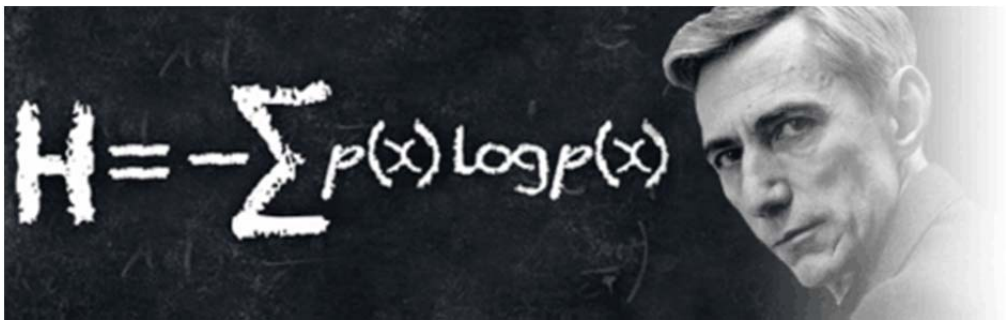
第三章　最大熵模型

1.什么是熵

1.1 概述

熵的概念起源于物理中的热力学，熵的英文原文为 entropy，最初由德国物理学家鲁道夫·克劳修斯提出，用来表示分子状态混乱程度的物理量。

1948 年，**香农**（Claude E. Shannon）引入**信息熵**概念，用来量化信息的不确定性。



什么是信息呢？信息论创始人香农（C.E.Shannon）于 1948 年从信息接收者的角度定义:“**信息是能够协助信息接收者消除事件不确定的因素**”。

假设一篇文章的标题叫做“黑洞到底吃什么”，包含词语分别是{黑洞，到底，吃什么}，我们现在要根据一个词语推测这篇文章的类别。哪个词语给予我们的信息最多？很容易就知道是“黑洞”，因为“黑洞”这个词语在所有的文档中出现的概率太低啦，一旦出现，就表明这篇文章很可能是在讲科普知识。而其他两个词语“到底”和“吃什么”出现的概率很高，给予我们的信息反而越少。

香农为此从概率论角度出发，引入**不确定程度**的概念。在日常生活中经常会遇到一些随机事件，这些事件的结果是事先不能完全肯定的。如果一个事件有 N 种可能性相等（均为 $1/N$）的结果，则结果结果未出现前的不确定程度和 N 有关。N 越大，事件的不确定度就越大，信息量也越大。不确定度应为 N 的单调上升函数，而当 $N=1$时，事件只有一个选择结果，不确定度为零，即必然事件不提供信息量。据此，香农把随机事件的信息量 I 定义为该事件发生概率 p（$=1/N$）的倒数的对数。即

$$I=\log\frac{1}{p}=-\log p$$

对数的底一般取 2。信息量 I 称为信息熵。单位为比特（bit）。举个例子，预测明天的天气，如果能 100%确定明天一定是晴天，即只有一种结果，那么熵就是 $-\log 1=0$,也就是说不确定性为零。如果说明天有天气有两种可能结果，50%概率晴天，50%概率下雨，那么明天晴天（下雨）的信息熵就是 $-\log_2 0.5=\log_2 2=1$。可以说明天晴天（下雨）的不确定性为 1 或者说信息量为 1 比特。而如果明天天气有 4 种可能结果，有 25%概率晴天，25%概率下雨，25%概率阴天，25%概率下雪，那么其中明天晴天的信息熵就是 $-\log_2 0.25=2$，也就是说**随着不确定程度的增加，熵也在不断地增大**。

世界杯决赛的两支球队巴西和南非中，哪支球队获得了冠军？在对球队实力没有任何了解的情况下，每支球队夺冠的概率都是 0.5，所以谁获得冠军这条信息的信息量是 $-\log 0.5 = 1\text{bit}$。

其实这正好对应了计算机对数字的表示，如果用二进制表示，每一位出现 0 和 1 的概率都是 1/2，所以每一位的信息量是 1 比特。如果用十六进制表示，每一位出现任意一个符号的概率是 1/16，所以每一位能表示 $-\log_2 \frac{1}{16} = 4$ 比特。所以 1 位十六进制的信息量，和 4 位二进制信息量是相同的。

这样就比较好理解另一个经典的例子，英文有 26 个字母，假设每个字母出现的概率是一样的，每个字母的信息量就是 $-\log_2 \frac{1}{26} = 4.7$ 比特；常用的汉字有 2500 个，每个汉字的信息量是 $-\log_2 \frac{1}{2500} = 11.3$ 比特。所以在信息量相同的情况下，使用的汉字要比英文字母要少——这其实就是十六进制和二进制的区别，在这个例子中，“apple“成了 5 位 26 进制的数值，信息量 4.7 * 5 = 23.5；而“苹果“成为 2 位 2500 进制的数值，信息量 11.3 * 2 = 22.6。

在实际的情况中，每种可能情况出现的概率并不是相同的，设事件事件 X 共有 m 个不同的结果，概率为 p_i，所以**熵就用来衡量平均信息量**，其计算公式如下

$$H(X) = \sum_{i=1}^{m} p_i \log \frac{1}{p_i} = -\sum_{i=1}^{m} p_i \log p_i$$

熵是平均信息量，也可以理解为不确定性。例如进行决赛的巴西和南非，假设根据经验判断，巴西夺冠的几率是 80%，南非夺冠的几率是 20%，则谁能获得冠军的信息量就变为 $-0.8 \times \log_2 0.8 - 0.2 \times \log_2 0.2 = 0.721 < 1$。经验减少了判断所需的信息量，消除了不确定性。

1.2 熵的数学定义

自信息

一个事件（消息）本身所包含的信息量，它是由事件的不确定性决定的。比如抛掷一枚硬币的结果是正面这个消息所包含的信息量。**随机事件的自信息量定义为该事件发生概率的倒数的对数。**

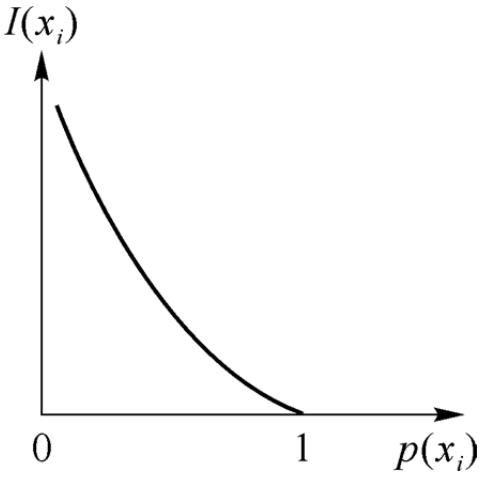
设一次随机事件（用随机变量 X 表示），它可能会有 $x_1, x_2, x_3, \cdots, x_m$ 共 m 个不同的结果（取值），设事件 x_i 的概率为 $p(x_i)$，则 x_i 的**自信息量**定义为

