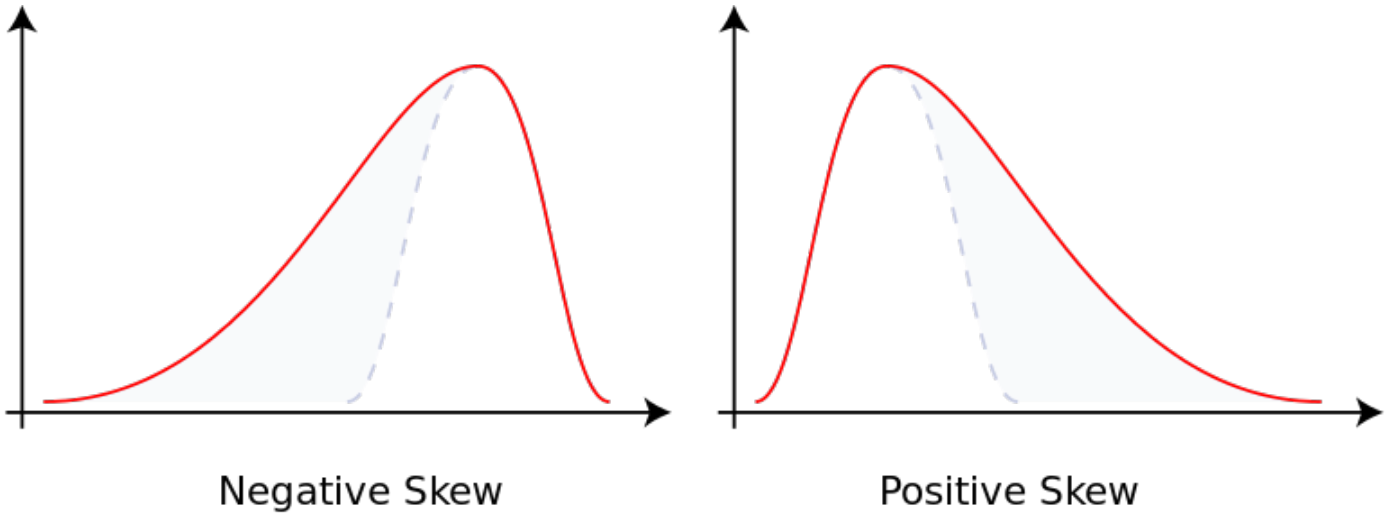

图 3: 左偏与右偏分布

峰度系数

峰度用于衡量数据分布陡峭或平滑的统计量，描述了数据在中心聚集程度，记为 K，是描述所有取值分布形态陡缓的统计量。

$$K = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4}$$

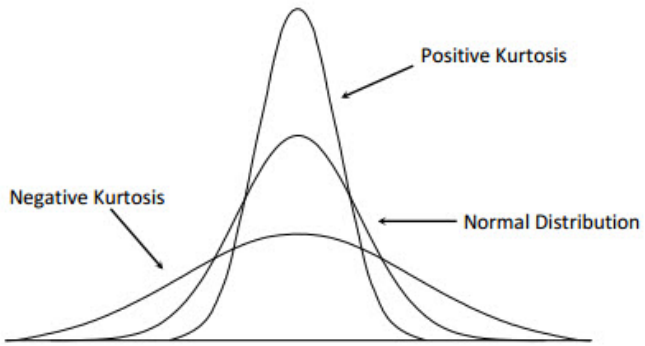

图 4: 不同峰度的分布比较

例如，正态分布的峰度系数为 3，当 K>3 时，更陡峭，有比正态分布更短的尾部；当 K<3 时，说明数据不那么集中，有比正太分布更长的尾部。

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性
- 连续空间向量的距离
- 字符串与集合的距离
- 变量与概率分布的距离

5 参考文献

数据的距离与相似性

在数据挖掘中，距离与相似性是最基础的研究工具。因为分类、回归与聚类算法，本质上是挖掘数据间的相关性与统计规律，从而得出结论。一方面是距离与数据相似性是数据挖掘算法的基石，另一方面距离与相似性也是评价模型性能与数据质量的重要指标。

图 5: 猫和狗的差异

距离使用的场景

- A, B 两市相距 $10KM$。
- A 同学身高 180CM, B 同学身高 176CM 的“距离”
- 向量 $\vec{A} = [1, 3]$ 与向量 $\vec{B} = [3, 1]$ 的距离
- 集合 $A = \{1, 2, 3\}$ 与集合 $B = \{1, 2\}$ 的距离
- 函数 $f(x)$ 与函数 $g(x)$ 的距离
- 概率分布 $pdf(x)$ 与概率分布 $pdf(y)$ 的距离
- 图像间的距离
- 句子间的距离“我喜欢上学”，“我喜欢读书”，“数学真难”
- 抽象实体间的距离，比如人与人之间的距离

