中信建投期货
CHINA FUTURES

CHINAFUTURES RESEARCH

信息量及信息增量的数学度量

专题报告

摘要：

信息熵是信息论和机器学习中非常重要的概念，应用极其广泛。本文对自信息、信息熵、基尼不纯度、交叉熵、相对熵、条件熵、互信息与信息增益等基本概念进行介绍，并对部分含义在金融市场中的应用进行举例说明。为后续模型介绍等做基础铺垫。

风险提示：策略失效风险、模型误设风险、历史统计规律失效等风险。

作者姓名：彭鲸桥

邮箱：pengjingqiao @csc.com.cn
电话：023-86769675
期货从业资格号：F3074348
期货投资咨询从业证书号：Z0012925

研究助理：姜慧丽

邮箱：jianghuili @csc.com.cn
电话：023-81157278
期货从业资格号：F3081375

发布日期： 2022 年 6 月 16 日

目录

一、自信息

通常我们在说“信息量很大”或者“没什么信息量”时，“信息量”是在表示什么呢？

上述的“信息量”，其实即是自信息的概念。自信息是由随机事件发生的概率计算而得的基本量，它量化了特定结果出现时带来的信息量的值。当某事件发生的概率非常小，但实际上却发生了，则此时的意外程度，即自信息（信息量）就非常大；反之，当某事件发生的概率非常大，并且实际上也发生了，则此时的意外程度，即自信息（信息量）比较小。

按照客观事实和习惯，自信息的定义应满足以下几个公理：

1. 概率为 1 的事件是完全确定的，当它发生时不会产生任何信息。

2. 事件发生的概率越小，它就越不确定，当它发生时产生的信息就越多。

3. 如果分别测量两个独立事件，则信息总量为两事件的自信息之和。

由此，当给定一个发生概率为 P 的事件 x，在数学上，其定义如下:

$$h(x) = -\log_2 P \tag{1}$$

例如，当我们预测 PTA 第二天上涨的概率为 30%，则上涨结果对应的自信息为 1.737；同理，当预测 PTA 第二天上涨的概率为 90%时，上涨结果对应的自信息为 0.152。注意，下文中 log 的底数均默认为 2。

自信息的概念在金融市场中也有体现，除了上述直观例子之外，通常市场对某一事件超预期的反应也基于该事件的自信息值比较大。比如，市场预期原油库存较上期变化不大，即原油库存大幅上行或下行的概率比较低，但实际原油库存下行幅度较大，则超出市场预期，即，该事件的信息值比较大。

二、信息熵

上述自信息描述的是随机变量的某特定事件发生所带来的信息量，而信息熵更具统计意义，它通常用来描述随机变量分布的混乱程度，分布越混乱，则信息熵越大，纯度越低，基于熵的机器学习算法则是基于这一思想。给定一个离散随机变量 X：

$$H(X) = -\sum_{i=1}^{n} p(x_i) logp(x_i) = E\big(-logp(X)\big) \tag{2}$$

继续以 PTA 为例，若第二天上涨的概率为 30%，下跌的概率为 70%，其信息熵为$-0.3 \times log_2 0.3 - 0.7 \times log_2 0.7 = 0.881$；若第二天上涨的概率为 90%，下跌的概率为 10%，其信息熵为$-0.9 \times log_2 0.9 - 0.1 \times log_2 0.1 = 0.469$。可知，第一种分布的混乱程度更大，纯度更低。

三、基尼不纯度

从一个数据集中随机选取子项，其被错误地划分到其他组的概率叫做 Gini 不纯度（基尼不纯度）。将上述信息熵的 -logp(x) 项进行泰勒展开后，忽略高阶项（近似于 0），就得到了基尼不纯度的公式：

$$Gini(X) = \sum_{i=1}^{n} p(x_i)\big(1 - p(x_i)\big) \tag{3}$$

由此可见，基尼不纯度与熵的含义类似，也是衡量混乱程度或者说不纯度的概念。

同上一节例子，其第一种情况基尼不纯度等于$0.3 \times 1 - 0.3 + 0.7 \times 1 - 0.7 = 0.42$，第二种情况为 0.18，同样可知第一种分布的混乱程度更大。

四、交叉熵

从熵的计算公式来看，若要计算某个事件的熵，则需要知道这个事件的概率分布。但往往事件的概率分布不得而知，在这种情况下若要计算熵，实际上是对熵的估计。

如果对某事件估计的概率分布是 q，同时我们知道它真正的分布 p，则 p 和 q 之间的交叉熵定义如下：

$$H(p,q) = -E_p logq = -\sum_{i=1}^{n} p(x_i) logq(x_i) \tag{4}$$

当 q 与 p 相同，也就是我们估计的概率分布与真实的概率分布完全相同时，交叉熵等于信息熵，否则交叉熵大于信息熵。由于这样的关系，交叉熵可以用作分类模型的损失函数。