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

自信息量也简称**自信息**。$I(x_i)$ 代表两种含义：**当事件发生以前，等于事件发生的不确定性的大小；当事件发生以后，表示事件所含有或所能提供的信息量。**

为什么定义为倒数的对数呢？自信息量是概率的函数，概率越小，事件发生的不确定性越大，事件发生以后所包含的自信息量越大；另外从直观概念上讲，由两个相对独立的不同的消息所提供的信息量应等于它们分别提供的信息量之和；极限情况下，$p=0$时，$I\to\infty$，$p=1$时，$I=0$。满足以上条件的函数就是倒数的对数。对数的好处是可以使自信息量可加，设想你同时做两个互相独立的实验，出现的结果分别是a和b，那么这件事发生的概率是$p(a)*p(b)$，你得到的信息量就是

$$\log\frac{1}{p(a)*p(b)}=\log\frac{1}{p(a)}+\log\frac{1}{p(b)}=I(a)+I(b)$$。



信息熵

事件集（用随机变量表示）所包含的平均信息量，它表示平均不确定性。比如抛掷一枚硬币的试验所包含的信息量。

离散随机变量 X 的每一个可能取值的自信息 $I(x_i)$ 的统计平均值定义为随机变量 X 的**平均自信息量**，通常称为随机变量 X 的**信息熵，**简称**熵**:

$$H(X)=E\left[I(x_i)\right]=\sum_{i=1}^{m}p(x_i)\log\frac{1}{p(x_i)}=-\sum_{i=1}^{m}p(x_i)\log p(x_i)$$

同样可以定义连续型随机变量的信息熵。若连续型随机变量 X 的概率密度函数为 $p(x)$，其信息熵定义为：

$$H(X)=-\int_x p(x)\log p(x)dx$$

有时也记为 $H(P)$，称为**概率分布 P 的熵**。

从通信角度来说，自信息量是信源发出某一具体消息所含有的信息量，发出的消息不同，所含有的信息量也不同。因此自信息量不能用来表征整个信源的不确定大小。定义平均自信息量（信息熵）来表征整个信源的不确定度。

自信息和信息熵定义中的对数的底通常采用 2，有时也可以采用其它的底。当底为 2 时，熵的单位是**比特**（bit）；当底为 e 时，熵的单位是**奈特**（Nat）；当底为 10 时，熵的单位是**哈特**（Hart）。

联合熵

一个随机变量的不确定性可以用信息熵来表示，这一概念可以直接推广到多个随机变量。

设 X，Y 为两个随机变量，$p\left(x_i,y_j\right)$ 表示其联合概率，用 $H\left(X,Y\right)$ 表示联合熵，计算公式为：

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

条件熵

设 X ，Y 为两个随机变量，在 X 发生的前提下，Y 发生的不确定程度即信息熵定义为 Y 的**条件熵**（Conditional Entropy），用 $H(Y|X)$ 表示，计算公式如下：

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

其含义是当变量 X 已知时，变量 Y 的平均不确定程度是多少。推导如下：

假设变量 X 取值有 m 个，那么 $H(Y|X=x_i)$ 是指变量 X 被固定为值 x_i 时的条件熵；$H(Y|X)$ 指变量 X 被固定时的条件熵。那么二者之间的关系是：

$$\begin{aligned}H(Y|X)&=p(x_1)\cdot H(Y|X=x_1)+\cdots+p(x_m)\cdot H(Y|X=x_m)\\&=\sum_{i=1}^{m}p(x_i)\cdot H(Y|X=x_i)\end{aligned}$$

进一步计算 Y 的条件熵：

$$\begin{aligned}H(Y|X)&=\sum_{i=1}^{m}p(x_i)\cdot H(Y|X=x_i)\\&=-\sum_{i=1}^{m}p(x_i)\cdot\left(\sum_{j=1}^{n}p(y_j|x_i)\cdot\log\ p(y_j|x_i)\right)\\&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(y_j,x_i)\cdot\log\ p(y_j|x_i)\end{aligned}$$

其中运用了公式 $p(y,x)=p(y|x)p(x)$。

条件熵、联合熵、熵之间的关系：

$$H(Y|X)=H(X,Y)-H(X)$$

公式推导如下：

$$\begin{aligned}H(X,Y)-H(X)&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot log_2p(x_i,y_j)+\sum_{i=1}^{m}\underline{p(x_i)}\cdot log_2p(x_i)\\&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot log_2p(x_i,y_j)+\sum_{i=1}^{m}\underline{\left(\sum_{j=1}^{n}p(x_i,y_j)\right)}\cdot log_2p(x_i)\\&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\left(log_2p(x_i,y_j)-log_2p(x_i)\right)\\&=-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot log_2p(y_j|x_i)\\&=H(Y|X)\end{aligned}$$

即 $H(X,Y)=H(X)+H(Y|X)$。这个关系可以推广到 N 个随机变量的情况：

$$H(X_1X_2\cdots X_N)=H(X_1)+H(X_2|X_1)+\cdots+H(X_N|X_1X_2\cdots X_{N-1})$$

称为熵函数的**链式规则。**

相对熵（交叉熵、KL 距离）

相对熵，又称为交叉熵、KL 距离、KL 散度（Kullback-Leibler Divergence），主要用

于衡量**相同事件空间**里的两个概率分布 $p(x)$ 和 $q(x)$ 的差异。用 $\mathrm{KL}(p\|q)$ 或 $\mathrm{D}(p\|q)$ 表示 KL 距离，定义如下：

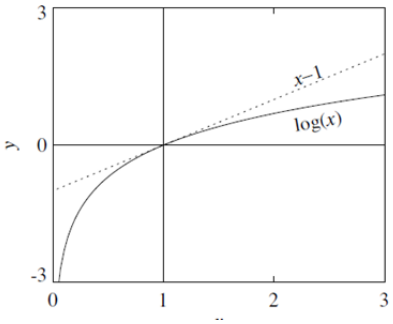
$$\mathrm{KL}(p\|q)=\sum_i p(x_i)\cdot\log\frac{p(x_i)}{q(x_i)}$$