距离的定义

一般而言，定义一个函数 $d(x, y)$，若它是一种“距离度量”，则需要满足一些基本性质：

- 非负性：$d(x, y) \geq 0$
- 统一性：$d(x, x) = 0$
- 对称性：$d(x, y) = d(y, x)$
- 三角不等式：从点 i 到 j 的直接距离小于等于途径 k 的距离之和

$$d(i, j) \leq d(i, k) + d(k, j)$$

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性
- 连续空间向量的距离
- 字符串与集合的距离
- 变量与概率分布的距离

5 参考文献

连续空间向量的距离

设向量 $\mathbf{x}$ 与向量 $\mathbf{y}$ 是向量空间 R^n 中的 n 维向量。则可以定义闵可夫斯基距离

$$d(i,j) = ||\mathbf{x} - \mathbf{y}||$$

- 向量 $\mathbf{x}$ 的范数，即向量模的大小，$||\mathbf{x}||^h$ 定义为：

$$||\mathbf{x}||^h = \sqrt[h]{\sum_{i=1}^{n} x_i^h}$$

闵可夫斯基距离

- 当 $h = 1$ 时，闵可夫斯基距离是曼哈顿距离：

$$d(i,j) = \sum_{i=1}^{n} |x_i - y_i|$$

- 当 $h = 2$ 时，闵可夫斯基距离是欧几里得距离：

$$d(i,j) = \sqrt[2]{\sum_{i=1}^{n} |x_i - y_i|^2}$$

- 当 $h \to \infty$，闵可夫斯基距离是切比雪夫距离：

$$d(i,j) = max_i(|x_i - y_i|)$$

闵可夫斯基距离的例子

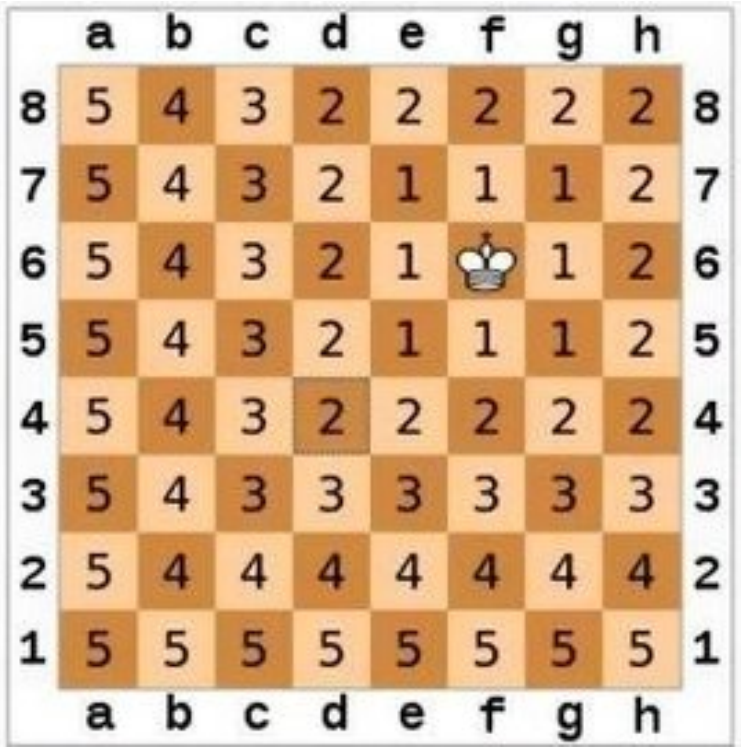

图 6: 切比雪夫距离 (棋盘距离)

余弦距离（相似度）

除了做差来定义距离，还可以利用向量的夹角定义距离或相似性。如常见的余弦相似度，广泛应用于数据挖掘，机器学习，数据科学等领域，定义如下：

$$\text{Similarity}\ (A, B) = \frac{\mathrm{A} \cdot \mathrm{B}}{\|\mathrm{A}\| \times \|\mathrm{B}\|} = \frac{\sum_{\mathrm{i}=1}^{\mathrm{n}} (\mathrm{A_i} \times \mathrm{B_i})}{\sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{A_i^2}} \times \sqrt{\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{B_i^2}}}$$

当 cos(A,B) = +1 或 -1 时，相关性最高，0 代表不相关。正数代表正相关，负数代表负相关。

例题

- 已知有 $A = [1, 1, 3, 1]$ 与 $B = [2, 1, 3, 2]$ 向量，分别求曼哈顿距离，欧几里得距离与切比雪夫距离。

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性
- 连续空间向量的距离
- 字符串与集合的距离
- 变量与概率分布的距离

5 参考文献

汉明距离 (二进制/字符串距离)

