七星聚首
7 stars together

互信息（Mutual Information）

本文根据以下参考资料进行整理：
1.维基百科：https://zh.wikipedia.org/wiki/%E4%BA%92%E4%BF%A1%E6%81%AF
2.新浪博客：http://blog.sina.com.cn/s/blog_6255d20d0100ex51.html

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（transinformation）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(Mutual Information)是度量两个事件集合之间的相关性(mutual dependence)。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。

1.互信息的定义

正式地，两个离散随机变量 X 和 Y 的互信息可以定义为：

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

其中 p(x,y) 是 X 和 Y 的联合概率分布函数，而p(x)和p(y)分别是 X 和 Y 的边缘概率分布函数。
在连续随机变量的情形下，求和被替换成了二重定积分：

$$I(X;Y) = \int_Y \int_X p(x,y) \log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) \, dx\, dy,$$

其中 p(x,y) 当前是 X 和 Y 的联合概率密度函数，而p(x)和p(y)分别是 X 和 Y 的边缘概率密度函数。

互信息量I(xi;yj)在联合概率空间P(XY)中的统计平均值。 平均互信息I(X;Y)克服了互信息量I(xi;yj)的随机性,成为一个确定的量。如果对数以 2 为基底，互信息的单位是bit。

直观上，互信息度量 X 和 Y 共享的信息：它度量知道这两个变量其中一个，对另一个不确定度减少的程度。例如，如果 X 和 Y 相互独立，则知道 X 不对 Y 提供任何信息，反之亦然，所以它们的互信息为零。在另一个极端，如果 X 是 Y 的一个确定性函数，且 Y 也是 X 的一个确定性函数，那么传递的所有信息被 X 和 Y 共享：知道 X 决定 Y 的值，反之亦然。因此，在此情形互信息与 Y（或 X）单独包含的不确定度相同，称作 Y（或 X）的熵。而且，这个互信息与 X 的熵和 Y 的熵相同。（这种情形的一个非常特殊的情况是当 X 和 Y 为相同随机变量时。）

互信息是 X 和 Y 联合分布相对于假定 X 和 Y 独立情况下的联合分布之间的内在依赖性。于是互信息以下面方式度量依赖性：I(X; Y) = 0 当且仅当 X 和 Y 为独立随机变量。从一个方向很容易看出：当 X 和 Y 独立时，p(x,y) = p(x) p(y)，因此：

$$\log \left(\frac{p(x,y)}{p(x)\,p(y)} \right) = \log 1 = 0.$$

此外，互信息是非负的（即 I(X;Y) ≥ 0; 见下文），而且是对称的（即 I(X;Y) = I(Y;X)）。

2.平均互信息量的物理含义

（1）观察者站在输出端

H(X/Y) —信道疑义度/损失熵.。Y关于X的后验不确定度。表示收到变量Y后,对随机变量X仍然存在的不确定度。代表了在信道中损失的信息。

H(X) —X的先验不确定度/无条件熵。

I(X;Y)—收到Y前后关于X的不确定度减少的量。从Y获得的关于X的平均信息量。

（2）观察者站在输入端

昵称： gatherstars
园龄： 6年10个月
粉丝： 11
关注： 2
+加关注

< 2022年6月 >

日	一	二	三	四	五	六
29	30	31	1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18
19	20	21	22	23	24	25
26	27	28	29	30	1	2
3	4	5	6	7	8	9

搜索
找找看
谷歌搜索

常用链接
我的随笔
我的评论
我的参与
最新评论
我的标签

我的标签
Linux(2)
统计基础(2)
数据分析(2)
机器学习(2)
算法(1)
C语言(1)
Tensorflow(1)
mac(1)
spark(1)

随笔档案 (9)
2018年1月(2)
2017年9月(1)
2016年11月(3)
2016年10月(2)
2015年7月(1)

阅读排行榜
1. ROC曲线与AUC值(135327)
2. 互信息（Mutual Information）(91908)
3. Linux中如何产生core文件?(14706)
4. 删除单向链表中的某一个节点(12978)
5. 更改Linux默认栈空间的大小(7655)

H(Y/X)—噪声熵。表示发出随机变量X后, 对随机变量Y仍然存在的平均不确定度。如果信道中不存在任何噪声, 发送端和接收端必存在确定的对应关系, 发出X后必能确定对应的Y, 而现在不能完全确定对应的Y, 这显然是由信道噪声所引起的。

I(Y;X) —发出X前后关于Y的先验不确定度减少的量。

（3）观察者站在通信系统总体立场上

H(XY)—联合熵.表示输入随机变量X, 经信道传输到达信宿, 输出随机变量Y。即收,发双方通信后,整个系统仍然存在的不确定度.