连续形式：

$$\mathrm{KL}(p\|q)=\int_{x\in X} p(x)\cdot\log\frac{p(x)}{q(x)}dx$$

可以看出，当两个概率分布完全相同时，KL 距离为 0。当两个概率分布的差别增加时，KL 距离也增加。相对熵越大，两个概率分布函数的差异性越大；反之，相对熵越小，两个概率分布函数的差异性越小。对于概率分布或者概率密度函数，如果取值均大于 0，相对熵可以用来度量两个随机分布的差异性。

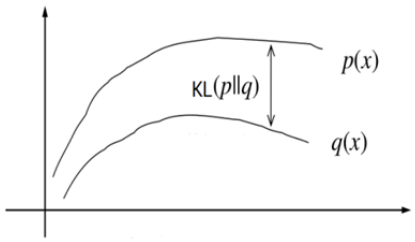
可以证明 KL 距离满足 $\mathrm{KL}(p\|q)\geq 0$（非负性），但不满足对称性和三角不等式。



KL 距离（相对熵）非负性的证明：从上图看出 $\log x\leq x-1$ 且仅当 $x=1$ 时等号成立。于是有：

$$\begin{aligned}\mathrm{KL}(p\|q)&=\sum_i p(x_i)\cdot\log\frac{p(x_i)}{q(x_i)}\\&=-\sum_i p(x_i)\cdot\log\frac{q(x_i)}{p(x_i)}\\&\geq-\sum_i p(x_i)\cdot\left(\frac{q(x_i)}{p(x_i)}-1\right)\\&=-\sum_i\left(q(x_i)-p(x_i)\right)\\&=\sum_i p(x_i)-\sum_i q(x_i)=1-1=0\end{aligned}$$

一般情形下，$p(x)$ 表示样本数据或观测值的分布，或数据的真实分布，$q(x)$ 表示数据的模型预测分布，或 $p(x)$ 的近似分布。相对熵也可以理解为，用分布 $q(x)$ 近似估计 $p(x)$，$p(x)$ 的不确定程度减少了多少。



举一个实际的例子：比如有四个类别，一个方法 A 得到 4 个类别的概率分别是 0.1,0.2,0.3,0.4。另一种方法 B（或者说是事实情况）是得到 4 个类别的概率分别是 0.4,0.3,0.2,0.1，那么这两个分布的 KL 距离为

$$0.1\times log_2\left(0.1/0.4\right)+0.2\times log_2\left(0.2/0.3\right)+$$
$$0.3\times log_2\left(0.3/0.2\right)+0.4\times log_2\left(0.4/0.1\right)=0.6585$$

相对熵最早是用在信号处理上，如果两个随机信号相对熵越小，说明这两个信号越接近，否则信号的差异性越大。后来用来衡量两段信息的相似程度，比如说如果一篇文章是照抄或者改写另一篇，那么这两篇文章中词频分布的相对熵就非常小，接近于 0。相对熵在自然语言处理中还有很多应用，比如用来衡量两个常用词（在语法和语义上）不同文本中的概率分布，看它们是否同义。另外，利用相对熵，还可以得到信息检索中最重要的一个概念：词频率-逆向文档频率（TF-IDF）。

我们可以利用 KL 距离来判定两个图像块的相似性。将 KL 距离引入到含噪图像块中，用 KL 解决的问题是，每个像素点受到噪声的污染，所以认为它是一个随机变量。这一对像素点服从一定的概率分布，那么用 KL 距离可以计算这两个随机变量的概率分布距离，如果两个像素点服从相同参数的同一概率分布（即它们相似），那么 KL 距离越小，以达到像素点之间相似性的判定。图像块的之间的 KL 距离就可以通过各像素的 KL 距离之和来求得。

互信息（Mutual Information）

如果说相对熵衡量的是相同事件空间里的两个事件的相似度大小，那么，互信息通常用来衡量**不同事件空间**里的两个随机事件的相关性大小。比如常识告诉我们，随机事件“今天广州下雨”与另一个随机变量“过去二十四个小时广州空气湿度”的相关性就很大，但是到底有多大呢？再比如，“过去二十四个小时广州空气湿度”与“旧金山的天气”似乎就相关性不大，如何度量这种相关性呢？香农提出了**互信息**（Mutual Information）作为两个随机事件相关性的度量。从信息量上讲，互信息是用来评价一个事件的出现对于另一个事件的出现所贡献的信息量，或者说，互信息是一个事件的出现消除或减少另一个事件的不确定性的度量。

设 X 和 Y 为两个离散随机变量，事件 $Y=y_j$ 的出现对于事件 $X=x_i$ 的出现的**互信息**（**交互信息量**的简称）$I(x_i,y_j)$ 定义为：

$$I(x_i;y_j)=\log\frac{p(x_i\mid y_j)}{p(x_i)}$$

即 y 的出现对 x 的互信息是 x 的后验概率与先验概率比值的对数。即

$$\begin{aligned}\text{交互信息量}&=\log\frac{\text{后验概率}}{\text{先验概率}}=\log\frac{1}{\text{先验概率}}-\log\frac{1}{\text{后验概率}}\\&=\text{先验不确定性（熵）}-\text{后验不确定性（熵）}\\&=\text{通信前后不确定性的减少}\end{aligned}$$

结合乘法公式 $p(x,y)=p(x\mid y)p(y)$，把上面定义中的分子分母同乘以 $p(y_j)$ 可得

$$I(x_i;y_j)=\log\frac{p(x_i\mid y_j)}{p(x_i)}=\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

对于事件 X 和 Y 来说，它们之间的**平均互信息**用 $I(X;Y)$ 表示。平均互信息简称为**互信息**。定义为：

$$I(X;Y)=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}$$

由此可知，$I(X;Y)=I(Y;X)$，这也是称为**交互信息量**的原因。

互信息和各类熵的关系:

1. $I(X;Y)=KL(P(X,Y)\|P(X)P(Y))$

   因为

   $$I(X;Y)=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}=KL(P(X,Y)\|P(X)P(Y))$$

   X，Y 独立时，即 $P(X,Y)=P(X)P(Y)$，$I(X;Y)=0$。互信息 $I(X;Y)$ 反映了随机变量 X 和 Y 的“独立程度”。

2. $I(X;Y)=H(X)-H(X|Y)$，

   推导如下：

$$\begin{aligned}H(X)-H(X|Y)&=\sum_{i=1}^{m}p(x_i)\cdot\log\frac{1}{p(x_i)}-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{1}{p(x_i|y_j)}\\&=\sum_{i=1}^{m}\left(\sum_{j=1}^{n}p(x_i,y_j)\right)\cdot\log\frac{1}{p(x_i)}-\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{1}{p(x_i|y_j)}\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\left(\log\ p(x_i|y_j)-\log\ p(x_i)\right)\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{p(x_i|y_j)}{p(x_i)}\\&=\sum_{i=1}^{m}\sum_{j=1}^{n}p(x_i,y_j)\cdot\log\frac{p(x_i,y_j)}{p(x_i)p(y_j)}\\&=I(X,Y)\end{aligned}$$