- Hamming Distance 汉明距离衡量通长字符串间的最小替换次数。已知字符串 x 与 y，

$$Hamming \quad Distance(x, y) = count_1(x \quad xor \quad y)$$

- 例如：01101, 10101 的汉明距离为 2
- 汉明距离广泛应用于通信，信道编码，密码学等领域。

编辑距离 Edit Distance

编辑距离（Edit Distance，Levenshtein Distance）是一个度量两个字符序列之间差异的字符串度量标准，两个单词之间的编辑距离是将一个字符串转换为另一个字符串所需的单字符编辑（插入、删除或替换）的最小数量。公式定义如下：

$$\mathrm{lev}_{a,b}(i,j) = \begin{cases} \max(i,j) & \text{if } \min(i,j) = 0 \\ \min \begin{cases} \mathrm{lev}_{a,b}(i-1,j) + 1 \\ \mathrm{lev}_{a,b}(i,j-1) + 1 \\ \mathrm{lev}_{a,b}(i-1,j-1) + 1_{(a_i \neq b_j)} \end{cases} & \text{otherwise} \end{cases}$$

可以采用动态规划的算法，给出代码实现。

杰卡德距离 Jaccard Distance

杰卡德距离 Jaccard Distance 定义为：两个集合 A 和 B 的交集与两集合并集的比例，用符号 $J(A,B)$ 表示：

$$\mathrm{J(A,B)} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

杰卡德相似系数定义为：

$$\mathrm{J}_\delta(\mathrm{A},\mathrm{B}) = 1 - \mathrm{J(A,B)} = \frac{|\mathrm{A} \cup \mathrm{B}| - |\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性
- 连续空间向量的距离
- 字符串与集合的距离
- 变量与概率分布的距离

5 参考文献

相关系数

统计学，数据挖掘，机器学习中，用的最大的相关系数有皮尔逊 Pearson 相关系数于斯皮尔曼 Spearman 相关系数两类来衡量随机变量的相关性。

- 皮尔逊相关系数 Pearson

$$\begin{aligned}\rho(X,Y) &= \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y} \\ &= \frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sqrt{\sum_{i=1}^{n}(X_i-\mu_X)^2}\sqrt{\sum_{i=1}^{n}(Y_i-\mu_Y)^2}}\end{aligned}$$

相关系数

- 斯皮尔曼相关系数 Spearman 又称秩相关系数（Coefficient of Rank Correlation），衡量了随机变量间的变换趋势方向和强度之间的关联。

定义：X和Y为两组数据，其斯皮尔曼（等级）相关系数：

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$$

其中，d_i为X_i和Y_i之间的等级差。

（一个数的等级，就是将它所在的一列数按照从小到大排序后，这个数所在的位置）

可以证明：r_s位于-1和1之间。

X	Y	X的等级	Y的等级	等级差	等级差的平方
3	5	2	1	1	1
8	10	5	4.5	0.5	0.25
4	8	3	3	0	0
7	10	4	4.5	-0.5	0.25
2	6	1	2	-1	1

注：如果有的数值相同，则将它们所在的位置取算术平均。

图 7: Spearman 相关系数

例题

- 已知有随机变量 X 和 Y 的结果，分别为 $[10, 1, 2, 11, 2, 3, 5]$ 与 $[9, 3, 1, 10, 2, 1, 5]$，分别求皮尔逊相关系数与斯皮尔曼相关系数。

分布的距离-信息熵回顾

- 自信息 [Gra11] $I(x_i)$:

$$I(x_i) = log\frac{1}{p(x_i)} = -logp(x_i)$$

- 信息熵 $H(x)$:

$$H(x) = \sum_{i=1}^{n} p(x_i) log\frac{1}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) logp(x_i) = -E_{p(x)} logp(x)$$

- 自信息给出了随机变量 x 取特定值时，本身自带的信息量。信息熵给出了随机变量 x 的在给定 $p(x)$ 分布下的平均编码长度。

交叉熵

- 交叉熵定义为在不知道随机变量真实分布 $p(x)$ 的情况下，假设随机变量服从 $q(x)$ 的分布下的平均编码长度。

$$H(p,q) = \sum_{i=1}^{n} p(x_i) log \frac{1}{q(x_i)} = -\sum_{i=1}^{n} p(x_i) log q(x_i) = -E_{p(x)} log q(x)$$

- 交叉熵由信息论提出，是机器学习领域常用的公式。
- 严格来说，交叉熵不是距离，因为不满足对称性，但是一定程度上描述了两个分布的差异情况，交叉熵越小，分布差距越小，交叉熵越大，分布差距越大。