I(X;Y) —通信前后整个系统不确定度减少量。在通信前把X和Y看成两个相互独立的随机变量, 整个系统的先验不确定度为X和Y的联合熵H(X)+H(Y); 通信后把信道两端出现X和Y看成是由信道的传递统计特性联系起来的, 具有一定统计关联关系的两个随机变量, 这时整个系统的后验不确定度由H(XY)描述。

以上三种不同的角度说明: 从一个事件获得另一个事件的平均互信息需要消除不确定度,一旦消除了不确定度,就获得了信息。

3.平均互信息量的性质

（1）对称性

I(X;Y)= I(Y;X)

由Y提取到的关于X的信息量与从X中提取到的关于Y的信息量是一样的。 I(X;Y)和 I(Y;X)只是观察者的立足点不同。

（2）非负性

I(X;Y)≥0

平均互信息量不是从两个具体消息出发, 而是从随机变量X和Y的整体角度出发, 并在平均意义上观察问题, 所以平均互信息量不会出现负值。或者说从一个事件提取关于另一个事件的信息, 最坏的情况是0, 不会由于知道了一个事件,反而使另一个事件的不确定度增加。

（3）极值性

I(X;Y)≤H(X)

I(Y;X)≤H(Y)

从一个事件提取关于另一个事件的信息量, 至多是另一个事件的熵那么多, 不会超过另一个事件自身所含的信息量。当X和Y是一一对应关系时: I(X;Y)=H(X), 这时H(X/Y)=0。从一个事件可以充分获得关于另一个事件的信息, 从平均意义上来说, 代表信源的信息量可全部通过信道。当X和Y相互独立时: H(X/Y)=H(X), I(Y;X)=0。 从一个事件不能得到另一个事件的任何信息,这等效于信道中断的情况。

（4） 凸函数性

平均互信息量是p(xi)和p(yj /xi)的函数,即I(X;Y)=f [p(xi), p(yj /xi)];

若固定信道,调整信源, 则平均互信息量I(X;Y)是p(xi)的函数,即I(X;Y)=f [p(xi)];

若固定信源,调整信道, 则平均互信息量I(X;Y)是p(yj /xi)的函数,即I(X;Y)=f [p (yj /xi)]。

平均互信息量I(X;Y)是输入信源概率分布p(xi)的上凸函数(concave function; or convext cap function)。

平均互信息量I(X;Y)是输入转移概率分布p(yj /xi)的下凸函数(convext function; or convext cup function)。

（5）数据处理定理

串联信道：在一些实际通信系统中, 常常出现串联信道。例如微波中继接力通信就是一种串联信道。信宿收到数据后再进行数据处理, 数据处理系统可看成一种信道, 它与前面传输数据的信道构成串联信道。

数据处理定理：当消息经过多级处理后,随着处理器数目的增多,输入消息与输出消息之间的平均互信息量趋于变小。即

I(X;Z)≤I(X;Y)

I(X;Z)≤I(Y;Z)

其中假设Y条件下X和Z相互独立。

两级串联信道输入与输出消息之间的平均互信息量既不会超过第Ⅰ级信道输入与输出消息之间的平均互信息量,也不会超过第Ⅱ级信道输入与输出消息之间的平均互信息量。

当对信号/数据/消息进行多级处理时, 每处理一次, 就有可能损失一部分信息, 也就是说数据处理会把信号/数据/消息变成更有用的形式, 但是绝不会创造出新的信息。这就是所谓的信息不增原理。

