加微信：642945106 发送“ 赠送 ”领取赠送精品课程
发数字“ 2 ”获取众筹列表

05 数学基础 | 万物皆数，信息亦然：信息论

2017-12-19 王天一

人工智能基础课 进入课程 >

讲述：王天一

时长 11:51 大小 5.43M

近年来的科学研究不断证实，不确定性才是客观世界的本质属性。换句话说，上帝还真就掷骰子。不确定性的世界只能使用概率模型来描述，正是对概率的刻画促成了信息论的诞生。

1948 年，供职于美国贝尔实验室的物理学家克劳德·香农发表了著名论文《通信的数学理论》（A Mathematical Theory of Communication），给出了对信息这一定性概念的定量分析方法，标志着信息论作为一门学科的正式诞生。

香农在《通信的数学理论》中开宗明义： “通信的基本问题是在一点精确地或近似地复现在另一点所选取的消息。消息通常有意义，即根据某种体系，消息本身指向或关联着物理上或概念上的特定实体。但消息的语义含义与工程问题无关，重要的问题是一条消息来自于一个所有可能的消息的集合。”

这样一来，所有类型的信息都被抽象为逻辑符号，这拓展了通信任务的范畴与信息论的适用性，也将信息的传播和处理完全剥离。

信息论使用“信息熵”的概念，对单个信源的信息量和通信中传递信息的数量与效率等问题做出了解释，并在世界的不确定性和信息的可测量性之间搭建起一座桥梁。

在生活中，信息的载体是消息，而不同的消息带来的信息即使在直观感觉上也是不尽相同的。比如，“中国男子足球队获得世界杯冠军”的信息显然要比“中国男子乒乓球队获得世界杯冠军”的信息要大得多。

究其原因，国足勇夺世界杯是如假包换的小概率事件（如果不是不可能事件的话），发生的可能性微乎其微；而男乒夺冠已经让国人习以为常，丢掉冠军的可能性才是意外。因此，以不确定性来度量信息是一种合理的方式。不确定性越大的消息可能性越小，其提供的信息量就越大。

香农对信息的量化正是基于以上的思路，他定义了“熵”这一信息论中最基本最重要的概念。“熵”这个词来源于另一位百科全书式的科学家约翰·冯诺伊曼，他的理由是没人知道熵到底是什么。虽然这一概念已经在热力学中得到了广泛使用，但直到引申到信息论后，**熵的本质才被解释清楚，即一个系统内在的混乱程度。**

在信息论中，如果事件 A 发生的概率为 $p(A)$，则这个事件的自信息量的定义为

$$h(A) = -\log_2 p(A)$$

如果国足闯进世界杯决赛圈，1:1000 的夺冠赔率是个很乐观的估计，用这个赔率计算出的信息量约为 10 比特；而国乒夺冠的赔率不妨设为 1:2，即使在这样高的赔率下，事件的信息量也只有 1 比特。两者之间的差距正是其可能性相差悬殊的体现。

根据单个事件的自信息量可以计算包含多个符号的信源的信息熵。信源的信息熵是信源可能发出的各个符号的自信息量在信源构成的概率空间上的统计平均值。如果一个离散信源 X 包含 n 个符号，每个符号 a_i 的取值为 $p(a_i)$，则 X 的信源熵为

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

信源熵描述了信源每发送一个符号所提供的平均信息量，是信源总体信息测度的均值。当信源中的每个符号的取值概率相等时，信源熵取到最大值 $\log_2 n$，意味着信源的随机程度最高。