   这个等式有时也用来定义互信息。互信息从不确定性发生改变的的角度也称为**信息增益**（Information gain）。为什么叫信息增益呢？Y 的出现使 X 的不确定减少，比如一个均匀骰子丢出 2 点（事件 X）的概率是 1/6，熵是 $\log_2 6$（$H(X)=\log_2 6$），已知丢出点数是双数（事件 Y）的情形下丢出 2 点的概率是 1/3，熵是 $\log_2 3$（$H(X|Y)=\log_2 3$），$\log_2 3<\log_2 6$，熵减少因为不确定程度减少了，其差 $\log_2 6-\log_2 3$ 就是“丢出点数是双数”这个信息的度量 $I(X;Y)=\log_2 6-\log_2 3$，熵和条件熵的差就是信息增益。

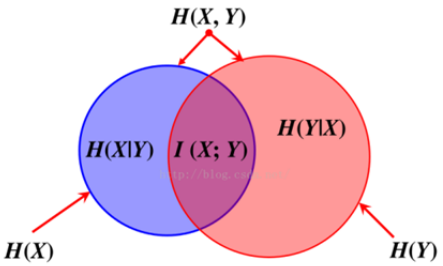
3. $I(X;Y)=H(X)+H(Y)-H(X,Y)$

   因为 $H(X,Y)=H(X)+H(Y|X)$

   所以

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

如右图，可以直观得到



2. 最大熵模型

在介绍最大熵模型之前，我们先了解一下最大熵原理，因为最大熵原理是选择最优概率模型的一个准则。

2.1 最大熵原理

在概率模型空间集合中，在满足给定约束条件的前提下，使信息熵最大化得到的概率模型，就是最优的模型，称为最大熵原理。

假设离散随机变量 X 的概率分布是 $P(X)$，其信息熵满足以下不等式：

$$0 \le H(X) \le \log |X|$$

其中，$|X|$ 是 X 的取值个数，当且仅当 X 的分布是均匀分布时右边的等号才成立。也就是说，当 X **服从均匀分布时，熵最大**。

证明如下：设离散随机变量 X 的分布律为 $p(X = x_i) = p_i$（ $i = 1, 2, ..., m$ ）。其信息熵表示如下

$$H(x_1, x_2, ..., x_m) = -\sum_{k=1}^{m} p_k \log_2 p_k$$

而约束条件为

$$g(p_1, p_2, ..., p_m) = \sum_{k=1}^{m} p_k = 1$$

要求函数 H 的最大值，根据**拉格朗日乘数法**，设

$$
\begin{aligned}
L(p_1, p_2, ..., p_m) &= H(X) + \lambda[g(p_1, p_2, ..., p_m) - 1] \\
&= -\sum_{k=1}^{m} p_k \log_2 p_k + \lambda(\sum_{k=1}^{m} p_k - 1)
\end{aligned}
$$

对所有的 p_k 求偏导数并令偏导数为零，得到

$$\frac{\partial}{\partial p_k}\left(-\sum_{k=1}^{m} p_k \log_2 p_k + \lambda(\sum_{k=1}^{m} p_k - 1) \right) = 0$$

得到 m 个等式

$$-\left(\frac{1}{\ln 2} + \log_2 p_k \right) + \lambda = 0$$

这说明所有的 p_k 都相等，最终解得

$$p_k = \frac{1}{m}$$

因此，当 X 服从**均匀分布**时，信息熵最大。

对于连续性随机变量可以得到类似的结论。**对概率密度函数未知的连续型随机变量，服从均匀分布的的随机变量的信息熵最大**。即概率密度函数为

$$p(x)=\frac{1}{b-a},\ a<x<b$$

最大熵是概率模型学习中一个准则，其思想为：在学习概率模型时，所有可能的模型中熵最大的模型是最好的模型；若概率模型需要满足一些约束，则最大熵原理就是在满足已知约束的条件集合中选择熵最大模型。最大熵原理指出，**对一个随机事件的概率分布进行预测时，预测应当满足全部已知的约束，而对未知的情况不做任何主观假设（称为无偏见原则）。在这种情况下，概率分布最均匀，预测的风险最小，因此得到的概率分布的熵最大。**

例如，投掷一个骰子，如果问"每个面朝上的概率分别是多少"，你会说是等概率，即各点出现的概率均为 1/6。因为对这个"一无所知"的骰子，什么都不确定，而假定它每一个朝上概率均等则是最合理的做法。

吴军《数学之美》中关于最大熵的论述：最大熵原理指出，当我们需要对一个随机事件的概率分布进行预测时，我们的预测应当满足全部已知的条件，而对未知的情况不要做任何主观假设。在这种情况下，概率分布最均匀，预测的风险最小。因为这时概率分布的信息熵最大，所以人们称这种模型叫“最大熵模型”。我们常说，不要把所有的鸡蛋放在一个篮子里，其实就是最大熵原理的一个朴素的说法，因为当我们遇到不确定性时，就要保留各种可能性。说白了，就是要保留全部的不确定性，将风险降到最小。

最大熵原理举例

我们通过一个简单的例子来介绍最大熵原理。

问题：假设随机变量 X 有 5 个取值 $\{A,B,C,D,E\}$，要估计各个取值的概率 $P(A),P(B),P(C),P(D),P(E)$。这个简单的例子也可能来自于实际问题，比如机器翻译，汉语词汇“打”翻译成英语的 5 个备选词汇:{fight,buy,break,dozen,since}，需要确定翻译成各备选词汇的概率。

首先这些概率只满足以下约束条件：

$$P(A)+P(B)+P(C)+P(D)+P(E)=1$$

满足这个约束条件的概率分布有无穷多个，但是在没有任何其它信息的情况下，根据最大熵原理和无偏见原则，选择熵最大时对应的概率分布，即各个取值概率相等是一个不错的概率估计方法。即有：

$$P(A)=P(B)=P(C)=P(D)=P(E)=\frac{1}{5}$$

等概率坚持了最大熵的无偏见原则，因为没有更多信息，此种判断是合理的。

现在从先验知识中得到一些信息：A 和 B 的概率值之和满足以下条件：

$$P(A)+P(B)=\frac{3}{10}$$

同样的，满足上面两个约束条件的概率分布仍有无穷多个。在缺少其它信息的情况下，坚持无偏见原则，得到：

$$P(A)=P(B)=\frac{3}{20}$$

$$P(C)=P(D)=P(E)=\frac{7}{30}$$

2.2 最大熵模型

最大熵原理是统计学习的一般原理，将它应用到分类问题中，即得到最大熵模型。

训练数据集 $D=\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(m)},y^{(m)})\}$，学习的目标是：**用最大熵原理选择最优的分类模型。**