举例来看，当我们使用历史数据预测今天 PTA 涨幅超过 5%的概率为 20%，下跌超过 5%的概率为 30%，其他情况为 50%，同时已知今天 PTA 实际涨幅超过 5%，即涨幅超过 5%的概率为 1，另外两种结果概率为 0。由此，我们可得其交叉熵为$-(1 \times log_2 0.2 + 0 \times log_2 0.3 + 0 \times log_2 0.5) = 2.32$。

五、相对熵

相对熵或称 Kullback-Leibler 散度$D_{KL}(p||q)$，衡量的是交叉熵与信息熵之间的差距。其定义如下：

$$D_{KL}(p||q) = E_p log\frac{p(x_i)}{q(x_i)} = -E_p log\frac{q(x_i)}{p(x_i)} = -\sum_{i=1}^{n} p(x_i) log\frac{q(x_i)}{p(x_i)} = H(p,q) - H_p(X) \tag{5}$$

一般而言，如果只需要评估损失函数的下降值，交叉熵可以满足要求，其计算量比相对熵更小；而如果需要通过算法对样本数据进行概率分布建模，则需要精确计算生成的分布和真实分布的差距，此时则考虑使用相对熵。

例如，当我们预测 PTA 明天的涨跌幅概率时，若真实概率为涨幅超过 5%的概率为 10%，下跌超过 5%的概率为 30%，其他情况为 60%；同时当我们使用历史数据预测今天 PTA 涨幅超过 5%的概率为 20%，下跌超过 5%的概率为 30%，其他情况为 50%，则相对熵为$-(0.1\times log_2 0.2+0.3\times log_2 0.3+0.6\times log_2 0.5)=1.354$。

六、条件熵

将信息熵推广到多元情况时称作联合熵，二元联合熵定义如下：

$$H(X,Y)=-\sum_{i=1}^{n}\sum_{j=1}^{n}p(x_i,y_j)log\,p(x_i,y_j) \tag{6}$$

条件熵则量化了在已知另一个随机变量 X 的情况下，随机变量 Y 的不确定性：

$$H(Y|X)=\sum_{i=1}^{n}p(x_i)H(Y|X=x)=-\sum_{i=1}^{n}\sum_{j=1}^{n}p(x_i,y_j)log\,\frac{p(x_i,y_j)}{p(x_i)}=H(X,Y)-H(X) \tag{7}$$

我们以 PTA 与乙二醇为例，若两者的联合概率分布如下：

PTA \ 乙二醇	上涨	下跌
上涨	20%	10%
下跌	20%	50%

则其联合熵为$-(0.2\times log_2 0.2+0.1\times log_2 0.1+0.2\times log_2 0.2+0.5\times log_2 0.5)=1.760$。

同时，若已知乙二醇实际情况为上涨，则$H(乙二醇|PTA)=H(乙二醇,PTA)-H(乙二醇)=1.760-(0.4log0.4+0.6log0.6)=0.790$。

七、互信息与信息增益

互信息度量了两个集合之间的相关性，是对总体信息量的评价。它量化了已知一个随机变量，对于另一个随机变量的不确定的减少程度，定义如下：

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(Y) + H(Y) - H(X,Y) \quad (8)$$

由上述定义可知，如果两个随机变量相互独立，则他们的互信息为 0。

信息增益的定义如下：

$$g(X,Y) = H(Y) - H(Y|X) \quad (9)$$

信息增益是同一集合分类后增加的信息量的度量。公式上看与互信息公式相同，但应用范围不同。通常信息增益是一个估计值。

另外有信息增益率是信息增益与固有信息的比值，定义为

$$g_r(X,Y) = g(X,Y)/H(Y) \quad (10)$$

例子同上一节，可得 PTA 与乙二醇的互信息/信息增益为$-0.3log0.3 - 0.7log0.7 - 0.79 = 0.09$。

八、总结

信息熵是信息论和机器学习中非常重要的概念，应用极其广泛。本文对自信息、信息熵、基尼不纯度、交叉熵、相对熵、条件熵、互信息与信息增益等基本概念进行介绍，并对部分含义在金融市场中的应用进行举例说明。为后续模型介绍等做基础铺垫。