在概率论中有**条件概率**的概念，将条件概率扩展到信息论中，就可以得到**条件熵**。如果两个信源之间具有相关性，那么在已知其中一个信源 X 的条件下，另一个信源 Y 的信源熵就会减小。条件熵 $H(Y|X)$ 表示的是在已知随机变量 X 的条件下另一个随机变量 Y 的不确定性，也就是在给定 X 时，根据 Y 的条件概率计算出的熵再对 X 求解数学期望：

$$H(Y|X) = \sum_{i=1}^{n} p(x_i) H(Y|X = x_i)$$

$$= -\sum_{i=1}^{n} p(x_i) \sum_{j=1}^{m} p(y_j|x_i) \log_2 p(y_j|x_i)$$

$$= -\sum_{i=1}^{n} \sum_{j=1}^{m} p(x_i, y_j) \log_2 p(y_j|x_i)$$

防止断更 请务必加
首发微信：17161436665

条件熵的意义在于先按照变量 X 的取值对变量 Y 进行了一次分类，对每个分出来的类别计算其单独的信息熵，再将每个类的信息熵按照 X 的分布计算其数学期望。

以上课为例，学生在教室中可以任意选择座位，那么可能出现的座位分布会很多，其信源熵也就较大。如果对座位的选择添加一个限制条件，比如男生坐左边而女生坐右边，虽然左边的座位分布和右边的座位分布依然是随机的，但相对于未加限制时的情形就会简单很多。这就是分类带来的不确定性的下降。

定义了条件信息熵后，就可以进一步得到**互信息**的概念

$$I(X;Y) = H(Y) - H(Y|X)$$

互信息等于 Y 的信源熵减去已知 X 时 Y 的条件熵，即由 X 提供的关于 Y 的不确定性的消除，也可以看成是 X 给 Y 带来的**信息增益**。互信息这个名称在通信领域经常使用，信息增益则在机器学习领域中经常使用，两者的本质是一样的。

在机器学习中，信息增益常常被用于分类特征的选择。对于给定的训练数据集 Y，$H(Y)$ 表示在未给定任何特征时，对训练集进行分类的不确定性；$H(Y|X)$ 则表示了使用特征 X 对训练集 Y 进行分类的不确定性。信息增益表示的就是特征 X 带来的对训练集 Y 分类不确定性的减少程度，也就是特征 X 对训练集 Y 的区分度。

显然，信息增益更大的特征具有更强的分类能力。但信息增益的值很大程度上依赖于数据集的信息熵 $H(Y)$，因而并不具有绝对意义。为解决这一问题，研究者又提出了**信息增益比**的概念，并将其定义为 $g(X,Y) = I(X;Y)/H(Y)$。

另一个在机器学习中经常使用的信息论概念叫作“**Kullback-Leibler 散度**”，简称**KL 散度**。KL 散度是描述两个概率分布 P 和 Q 之间的差异的一种方法，其定义为

$$D_{KL}(P||Q) = \sum_{i=1}^{n} p(x_i) \log_2 \frac{p(x_i)}{q(x_i)}$$

KL 散度是对额外信息量的衡量。给定一个信源，其符号的概率分布为 $P(X)$，就可以设计一种针对 $P(X)$ 的最优编码，使得表示该信源所需的平均比特数最少（等于该信源的信源熵）。

可是当信源的符号集合不变，而符合的概率分布变为 $Q(X)$ 时，再用概率分布 $P(X)$ 的最优编码对符合分布 $Q(X)$ 的符号编码，此时编码结果的字符数就会比最优值多一些比特。

KL 散度就是用来衡量这种情况下平均每个字符多用的比特数，也可以表示两个分布之间的距离。

KL 散度的两个重要性质是非负性和非对称性。

非负性是指 KL 散度是大于或等于 0 的，等号只在两个分布完全相同时取到。

非对称性则是指 $D_{KL}(P||Q) \neq D_{KL}(Q||P)$，即用 $P(X)$ 去近似 $Q(X)$ 和用 $Q(X)$ 去近似 $P(X)$ 得到的偏差是不同的，因此 KL 散度并不满足数学意义上对距离的定义，这一点需要注意。

事实上，$D_{KL}(P||Q)$ 和 $D_{KL}(Q||P)$ 代表了两种不同的近似方式。要让 $D_{KL}(P||Q)$ 最小，需要让 $Q(X)$ 在 $P(X)$ 不等于 0 的位置同样不等于 0；要让 $D_{KL}(Q||P)$ 最小，则需要让 $Q(X)$ 在 $P(X)$ 等于 0 的位置同样等于 0。

