信息熵

参考资料：https://www.cnblogs.com/kyrieng/p/8694705.html

1. 信息量度量的本质

信息量的度量就等于不确定性的多少

> e.g. alpha go刚出世时，大家都非常关心它和李世石的对战结果，因为这时我们还不知道alpha go的水平如何，输赢的概率都是一半，因此对战结果包含的信息量是很大的；如果现在再让alpha go和李世石对战，我们不会有任何兴趣去了解结果如何，因为我们有接近100%的把握认为alpha go会赢，结果是确定的，所以不会给我们提供任何有帮助的信息。

2. 为什么是对数（$\log x$）的形式？（从自信息说起）

考虑一个**离散的随机变量** X，由上面的例子可知，信息的量度应该依赖于概率分布 *p*(*x*)，因此我们想要寻找一个函数 *I*(*x*)，它是概率 *p*(*x*) 的单调函数，表达了信息的内容。怎么寻找呢？如果我们有两个不相关的事件 *x* 和 *y*，那么观察两个事件同时发生时获得的信息量应该等于观察到事件各自发生时获得的信息之和，即

$$I(x,y) = I(x) + I(y)$$

因为两个事件是独立不相关的，因此

$$p(x,y) = p(x) * p(y)$$

根据这两个关系，很容易看出 **I(x)一定与 p(x)的对数有关。**因此，我们有

$$I(x) = -\log p(x)$$

其中负号是用来保证信息量大于等于0。而 log 函数基的选择是任意的（信息论中基常常选择为2，因此信息的单位为比特bits；而机器学习中基常常选择为自然常数e，因此单位常常被称为奈特nats）。$I(x)$也被称为随机变量 X的**自信息 (self-information)**，描述的是随机变量的某个事件发生所带来的信息。

3. 信息熵的定义及性质

最后，我们正式引出信息熵。 现在假设一个发送者想传送一个随机变量的值给接收者。那么在这个过程中，他们传输的平均信息量可以通过求 $I(x) = -\log p(x)$关于概率分布 *p*(*x*)的期望得到，即

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

$H(x)$就被称为随机变量 X的**熵**,它是表示随机变量不确定的度量，是对所有可能发生的事件产生的信息量的期望。从公式可得，**随机变量的取值个数越多，状态数也就越多，信息熵就越大，混乱程度就越大。当随机分布为均匀分布时，熵最大。**

> 1、熵只依赖于随机变量的分布,与随机变量取值无关，所以也可以将 X的熵记作$H(p)$
>
> 2、定义$0\log 0 = 0$(因为某个取值概率可能为0)

熵的取值范围满足以下不等式

$$0 \leq H(X) \leq \log n$$

证明（使用拉格朗日乘子法）：

目标函数：

$$f(p(1),p(2),\ldots,p(n)) = -(p(1)\log p(1) + p(2)\log p(2) + \ldots + p(n)\log p(n))$$

约束条件：

$$p(1) + p(2) + \ldots + p(n) = 1$$

构造拉格朗日函数：

$$\begin{aligned} L(p(1),p(2),\ldots,p(n),\lambda) = &-(p(1)\log p(1) + p(2)\log p(2) + \ldots + p(n)\log p(n)) \\ &+ \lambda(p(1) + p(2) + \ldots + p(n) - 1) \end{aligned}$$

分别对$p(1),p(2)\ldots,p(n),\lambda$求偏导，得到

$$\begin{aligned} \lambda - log(e \cdot p(1)) &= 0 \\ \lambda - log(e \cdot p(2)) &= 0 \\ &\cdots \\ \lambda - log(e \cdot p(n)) &= 0 \\ p(1) + p(2) + \ldots + p(n) &= 1 \end{aligned}$$