假设分类模型是一个条件概率分布 $P(y|x), x\in X\subseteq R^n$ 表示输入（特征向量），$y\in Y$，X 和 Y 分别是输入（特征向量）和输出（标签）的集合。这个模型表示的是对于给定的输入 X，以**条件概率** $P(y|x)$ 计算得到**标签** y。

首先，考虑模型应满足的**约束条件**。约束条件是靠**特征函数**来引入的。

特征函数 $f(x,y)$ 描述 x 与 y 之间的某一事实，其定义如下：

$$f(x,y)=\begin{cases}1, & x,y\text{满足某一事实}\\ 0, & \text{不满足该事实}\end{cases}$$

这是一个二值函数，当 x 与 y 满足某一事实时取值为 1，否则为 0。比如对于以下数据集：

Outdoor	Cloudy	Happy	Humid
Outdoor	Cloudy	Sad	Humid
Outdoor	Cloudy	Sad	Humid
Indoor	Rainy	Happy	Humid
Indoor	Rainy	Happy	Dry

其中第一列为标签 Y，其余列为特征 X，可以为该数据集写出特征函数，形式如下：

$$f(x,y)=\begin{cases}1, & \text{if } x=\text{Cloudy and } y=\text{Outdoor}\\ 0, & \text{else}\end{cases}$$

需要为每个<特征，标签>对都定义一个类似的特征函数。对于前面提到的机器翻译问题，特征函数可设置为

$$f(x,y)=\begin{cases}1, & \text{if } y=\text{buy and “打”后面接的词汇是“酱油”}\\ 0, & \text{else}\end{cases}$$

给定训练集，可以计算得到总体的**联合分布** $P(X,Y)$ 和**边缘分布** $P(X)$ 的**经验分布**，分别以 $\tilde{P}(X,Y)$ 和 $\tilde{P}(X)$ 表示，即：

$$\tilde{P}(X=x,Y=y)=\frac{count(X=x,Y=y)}{m}$$

$$\tilde{P}(X=x)=\frac{count(X=x)}{m}$$

其中，$count(X=x,Y=y)$ 表示训练集中样本 (x,y) 出现的**频数**，$count(X=x)$ 表示训练集中输入 x 出现的频数，m 表示训练集中的样本个数。

- 特征函数 $f(x,y)$ 关于经验分布 $\tilde{P}(X,Y)$ 的期望值，用 $E_{\tilde{P}}(f)$ 表示如下：

$$E_{\tilde{P}}(f)=\sum_{x\in X,y\in Y}\tilde{P}(x,y)\cdot f(x,y)$$

- 特征函数 $f(x,y)$ 关于基于模型 $P(y|x)$ 计算的总体联合分布 $P(X,Y)$ 的近似期望值，用 $E_P(f)$ 表示如下：

$$\begin{aligned}E_P(f)&=\sum_{xx\in X,y\in Y}P(x,y)\cdot f(x,y)\\&=\sum_{x\in X,y\in Y}P(x)\cdot P(y|x)\cdot f(x,y)\\&\approx\sum_{x\in X,y\in Y}\tilde{P}(x)\cdot P(y|x)\cdot f(x,y)\end{aligned}$$

这里因为 $P(x)$ 不好求，所以用样本中 x 出现的概率 $\tilde{P}(X)$ 代替 x 在总体中的分布概率 $P(x)$

我们希望模型能够符合（拟合）样本数据，那么就可以假设这两个期望值相等。即：

$$E_P(f) = E_{\tilde{P}}(f)$$

得到等式

$$\sum_{x\in X, y\in Y} \tilde{P}(x)\cdot P(y\mid x)\cdot f(x,y) = \sum_{x\in X, y\in Y} \tilde{P}(x,y)\cdot f(x,y)$$

上式即为最大熵模型需要满足约束条件，给定 n 个特征函数 $f_i(x, y)$，则有 n 个约束条件，用 C 表示满足约束的模型集合：

$$C = \left\{P \mid E_P(f_i) = E_{\tilde{P}}(f_i), I = 1,2,\cdots,n\right\}$$

从满足约束的模型集合 C 中找到使得 $P(Y\mid X)$ 的熵最大的分布就是最大熵模型了。

最大熵模型

假设满足所有约束条件的模型集合为：

$$C = \left\{P \mid E_P(f_i) = E_{\tilde{P}}(f_i), I = 1,2,\cdots,n\right\}$$

定义在条件概率分布 $P(Y\mid X)$ 上的条件熵为：

$$H(P) = -\sum_{x,y} \tilde{P}(x)\cdot P(y\mid x)\cdot \log P(y\mid x)$$

模型集合 $\mathcal{C}$ 中条件熵 $H(P)$ 最大的模型称为最大熵模型。

$$P^* = arg\max_{P\in C} H(P) \quad 或 \quad P^* = arg\min_{P\in C} -H(P)$$

综上给出形式化的**最大熵模型**：

给定数据集 $D = \{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \cdots, (x^{(m)}, y^{(m)})\}$，特征函数 $f_i(x, y)$ $i = 1,2\ldots,n$，根据经验分布得到满足约束集的模型集合 C：

$$\min_{P\in C} \sum_{x,y} \tilde{P}(x)P(y\mid x)\log P(y\mid x)$$

$$s.t. \quad E_p(f_i) = E_{\tilde{P}}(f_i), i = 1,2,\cdots n$$

$$\sum_y P(y\mid x) = 1$$

2.3 最大熵模型的求解

我们先讨论前面最大熵原理中的讲过的例子。为了简便，这里分别以 y_1, y_2, y_3, y_4, y_5 表示 A, B, C, D 和 E，最大熵模型学习的最优化问题可以表示为：

$$min \quad -H(P) = \sum_{i=1}^{5} P(y_i)\cdot logP(y_i)$$

$$s.t. \quad P(y_1) + P(y_2) = \tilde{P}(y_1) + \tilde{P}(y_2) = \frac{3}{10}$$

$$\sum_{i=1}^{5} P(y_i) = \sum_{i=1}^{5} \tilde{P}(y_i) = 1$$

将带约束优化问题转化为无约束优化问题：引入拉格朗日乘子 w_0, w_1 ，定义拉格朗日函数：

$$L(P,w)=\sum_{i=1}^{5}P(y_i)\log P(y_i)+w_1\left(P(y_1)+P(y_2)-\frac{3}{10}\right)+w_0\left(\sum_{i=1}^{5}P(y_i)-1\right)$$