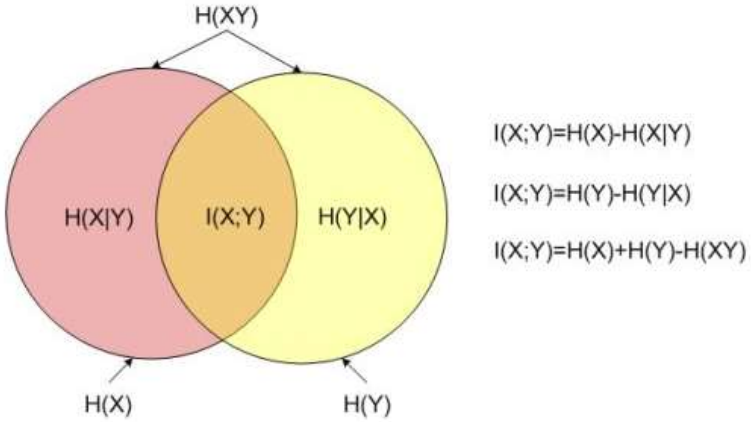
当已用某种方式取得Y后, 不管怎样对Y进行处理, 所获得的信息不会超过I(X;Y)。每处理一次, 只会使信息量减少, 至多不变。也就是说在任何信息流通系统中, 最后获得的信息量,至多是信源提供的信息。一旦在某一过程中丢失了一些信息, 以后的系统不管怎样处理, 如果不能接触到丢失信息的输入端, 就不能再恢复已丢失的信息。

4.与其他量的关系

互信息又可以等价地表示成

$$\begin{aligned} I(X;Y) &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X,Y) - H(X|Y) - H(Y|X) \end{aligned}$$

其中H(X)和H(Y) 是边缘熵，H(X|Y)和H(Y|X)是条件熵，而H(X,Y)是X和Y的联合熵。注意到这组关系和并集、差集和交集的关系类似，用Venn图表示：



于是，在互信息定义的基础上使用琴生不等式，我们可以证明 I(X;Y) 是非负的，因此 H(X)>=H(X|Y)，这里我们给出 I(X;Y) = H(Y) - H(Y|X) 的详细推导:

$$\begin{aligned} I(X;Y) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)} - \sum_{x,y} p(x,y) \log p(y) \\ &= \sum_{x,y} p(x)p(y|x) \log p(y|x) - \sum_{x,y} p(x,y) \log p(y) \\ &= \sum_{x} p(x) \left(\sum_{y} p(y|x) \log p(y|x) \right) - \sum_{y} \log p(y) \left(\sum_{x} p(x,y) \right) \\ &= -\sum_{x} p(x) H(Y|X=x) - \sum_{y} \log p(y) p(y) \\ &= -H(Y|X) + H(Y) \\ &= H(Y) - H(Y|X). \end{aligned}$$

上面其他性质的证明类似。

直观地说，如果把熵 H(Y) 看作一个随机变量不确定度的量度，那么 H(Y|X) 就是 X 没有涉及到的 Y 的部分的不确定度的量度。这就是“在 X 已知之后 Y 的剩余不确定度的量”，于是第一个等式的右边就可以读作“Y的不确定度，减去在 X 已知之后 Y 的剩余不确定度的量”，此式等价于“移除知道 X 后 Y 的不确定度的量”。这证实了互信息的直观意义为知道其中一个变量提供的另一个的信息量（即不确定度的减少量）。

注意到离散情形 H(X|X) = 0，于是 H(X) = I(X;X)。因此 I(X;X) ≥ I(X;Y)，我们可以制定“一个变量至少包含其他任何变量可以提供的与它有关的信息”的基本原理。

互信息也可以表示为两个随机变量的边缘分布 X 和 Y 的乘积 p(x) × p(y) 相对于随机变量的联合熵 p(x,y) 的相对熵:

$$I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y)).$$

此外，令 p(x|y) = p(x, y) / p(y)。则

$$\begin{aligned} I(X;Y) &= \sum_{y} p(y) \sum_{x} p(x|y) \log_2 \frac{p(x|y)}{p(x)} \\ &= \sum_{y} p(y)\ D_{\mathrm{KL}}(p(x|y) \| p(x)) \\ &= \mathbb{E}_Y \{ D_{\mathrm{KL}}(p(x|y) \| p(x)) \}. \end{aligned}$$

注意到，这里相对熵涉及到仅对随机变量 X 积分，表达式

$$D_{\mathrm{KL}}(p(x|y) \| p(x))$$

现在以 Y 为变量。于是互信息也可以理解为相对熵 X 的单变量分布 p(x) 相对于给定 Y 时 X 的条件分布 p(x|y)： 分布 p(x|y) 和 p(x) 之间的平均差异越大，信息增益越大。

标签： 机器学习， 数据分析， 统计基础

gatherstars
关注 - 2
粉丝 - 11
+加关注

7 0

« 上一篇： Spark安装指南
» 下一篇： 更改Linux默认栈空间的大小

posted @ 2016-10-27 15:15 gatherstars 阅读(91911) 评论(0) 编辑 收藏 举报