联系我们

中信建投期货总部

重庆市渝中区中山三路131号希尔顿商务中心27楼、30楼

电话：023-86769605

上海分公司

地址：中国（上海）自由贸易试验区浦电路490号，世纪大道1589号8楼10-11单元

电话：021-68765927

济南分公司

地址：济南市历下区泺源大街150号中信广场A座六层611、613室

电话：0531-85180636

湖南分公司

地址：长沙市岳麓区观沙岭街道茶子山东路112号滨江金融中心C座2127、2128室

电话：0731-82681681

大连分公司

地址：大连市沙河口区会展路129号大连国际金融中心A座大连期货大厦2901号房间

电话：0411-84806336

河南分公司

地址：郑州市未来路69号未来大厦2205、2211、1910房，未来公寓1306、1506、1806房

电话：0371-65612397

河北分公司

地址：廊坊市广阳区吉祥小区20-11号门市一至三层、20-1-12号门市第三层

电话：0316-2326908

深圳分公司

地址：深圳市福田区深南大道和泰然大道交汇处绿景纪元大厦11I

电话：0755-33378759

杭州分公司

地址：浙江省杭州市江干区钱江国际时代广场3幢702室

电话：0571-87380613

宁波分公司

地址：浙江省宁波市鄞州区和济街180号国际金融中心F座1809室

电话：0574-89071681

西安分公司

地址：陕西省西安市高新区科技路38号林凯国际大厦十九层1905、1906、1907室

电话：029-85725585

重庆渝北分公司

地址：重庆市渝北区龙山街道新南路439号中国华融现代广场3幢19-1/2号

电话：023-67380500

上海浦东分公司

地址：中国（上海）自由贸易试验区浦东南路528号2202室

电话：021-68597013

四川分公司

地址：成都市武侯区科华北路62号力宝大厦南楼1801、1802、1803室

电话：028-62818710

重庆分公司

地址：重庆市渝中区中山三路107号上站大楼平街名义层11-A4-A6

电话：023-61361140

海南分公司

地址：海南省海口市龙华区滨海大道77号中环国际广场10层1002号

电话：0898-68538536

北京朝阳门北大街营业部

地址：北京市东城区朝阳门北大街6号首创大厦207室

电话：010-85282866

南昌营业部

地址：江西省南昌市红谷滩新区红谷中大道998号绿地中央广场A1#办公楼-3404室

电话：0791-82082702

广州东风中路营业部

地址：广州市越秀区东风中路410号第16层自编1605C、1605B、1606房

电话：020-28325286

漳州营业部

地址：福建省漳州市龙文区九龙大道以东漳州碧湖万达广场A2地块9幢1203号

电话：0596-6161588

合肥营业部

地址：安徽省合肥市包河区马鞍山路130号万达广场C区6幢1903、1904、1905室

电话：0551-2889767

上海徐汇营业部

地址：上海市徐汇区斜土路2899甲号1幢1601室

电话：021-64040178

武汉营业部

地址：武汉市江汉区香港路193号中华城A写字楼栋/单元36层3601号02-03室

电话：027-59909521

南京营业部

地址：南京市黄埔路 2 号黄埔大厦 11 层 D1、D2 座

电话：025-86951881

北京北三环西路营业部

地址：北京市海淀区中关村南大街 6 号 9 层 912

电话：010-82129971

太原营业部

地址：山西省太原市小店区长治路 103 号阳光国际商务中心 A 座 902 室

电话：0351-8366898

广州黄埔大道营业部

地址：广州市天河区黄埔大道西 100 号富力盈泰大厦 B 座 1406

电话：020-22922102

北京国贸营业部

地址：北京市朝阳区光华路 8 号 17 幢一层 A113 房间

电话：010-85951101

福州营业部

地址：福建省福州市台江区宁化街道振武路 70 号（原江滨西大道北侧）福晟·钱隆广场 18 层 01 商务办公

电话：0591-83625596

方顿物产（重庆）有限公司

地址：重庆市渝中区中山三路 131 号希尔顿商务中心 2603 室

电话：023-86769662

重要声明

本报告观点和信息仅供符合《证券期货投资者适当性管理办法》规定可参与期货交易的投资者参考。中信建投不因任何订阅或接收本报告的行为而将订阅人视为中信建投的客户。

本报告发布内容如涉及或属于系列解读，则投资者若使用所载资料，有可能会因缺乏对完整内容的了解而对其中假设依据、研究依据、结论等内容产生误解。提请投资者参阅中信建投已发布的完整系列报告，仔细阅读其所附各项声明、数据来源及风险提示，关注相关的分析、预测能够成立的关键假设条件，关注研究依据和研究结论的目标价格及时间周期，并准确理解研究逻辑。

中信建投对本报告所载资料的准确性、可靠性、时效性及完整性不作任何明示或暗示的保证。本报告中的资料、意见等仅代表报告发布之时的

判断，相关研究观点可能依据中信建投后续发布的报告在不发布通知的情形下作出更改。

中信建投的销售人员、交易人员以及其他专业人士可能会依据不同假设和标准、采用不同的分析方法而口头或书面发表与本报告中资料意见不一致的市场评论和/或观点。本报告发布的内容并非投资决策服务，在任何情形下都不构成对接收本报告内容投资者的任何投资建议，投资者应充分了解各类投资风险并谨慎考虑本报告发布内容是否符合自身特定状况，自主做出投资决策并自行承担投资风险。投资者根据本报告内容做出的任何决策与中信建投或相关作者无关。

全国统一客服电话：400-8877-780

网址：www.cfc108.com