根据拉格朗日对偶性，可以通过求解对偶最优化问题得到原始最优化问题的解（见附录），所以求解对偶问题：

$$\max_w \min_P L(P,w)$$

首先求解 $L(P,w)$ 关于 P 的极小化问题。为此，固定 w_0, w_1 ，求偏导数：

$$\frac{\partial L(P,w)}{\partial P(y_1)}=1+log_2P(y_1)+w_1+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_2)}=1+log_2P(y_2)+w_1+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_3)}=1+log_2P(y_3)+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_4)}=1+log_2P(y_4)+w_0$$

$$\frac{\partial L(P,w)}{\partial P(y_5)}=1+log_2P(y_5)+w_0$$

令各偏导数等于 0，可解得：

$$P(y_1)=P(y_2)=e^{-w_1-w_0-1}$$

$$P(y_3)=P(y_4)=P(y_5)=e^{-w_0-1}$$

于是，极小化结果为：

$$\min_P L(P,w)=L(P_w,w)=-2e^{-w_1-w_0-1}-3e^{-w_0-1}-\frac{3}{10}w_1-w_0$$

下面再求解对偶函数 $L(P_w,\boldsymbol{w})$ 关于 $\boldsymbol{w}$ 的极大化问题：

$$\max_w L(P_w,w)=-2e^{-w_1-w_0-1}-3e^{-w_0-1}-\frac{3}{10}w_1-w_0$$

分别求 $L(P_w,\boldsymbol{w})$ 对 w_0, w_1 的偏导数，并令其为 0，得到：

$$e^{-w_1-w_0-1}=\frac{3}{20}$$

$$e^{-w_0-1}=\frac{7}{30}$$

于是得到所求的概率分布为

$$P(y_1)=P(y_2)=\frac{3}{20}$$

$$P(y_3)=P(y_4)=P(y_5)=\frac{7}{30}$$

***最大熵模型求解的一般流程**

对于给定训练数据集 $D=\{(x^{(1)},y^{(1)}),(x^{(2)},y^{(2)}),\cdots,(x^{(m)},y^{(m)})\}$，最大熵模型的求解等价于带约束的最优化问题（ $f(x,y)$ 为特征函数）：

$$\min_{P\in\mathcal{C}} \quad \sum_{x,y}\tilde{P}(x)\cdot P(y\mid x)\cdot \log P(y\mid x)$$

$$s.t. \quad E_P(f_i)=E_{\tilde{P}}(f_i), i=1,2,\cdots,n$$

$$\sum_y P(y\mid x)=1$$

将约束最优化的原始问题转换为无约束最优化的对偶问题。具体推导过程如下：

- **首先，引入拉格朗日乘子 $w_0,w_1,\cdots,w_n$，定义拉格朗日函数 $L(P,w)$**

$$\begin{aligned} L(P,w) &= -H(P)+w_0\cdot\left(1-\sum_y P(y\mid x)\right)+\sum_{i=1}^{n} w_i\cdot\left(E_{\tilde{P}}(f_i)-E_P(f_i)\right) \\ &= \sum_{x,y}\tilde{P}(x)\cdot P(y\mid x)\cdot \log P(y\mid x)+w_0\cdot\left(1-\sum_y P(y\mid x)\right) \\ &\quad +\sum_{i=1}^{n} w_i\cdot\left(\sum_{x,y}\tilde{P}(x,y)\cdot f_i(x,y)-\sum_{x.y}\tilde{P}(x)\cdot P(y\mid x)\cdot f_i(x,y)\right) \end{aligned}$$

最优化的**原始问题**是：

$$\min_{P\in\mathcal{C}}\max_w L(P,w)$$

对偶问题是：

$$\max_w \min_{P\in\mathcal{C}} L(P,w)$$

由于最大熵模型对应的朗格朗日函数 $L(P,w)$ 是参数 P 的凸函数，所以**原始问题的解与对偶问题的解是等价的**。因此，可以通过求解对偶问题来得到原始问题的解。

- **其次，求对偶问题内部的极小化问题 $\min_{P\in\mathcal{C}} L(P,w)$**

  $\min_{P\in\mathcal{C}} L(P,w)$ 是乘子 w 函数，将其记作：

$$\Psi(w)=\min_{P\in\mathcal{C}} L(P,w)=L(P_w,w)$$

  将其解记作：

$$P_w=arg\min_{P\in\mathcal{C}} L(P,w)=P_w(y\mid x)$$

具体地，固定 w_i，求 $L(P,w)$ 对 $P(y\mid x)$ 的偏导数：

$$\begin{aligned} \frac{\partial L(P,w)}{\partial P(y\mid x)} &= \sum_{x,y}\tilde{P}(x)\cdot\left(logP(y\mid x)+1\right)-\sum_y w_0-\sum_{x,y}\left(\tilde{P}(x)\cdot\sum_{i=1}^{n} w_i\cdot f_i(x,y)\right) \\ &= \sum_{x,y}\tilde{P}(x)\cdot\left(logP(y\mid x)+1-w_0-\sum_{i=1}^{n} w_i\cdot f_i(x,y)\right) \end{aligned}$$

令偏导数等于 0，在 $\tilde{P}(x)>0$ 的情况下，求得：

$$P(y\mid x)=\exp\left(\sum_{i=1}^{n} w_i\cdot f_i(x,y)+w_0-1\right)=\frac{\exp\left(\sum_{i=1}^{n} w_i\cdot f_i(x,y)\right)}{\exp(1-w_0)}$$

由于 $\sum_y P(y\mid x)=1$，可得：

$$P_w(y|x)=\frac{1}{Z_w(x)}\exp\left(\sum_{i=1}^{n}w_i\cdot f_i(x,y)\right)$$

其中，

$$Z_w(x)=\sum_y\exp\left(\sum_{i=1}^{n}w_i\cdot f_i(x,y)\right)$$

$Z_w(x)$ 称为归一化因子； $f_i(x,y)$ 是特征函数； w_i 是第 i 个参数（特征权值）。模型 $P_w=P_w(y|x)$ 就是最大熵模型（ $\boldsymbol{w}$ 是最大熵模型中的参数向量）。