交叉熵的功能

- 在分类任务中，我们常采用交叉熵损失函数来优化我们模型。
- 以下是动物分类的模型结果：

类别	猫	狗	大象
模型 1 输出的概率值	0.2	0.1	0.7
模型 2 输出的概率值	0.6	0.2	0.2
真实标签	0	0	1

表 1: softmax 分类与交叉熵

$$H(p,q) = -1 * log_2(0.7) - 0 * log_2(0.2) - 0 * log_2(0.1) \approx 0.51$$

$$H(p,q') = -1 * log_2(0.2) - 0 * log_2(0.2) - 0 * log_2(0.1) \approx 2.3$$

KL（Kullback-Leibler）散度

- KL 散度定义如下：

$$D_{KL}(p||q) = H(p,q) - H(p) = \sum_{i=1}^{n} p(x_i)(logp(x_i) - logq(x_i))$$

$$= E_{p(x)}[logp(x) - logq(x)] = E_{p(x)} log\frac{p(x)}{q(x)}$$

- KL 散度衡量了近似分布与真实分布的差异，损失了多少信息。
- KL 散度具有的性质
  1. 非对称性：$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$
  2. 非负性：$D_{KL}(P\|Q) \geqslant 0$，当且仅当 $P = Q$ 时等于 0

KL（Kullback-Leibler）散度

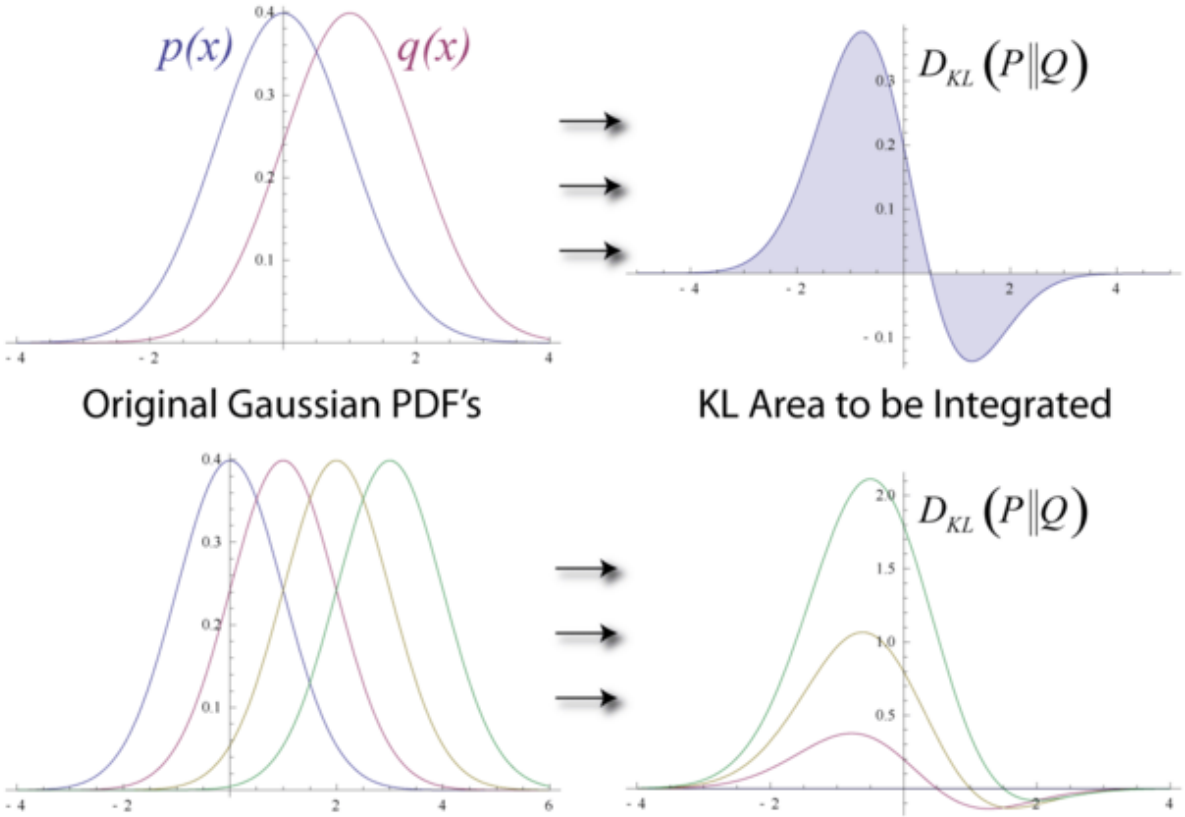

图 8: KL 散度可视化

1 数据基本概念

2 数据的统计特征

3 Iris 可视化案例

4 数据的距离与相似性

5 参考文献

[Gra11] Robert M Gray.
Entropy and information theory.
Springer Science & Business Media, 2011.

Thanks!