除了以上定义的指标之外，信息论中还有一个重要定理，叫作“**最大熵原理**”。**最大熵原理是确定随机变量统计特性时力图最符合客观情况的一种准则。对于一个未知的概率分布，最坏的情况就是它以等可能性取到每个可能的取值。**这个时候的概率分布最均匀，也就是随机变量的随机程度最高，对它进行预测也就最困难。

从这个角度看，最大熵原理的本质在于在推断未知分布时不引入任何多余的约束和假设，因而可以得到最不确定的结果，预测的风险也就最小。投资理财中的名言“不要把所有鸡蛋放在同一个篮子里”，就可以视为最大熵原理的一个实际应用。

将最大熵原理应用到分类问题上就可以得到**最大熵模型**。在分类问题中，首先要确定若干特征函数作为分类的依据。为了保证特征函数的有效性，其在模型真实分布 $P(X)$ 上的数学期望和在由训练数据集推导出的经验分布 $\tilde{P}(X)$ 上的数学期望应该相等，即对给定特征函数数学期望的估计应该是个无偏估计量。

这样一来，每一个特征函数就对应了一个约束条件。分类的任务就是在这些约束条件下，确定一个最好的分类模型。由于除了这些约束条件之外，没有任何关于分类的先验知识，因而需要利用最大熵原理，求解出不确定性最大的条件分布，即让以下函数的取值最大化

$$H(p) = -\sum_{x,y} \tilde{p}(x)p(y|x)\log_2 p(y|x)$$

式中的 $p(y|x)$ 就是分类问题要确定的目标条件分布。计算上式的最大值实质上就是一个约束优化问题，由特征函数确定的约束条件可以通过**拉格朗日乘子**的引入去除其影响，转化为无约束优化问题。从数学上可以证明，这个模型的解是存在且唯一的。