- **最后，求解对偶问题外部的极大化问题**

对偶问题外部极大化表达式

$$\max_w \Psi(\boldsymbol{w})$$

将其解记作 $\boldsymbol{w}^*$ ，即

$$\boldsymbol{w}^*=arg\max_w \Psi(\boldsymbol{w})$$

也就是说，可以应用最优化算法求对偶函数 $\Psi(\boldsymbol{w})$ 的极大化，得到 $\boldsymbol{w}^*$ ，用其表示 $P^*=P_{w^*}=P_{w^*}(y|x)$ 是学习到的最优模型（最大熵模型）。可以证明，对偶函数的极大化等价于最大熵模型的极大似然估计。

附录

拉格朗日乘子法、拉格朗日对偶性、KKT 条件

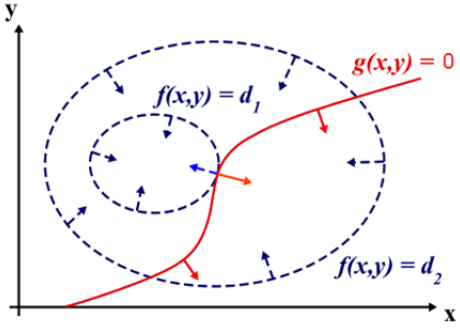
拉格朗日乘子法(Lagrange Multiplier)

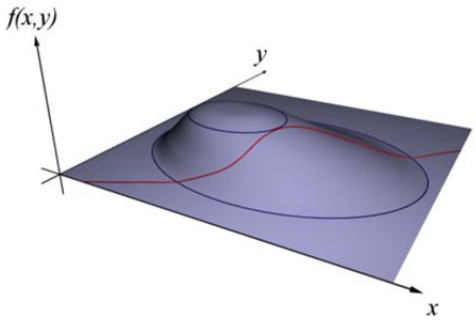
拉格朗日乘子法是一种寻找有等式约束条件的函数的最优值(最大或者最小)的最优化方法.在求取函数最优值的过程中,约束条件通常会给求取最优值带来困难,而拉格朗日乘子法就是解决这类问题的一种强有力的工具.

有等式约束的最优化问题

我们先考虑以下的二维单约束优化问题:

$$\begin{aligned}\max\quad & f(x,y)\\ s.t.\quad & g(x,y)=0\end{aligned}$$





我们在等高线图上考虑，先任取 $f(x,y)=d_2$。此时，等高线与约束曲线的交点既是满足约束且取值为 d_2 的点。下面，我们逐渐增大 $f(x,y)$ 的取值，即 d 逐渐由 d_2 变为 d_1，这个过程中，$f(x,y)$ 的值逐渐增大，当达到 d_1 后，再增大一点点，即不满足约束，因此 d_1 既是既是 $f(x,y)$ 在该约束下的最大值。而这个时候，约束曲线正好和等高线相切。也就是说，在这一点，d_1 就是在该约束下的最大值。而这个时候，约束曲线正好和等高线相切。在这一点，$f(x,y)$ 和和 $g(x,y)$ 的**梯度方向共线**，即存在一个数 λ，使得

$$\nabla f(x,y)=-\lambda\nabla g(x,y)$$

即

$$\nabla f(x,y)+\lambda\nabla g(x,y)=0$$

所以可以构造一个函数，称为**拉格朗日函数** L：

$$L(x,y,\lambda)=f(x,y)+\lambda g(x,y)$$

原优化问题取得极值，等价于拉格朗日函数取得驻点。即：

$$\nabla L(x,y,\lambda)=0$$

这样，就将一个**带约束优化问题，转化成了一个无约束的极值问题**。

上述二元函数在一个等式约束条件下的极值问题可以推广到 n 元函数 $f(x_1,...,x_n)$ 在 M 个等式约束 $g_k(x_1,...,x_n)=0$，$k=1,...,M$ 的一般情形。这时候拉格朗日乘子法所寻找的点对应的梯度并不是 f 某个约束的梯度的倍数，而是所有约束的梯度的线性组合。

$$\nabla f(x_1,...,x_n)=-\sum_{k=1}^{M}\lambda_k\nabla g_k(x_1,...,x_n)$$

$$L(x_1,...,x_n,\lambda_1,...,\lambda_M)=f(x_1,..,x_n)+\sum_{k=1}^{M}\lambda_k g_k(x_1,...,x_n),$$

下面讨论更一般的情形，其中约束条件不但包含等式约束，还包含不等式约束。

拉格朗日对偶性(Lagrange duality)

在约束最优化问题中，常常利用**拉格朗日对偶性**将原始问题转化为对偶问题。通过解对偶问题而得到原始问题的解.

- **原始问题**(primal problem)

假设 $f(x),g_i(x),h_j(x)$ 是定义在 R^n 上的连续可微函数。考虑如下最优化问题

$$\min_{x\in R^n}\ f(\boldsymbol{x})$$
$$s.t.\quad g_i(x)\le 0,\quad i=1,2,\cdots,k$$
$$h_j(x)=0,\quad j=1,2,\cdots,l$$

称此约束最优化问题为**原始最优化问题**或**原始问题**.

引入**广义拉格朗日函数**

$$L(x,\alpha,\beta)=f(x)+\underbrace{\sum_{i=1}^{k}\alpha_i g_i(x)}_{\text{不等式约束}}+\underbrace{\sum_{j=1}^{l}\beta_j h_j(x)}_{\text{等式约束}}$$

这里，α_i,β_j 是拉格朗日乘子，且 $\alpha_i\ge 0$（因为不等式有方向性）。考虑 $\boldsymbol{x}$ 的函数：

$$\theta_P(x) = \max_{\alpha,\beta;\alpha_i \geq 0} L(x,\alpha,\beta)$$

这里下标 P 表示**原始问题**.

下面通过 $\boldsymbol{x}$ 是否满足约束条件两方面来分析这个函数:

1. 考虑某个 $\boldsymbol{x}$ 违反了原始的约束，即某个 $g_i(x) > 0$ 或 $h_j(x) \neq 0$ ，那么

$$\theta_P(x) = \max_{\alpha,\beta;\alpha_i \geq 0} [f(x) + \sum_{i=1}^{k} \alpha_i g_i(x) + \sum_{j=1}^{l} \beta_j h_j(x)] = +\infty$$

因为若 $g_i(x) > 0$ ，则令 $\alpha_i \to +\infty$ ，若 $h_j(x) \neq 0$ ，则很容易取值 β_j ，使 $\beta_j h_j(x) \to +\infty$ 。

2. 考虑某个 $\boldsymbol{x}$ 满足原始的约束，则

$$\theta_P(x) = \max_{\alpha,\beta;\alpha_i \geq 0} [f(x)] = f(x)$$

综上可得：

$$\theta_P(x) = \begin{cases} f(x) & x\text{满足原始问题约束} \\ +\infty & \text{其它} \end{cases}$$