解得

$$p(1) = p(2) = \ldots = p(n) = \frac{1}{n}$$

此时

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

Jensen不等式

若$f(x)$是区间上的凹函数，则有

$$f(\mathbb{E}[x]) \leq \mathbb{E}[f(x)]$$

指数分布族

我们说如果一个分布是指数族分布，那么它可以用以下形式表示：

$$p(y;\eta) = b(y)exp(\eta^T T(y) - a(\eta))$$

这里，η被称为分布的自然参数（也称为规范参数）；T(y)是充分统计量（对于我们所考虑的分布，通常情况下有T(y)=y）；a(η)被称为对数划分函数。这一项

$$e^{-a(\eta)}$$

本质上是起到了正则化常数的作用，确保了分布p(y;η)的总和或是积分在y到1上。

固定T，a和b，我们定义一族以η为参数的分布；随着η的变化，我们可以在这个族中得到不同的分布。

常见的一些分布，如Bernouli分布和Gaussian分布，都属于指数分布族。

- Bernouli

$$\begin{aligned} p(y;\phi) =& \phi^y(1-\phi)^{1-y} \\ =& exp(y\log\phi + (1-y)\log(1-\phi)) \\ =& exp(y\log(\frac{\phi}{1-\phi}) + \log(1-\phi)) \end{aligned}$$

自然参数由$\eta = \log(\frac{\phi}{1-\phi})$给出，可以得出$\phi = \frac{1}{1+e^{-\eta}}$，也就是Sigmoid函数的形式。此时容易得到

$$\begin{aligned} T(y) &= y \\ a(\eta) &= -\log(1-\phi) = \log(1+\phi) \\ b(y) &= 1 \end{aligned}$$

- Gaussian

$$\begin{aligned} p(y;\mu) =& \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}(y-\mu)^2) \\ =& \frac{1}{\sqrt{2\pi}}exp(-\frac{1}{2}y^2)\exp(\mu y - \frac{1}{2}\mu^2) \end{aligned}$$

易得

$$\begin{aligned} \eta &= \mu \\ T(y) &= y \\ a(\eta) &= \mu^2/2 = \eta^2/2 \\ b(y) &= \frac{1}{\sqrt{2\pi}}exp(-\frac{y^2}{2}) \end{aligned}$$

拉格朗日乘子法

参考资料：https://www.cnblogs.com/sddai/p/5728195.html

在求解最优化问题中，**拉格朗日乘子法**（Lagrange Multiplier）和**KKT**（Karush Kuhn Tucker）**条件**是两种最常用的方法。在有等式约束时使用拉格朗日乘子法，在有不等式约束时使用KKT条件。

> 最优化问题：通常是指对于给定的某一函数，**求其在指定作用域上的全局最小值(最大值问题可以转化成最小值问题)。**

一般情况下，最优化问题会碰到一下三种情况：

（1）无约束条件

这是最简单的情况，解决方法通常是函数对变量求导，导数等于0的点可能是极值点，将结果带回原函数进行验证即可。

（2）等式约束条件

设目标函数为f(x)，约束条件为$h_k(x)$，形如:

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

解决方法是**消元法**或者**拉格朗日法。**消元法比较简单不在赘述，这里主要讲拉格朗日法，因为后面提到的KKT条件是对拉格朗日乘子法的一种泛化。

> 算法-拉格朗日乘子法

1. 构造拉格朗日函数（在每个约束条件前引入一个系数λ，称为拉格朗日乘子）

$$\mathcal{L}(x,\lambda)=f(x)+\sum_k \lambda_k(h_k(x))$$

2. 对函数中每个变量求偏导

$$\begin{aligned} \frac{\partial \mathcal{L}(x,\lambda)}{\partial x}&=0 \\ \frac{\partial \mathcal{L}(x,\lambda)}{\partial \lambda_k}&=0 \end{aligned}$$

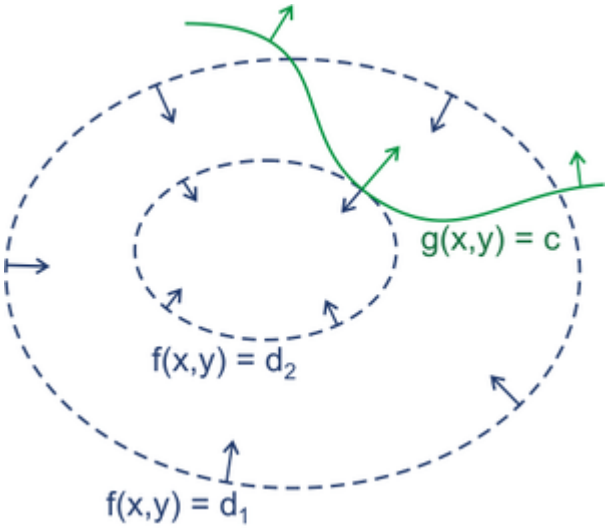
若有l个约束条件，那么一共可以得到l+1个等式，求出的方程组的解就是原始优化函数的极值，将结果带回原方程验证即可。