今天我和你分享了人工智能必备的信息论基础，着重于抽象概念的解释而非数学公式的推导，其要点如下：

信息论处理的是客观世界中的不确定性；

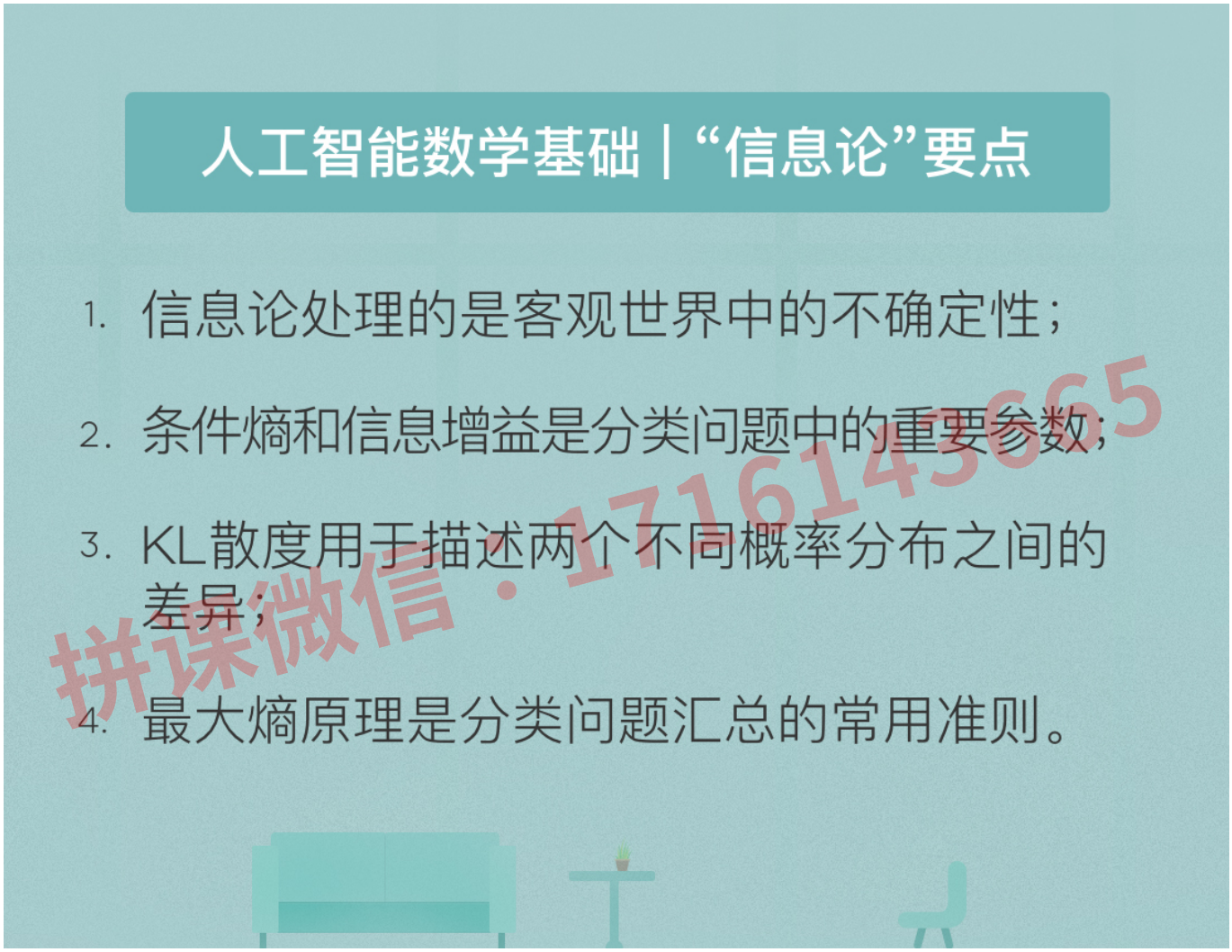
条件熵和信息增益是分类问题中的重要参数；

KL 散度用于描述两个不同概率分布之间的差异；

最大熵原理是分类问题中的常用准则。

信息论建立在概率的基础上，但其形式并不唯一，除了香农熵外也有其他关于熵的定义。那么概率与信息之间的关系对人工智能有什么启示呢？

欢迎发表你的观点。



拼课微信：17161456665

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　04 数学基础 | 不畏浮云遮望眼：最优化方法

下一篇　06 数学基础 | 明日黄花迹难寻：形式逻辑

精选留言 (17)

写留言

井中月
2018-03-03

4

王老师，感谢您的回复。但是我还有点疑惑，X表示的是训练集的某个特征，Y相当于是训练集中需要被分类的变量，那么这样的话H(Y)就是一个定值，用它做分母和直接使用信息增益进行特征选择不就是一样的吗？

展开

作者回复: 感谢你指出，这里的符号写的不够清晰，H(Y)其实应该写成H_X(Y)。H(Y)是直接用数据的分类结果计算出来的信息熵，H_X(Y)的下标X表示的是以特征X的取值为变量对数据集计算出的信息熵。所以当关注的特征X不同时，H_X(Y)也是不一样的。
信息增益比主要用在决策树当中，作用是消除多个取值的特征导致的偏差，因为多值特征的信息增益很大，但泛化性能却很差。比如，使用姓名作为特征可以得到较大的信息增益，因为它基本

可以把每个人区分开来，但这种区分对于分类显然没什么帮助。这时就可以用信息增益比来一定程度上消除对多值属性的偏向性，但也不能完全消除。

刘祯
2018-01-04

3

看完之后，我努力应用如下：

消息是今天我学会了专栏的信息论部分，因为可能性较低，因而信息量较大，信息熵也就越大。

...

展开

星运里的错
2018-05-19

1

信息增益表示的就是特征 X带来的对训练集 Y 分类不确定性的减少程度，也就是特征 X 对训练集 YY的区分度。
这句话意思是 总体熵-某个特征下的熵 = 去除某个特征影响的熵 老师。这个公式对么？
我的理解是 熵对应着信息量的多少。熵大，意味着信息量大，信息混杂，也就是不确定性大。...

展开

作者回复: 总体熵 - 特征分类之后每个类别的熵的总和 = 特征的信息增益
这里的信息增益表示的是分类之后残留的不确定度。如何一个特征能够将两个类别完全正确地分开，那它的信息增益是最大的，就等于数据集的熵。

夜星辰
2018-03-05

1

有一点理解上的困惑希望王老师帮忙解答下

1. 熵表示的是信息量大小，从公式中知道随着概率增大，熵会变小。而机器学习中常用交叉熵作为目标函数，学习的过程是不断求取最小熵，也就是求取概率最大的参数，等价于极大似然估计法进行参数估计。...