那么则满足原始问题约束时，

$$\min_{\boldsymbol{x}} \theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\boldsymbol{x}} f(\boldsymbol{x})$$

即 $\min_x \theta_P(x)$ 与原始优化问题等价。于是，**原始约束最优化问题就变成了无约束的拉格朗日函数的极小极大问题**。

定义**原始问题（primal problem）的最优值为** p^* ，即

$$p^* = \min_{\boldsymbol{x}} \theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

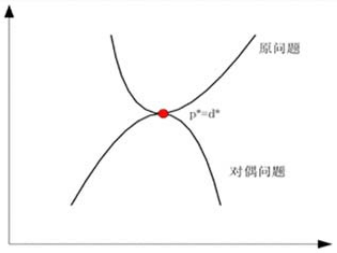
- **对偶问题(dual problem)**

定义关于 $\boldsymbol{\alpha},\boldsymbol{\beta}$ 的函数

$$\theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

再考虑极大化上式，即

$$\max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} \theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$



这就是原始问题的**对偶问题。**再把原始问题写出来

$$\min_{\boldsymbol{x}} \theta_P(\boldsymbol{x}) = \min_{\boldsymbol{x}} \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

形式上可以看出很对称，只不过原始问题是先固定 $L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$ 中的 $\boldsymbol{x}$ ，优化出参数 $\boldsymbol{\alpha},\boldsymbol{\beta}$ ，再优化最优 $\boldsymbol{x}$ ，而对偶问题是先固定 $\boldsymbol{\alpha},\boldsymbol{\beta}$ ，优化出最优 $\boldsymbol{x}$ ，然后再确定参数 $\boldsymbol{\alpha},\boldsymbol{\beta}$ 。

定义**对偶问题的最优值为** d^* ，即

$$d^* = \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} \theta_D(\boldsymbol{\alpha},\boldsymbol{\beta}) = \max_{\boldsymbol{\alpha},\boldsymbol{\beta};\alpha_i \geq 0} \min_{\boldsymbol{x}} L(\boldsymbol{x},\boldsymbol{\alpha},\boldsymbol{\beta})$$

由于

$$\theta_D(\alpha,\beta) = \min_x L(x,\alpha,\beta) \leq L(x,\alpha,\beta) \leq \max_{\alpha,\beta;\alpha_i \geq 0} L(x,\alpha,\beta) \leq \theta_p(x)$$

所以

$$\theta_D(\alpha,\beta) \leq \theta_p(x)$$

上面这是式子说明，$\theta_D(\alpha,\beta)$的所有的解，都不大于$\theta_p(x)$的解。那么，对偶问题的最优解和原始问题的最优解也满足这个式子：

$$d^* = \max_{\alpha,\beta;\alpha_i\geq 0} \theta_D(\alpha,\beta) \leq \min_x \theta_p(x) = p^*$$

即

$$d^* \leq p^*$$

即对偶问题的最优值为原始问题的最优值的下界。这个不等式称为**弱对偶性**(weak duality)。

在一些常见的问题中，只要满足一定的条件，就可以使得$d^* = p^*$，这种情形称为**强对偶性**。比如支持向量机算法中的凸二次规划问题。下面的定理 1 和定理 2 分别给出了$d^* = p^*$需要满足的充分条件和必要条件。

定理 1（强对偶性的充分条件）

考虑原始问题和对偶问题. 设原始问题是**凸优化**问题，即函数$f(x)$和$g_i(x)$是**凸函数**，$h_j(x)$是**仿射函数**（1 阶多项式函数）；并且假设不等式约束$g_i(x)$是**严格可行**的，即存在x，对所有i有$g_i(x)<0$，则存在x^*和α^*,β^*，使x^*是原始问题的最优解，α^*,β^*是对偶问题的最优解，并且

$$p^* = d^* = L(x^*,\alpha^*,\beta^*)$$

定理 2（强对偶性的必要条件）

对于原始问题和对偶问题，x^*和α^*,β^*分别是原始问题和对偶问题的最优解且 满足

$$p^* = d^* = L(x^*,\alpha^*,\beta^*)$$

的必要条件是x^*和α^*,β^*满足 **KKT 条件**（ Karush-Kuhn-Tucker Conditions）：

$$(1)\quad \nabla_x L(x^*,\alpha^*,\beta^*) = 0$$

$$(2)\quad g_i(x^*) \leq 0; \quad i = 1,2...k$$

$$\alpha_i^* \geq 0; \quad i = 1,2...k$$

$$h_j(x^*) = 0; \quad j = 1,2,...l$$

$$(3)\quad \alpha_i^* g_i(x^*) = 0; \quad i = 1,2...k$$

上述的 KKT 条件，可以这样理解：函数$L(x,\alpha,\beta)$ 是以x,α,β为参数的，那么其最优解x^*必然满足函数$L(x,\alpha,\beta)$ 的梯度为 0，这就是 KKT 条件的条件（1）：

$$\nabla_x L(x^*,\alpha^*,\beta^*) = 0$$

又因为说最优解必须满足所有等式及不等式限制条件，也就是说最优解必须是一个可行解。这样得到条件（2）：

$$g_i(x^*) \leq 0; \quad i = 1,2...k$$

$$\alpha_i^* \geq 0; \quad i = 1,2...k$$

$$h_j(x^*) = 0; \quad j = 1,2,...l$$

条件（3）推导如下：

因为

$$
\begin{aligned}
f(x^*) &= \theta_D(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) \\
&= \min_x \left(f(x) + \sum_{i=1}^{k} \alpha_i^* g_i(x) + \sum_{i=1}^{l} \beta_i^* h_i(x) \right) \\
&\le f(x^*) + \sum_{i=1}^{k} \alpha_i^* g_i(x) + \sum_{i=1}^{l} \beta_i^* h_i(x) \le f(x^*)
\end{aligned}
$$

由于两头是相等的，所以这一系列的式子里的不等号全部都可以换成等号。又因为 $\alpha_i^* g_i(x^*) \le 0$ ，因此可以得到

$$
\alpha_i^* g_i(x^*) = 0; \quad i = 1, 2 \ldots k
$$

条件（3）也被称作**互补性条件**。

因为 $\alpha_i^* g_i(x^*) = 0$，所以 α_i^* 与 $g_i(x^*)$ 只要一个不为 0，另一个就必为 0。当 $g_i(x^*) < 0$ 时，x^* 是处于可行域的内部，这时不等式约束不起作用（因为 $\alpha_i^* = 0$ ）。而 $\alpha_i^* > 0$ 的点肯定是可行域边界的点（因为 $g_i(x^*) = 0$ ），这时候不等式约束才是**有效约束**。