> 算法原理

以二元优化问题为例

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

下图画出了z=f(x,y)的等高线：



绿线标出的是约束$g(x,y)=c$的点的轨迹。蓝线是f(x,y)的等高线。箭头表示斜率，和等高线的法线平行。从梯度的方向上来看，显然有$d_1 > d_2$。绿色的线是约束，也就是说，只有正好落在这条绿线上的点才可能是满足要求的点。如果没有这条约束，$f(x,y)$的最小值应该会落在最小那圈等高线内部的某一点上。

考虑约束的情况下，最小值点应该在哪里呢？显然应该是在**$\mathbf{f(x,y)}$的等高线正好和约束线相切的位置**，因为如果只是相交意味着肯定还存在其它的等高线在该条等高线的内部或者外部，使得新的等高线与目标函数的交点的值更大或者更小，只有当等高线与目标函数的曲线相切的时候，才可能取得最优值。

如果我们对约束也求梯度$\nabla g(x,y)$，则其梯度如图中绿色箭头所示。很容易看出来，**要想让目标函数的$\mathbf{f(x,y)}$等高线和约束相切，则他们切点的梯度一定平行。**

也即在最优化解的时候：$\nabla f(x,y) = \lambda \nabla(g(x,y))$ （其中∇为梯度算子; 即：f(x)的梯度 = λ* g(x)的梯度，λ是常数,可以是任何非0实数，表示左右两边同向）。于是有

$$\nabla_{x,y}\mathcal{L}(x,y,\lambda) = \nabla_{x,y}[f(x,y) + \lambda(g(x,y) - C)] = 2\nabla_{x,y} f(x,y)$$

拉格朗日函数对x,y的极值点也是$f(x,y)$的极值点，而对乘子λ的偏导为0保证了约束条件$g(x,y)=C$成立。因此，求解拉格朗日函数的极值点等价于求解原始的最优化问题。

（3）不等式约束条件

设目标函数f(x)，不等式约束为g(x)，有的教程还会添加上等式约束条件h(x)。此时的约束优化问题描述如下：

$$\begin{aligned} & \min f(x) \\ s.t. \quad & h_k(x) = 0 \quad k = 1,2,\ldots,K \\ & g_l(x) \le 0 \quad l = 1,2,\ldots,L \end{aligned}$$

我们定义不等式约束下的拉格朗日函数为：

$$\mathcal{L}(x,\lambda,\alpha) = f(x) + \sum_k \lambda_k h_k(x) + \sum_l \alpha_l g_l(x), \quad \alpha \ge 0$$

KKT条件是说**最优值**必须满足以下条件：

1）$\frac{\partial \mathcal{L}(x,\lambda,\mu)}{\partial x} = 0$

2）$\frac{\partial \mathcal{L}(x,\lambda,\mu)}{\partial \lambda_k} = 0$即 $h_k(x) = 0$

3) $\alpha_l g_l(x) = 0$

求取这些等式之后就能得到候选最优值。其中第三个式子非常有趣，因为$\alpha \geq 0, g(x) <= 0$，如果要满足这个等式，必须$\alpha = 0$或$g(x) = 0$. 这是SVM的很多重要性质的来源，如支持向量的概念。详细推导略。

当约束条件为$g(x) \geq 0$时，我们可以通过对$g(x)$取负使其满足上述形式，此时拉格朗日函数为

$$\mathcal{L}(x, \lambda, \alpha) = f(x) + \sum_k \lambda_k h_k(x) - \sum_l \alpha_l g_l(x), \quad \alpha \geq 0$$