展开

作者回复: 最大熵表示的是对未知的部分不做任何多余的假设，所以要选择符合已有知识但不确定性最大，也就是熵最大的分布，通俗说就是不要不懂装懂。对交叉熵的最小化意味着数据训练的模型要尽可能地接近真实模型，而真实模型又是建立在最大熵的前提下的。所以在优化时，要不断地调整训练的模型，以期更接近真实情况。

秦龙君
2017-12-29

1

学习了。这篇很难，后半部分暂时还看不懂。

展开

Naraka,
2019-03-25

老师，不知道现在提问还会不会回答，
“从这个角度看，最大熵原理的本质在于在推断未知分布时不引入任何多余的约束和假设，因而可以得到最不确定的结果，预测的风险也就最小。”
这一段没有看懂，为什么得到最不确定的结果，预测风险会最小？最不确定的，可能性很多，预测的结果不也更吗？

展开

囊子
2019-01-21

可以参考数学之美第二版，第六章 - 信息的度量和作用

展开

Snail@AI_M...
2019-01-10

非常棒，深入浅出，对照了培训课程之后，有一个更清晰的思路，虽然理解程度可能不够深，但觉得目前够用了，安利一波😁

Shawn
2018-12-29

这一张看完特别熟悉，翻了好几本书找到了数学之美

展开

梦帝

2018-12-25

老师你好，虽然留言里面提到了，但有一个问题还是不太明白，关于最大熵模型和交叉熵的。在网上看max最大熵模型的解时发现，其实max最大熵模型，就是max模型的最大似然估计，也就是说如果以logistic regression为例的化，max最大熵模型和max logistic regression的最大似然估计是一样的，而max logistic regression的最大似然估计其实就是minimize对应的cross entropy，所以其实最大熵模型和最小化cross entropy是不是...

展开

Mr.Button

2018-08-13

为什么log以2为底的函数这么常见...这里为什么取2

展开

作者回复: 以2为底计算出的单位就是二进制的比特。

水木竹水

2018-07-06

首先感谢老师讲的非常好。有个疑惑问下老师，信息增益是H(Y)-H(Y|X)，后者是已知X情况下Y的不确定性，信息增益就是X对Y的确定性消除。H(Y|X)越小，说明X对Y的分类效果越好，为何决策树不直接用H(Y|X)选取主要特征，而用信息增益，H(Y)是变化的吗？

作者回复: 数据集确定了，总体的信息熵H(Y)就是常量，所以两个其实是等效的。之所以选信息增益一方面在于它和信息论一脉相承，意义清晰；另一方面，在取值上信息增益是越大越好，如果选一个越小越好的指标，有些反直觉。

井中月

2018-03-01

王老师，您好，我有个疑问，信息增益比里面的分母是不是应该是H(X)？

展开

作者回复: 分母是训练数据集的信息熵，因为这里把训练集定为Y，所以分母就是H(Y)。

卡斯瓦德

2018-02-01

看完这篇，突然觉得所谓的奇迹，其实就是信息熵不对等的结果，从某个面如何环境，物质看概率为百万分之一，从另一个面如自主意念等，概率可能就是十分之一，那么事件成就的结果其实就是KL后，不同的结果，饿可能总结有点问题，但是有那么个方向的感觉

作者回复: 奇迹其实就是小概率事件的发生

wolfog

2018-01-16

这个推荐大家可以看看吴军老师的数学之美其中就有关于最大熵和互信息等的介绍，讲的更加详细和通俗一些

wolfog

2018-01-16

之前看过吴军老师的《数学之美》，这一张还听得有点眉目，加油。

展开

chucklau

2017-12-30

嗯，这篇的内容很难理解，希望有其它更多的相关资料，谢谢老师。

展开

作者回复: 可以参考MacKay的《信息论，推理与学习算法》