Fundamental of Big Data Inference

Yin Tianyou

Shandong University
School of Journalism and Communication

June 28, 2023

Randomness
数据的离散

1 KL 散度

KL散度衡量的是两个分布$p(x)$和$q(x)$之间的相似程度，其定义如下

$$D\left(p(x)||q(x)\right)=\sum p_X\left(x\right)\log\frac{p_X\left(x\right)}{q_X(x)}=\mathbf{E}_{pX}\left[\log\frac{p_X(x)}{q_X(x)}\right] \tag{1}$$

需要注意的是，log的底数决定了单位. 底数是2，单位是比特（$bits$）；底数是e，单位是奈特（$nats$）.
我们可以通过推断来反映信息缺失. 结果越不确定，随机变量越不可被预测，推断更不准确.
KL散度永远大于等于零. 当$p(x)=q(x)$时等于零.
KL散度具有不对称性，即$D(p(x)||q(x))$不等于$D(q(x)||p(x))$.

2 香农熵

2.1 香农熵

香农熵是一个数学上颇为抽象的概念，在这里不妨把香农熵理解成某种特定信息的出现概率（离散随机事件的出现概率）. 一个系统越是有序，香农熵就越低；反之，一个系统越是混乱，香农熵就越高. 香农熵也可以说是系统有序化程度的一个度量. 在信息论里，香农熵可以用来衡量信息不确定性的大小. 香农熵的定义式为：

$$H(X)=-\sum_{i=1}^{n}P(x_i)\log_2P(x_i) \tag{2}$$

相比只能刻画数字离散程度的方差，香农熵的适用范围更广. 它只考虑各个变量出现的概率.
香农熵永远大于等于零. 当X是确定的时候，$H(X)=0$.

2.2 联合熵与条件熵

对于联合变量X,Y，我们定义联合熵为

$$H(X,Y)=-\sum_{x,y}p_{X,Y}\left(x,y\right)\log_{X,Y}\left(x,y\right) \tag{3}$$

对于条件变量$X|Y$，我们定义条件熵为

$$H(X|Y)=-\sum_{x}p_{X|Y}\left(x|y\right)\log_{X|Y}\left(x|y\right) \tag{4}$$

条件熵$H(X|Y)$衡量了X的平均离散程度.
对于所有的y,我们有

$$H\left(X|Y\right)=\sum_{y}H\left(X|Y=y\right)=-\sum_{x,y}p_{X,Y}\log p_{X|Y}\left(x|y\right)$$

$$H(X)-H(X|Y)=D(p_{X,Y}(x,y)||p_X(x)p_Y(y))\geq0$$

当且仅当X与Y相互独立、$p_{X,Y}(x,y)=p_X(x)p_Y(y)$的时候有

$$D\left(X||Y\right)=0$$

联合熵与条件熵有如下关系：

$$H\left(X,Y\right)=H\left(X\right)+H\left(Y|X\right)=H\left(Y\right)+H\left(X|Y\right)$$

此外，

$$H(X,Y) \leq H(X)+H(Y)$$

上式在$H(X)=H(Y)$时取等号，此时$H(X|Y) \leq H(X)$，$H(Y|X) \leq H(Y)$.

3 平均互信息(MI)

平均互信息可以衡量两个变量的相关程度. 平均互信息可以看成是一个随机变量所包含的关于另外一个随机变量的信息量，或者是一个随机变量由于已知的另一个随机变量而减少的不确定性.

平均互信息用$I(X;Y)$表示：

$$I(X;Y)=\sum_{X,Y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{5}$$

此外，有

$$I(X;Y)=H(X)-H(X|Y)=D(p(x,y)||p(x)p(y))$$

平均互信息有对称性。即$I(X;Y)=I(Y;X)$.

当且仅当X与Y独立的时候有$I(X;Y)=0$. 反之亦然.

我们有$I(X;X)=H(X)$.

我们可以使用韦恩图Fig. 1和维拉图Fig. 2来表示平均互信息：

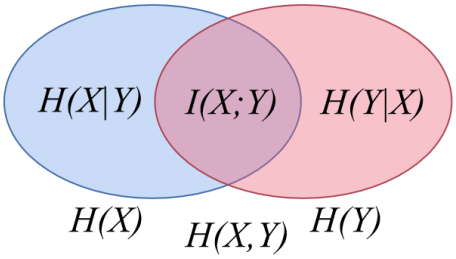

Fig. 1. 韦恩图表示

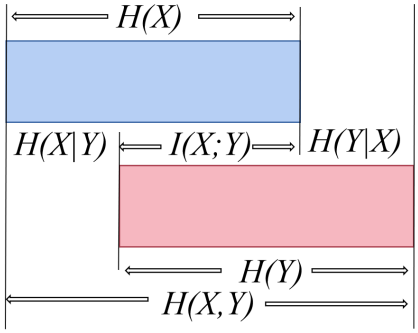

Fig. 2. 维拉图表示

由图可知

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

References

1 徐政五, 甘露, 汪利辉. 信息论导引（第2版）[M]. 电子科技大学出版社, 2017.

Decision-Making
决策推断

决策推断（$Decision-Making$）是由y反推回x的过程. 这一过程需要寻找一个$\hat{x}$，我们称之**“猜测”**（$guess$）. 我们的目标是寻找一个“$Intelligent\ guess$”，使得$\hat{x}$ **在平均意义上尽可能地接近**x. 如果x已经被观察到了（$observed$），那么我们有$\hat{x} \to x$（观测数据由隐随变量产生；在引入观测数据后，X的分布就变成了条件分布）. 反之，我们则需要寻找一个尽可能接近所有可能的样本值（$Sample\ value$）的x. 这个时候可以考虑x的分布列，寻找概率最大的取值.

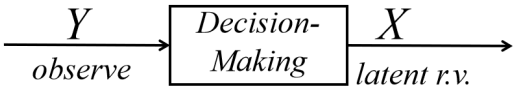

Fig. 1. 关于决策推断的图示. Obs.Y should help us do a better job in our Decision-Making——消除不确定性；提高精准度.

1 代价函数

人们需要从已知的数据当中获取有效信息，并以此为根据帮助自己进行更好的决策，而决策的质量又与数据质量还有算法的质量有关。而当我们进行猜测的时候猜测值与真实值之间必定会产生一个误差值，由此我们引入了代价函数（$cost\ f.$）与风险函数（$risk\ f$）.

1.1 代价函数（$Cost\ function$）

代价函数（$Cost\ function$）是表示估计值$\hat{x}$ 与真实值x 之间的差异的函数. 又叫损失函数或成本函数，一个优化问题试图最小化代价函数. 代价函数有助于我们弄清楚如何把最有可能的函数与我们的数据相拟合. 代价函数一般为下面的形式：

$$c(x,\hat{x}) \tag{1}$$

1.2 风险函数（$Risk\ function$）

风险函数（$risk\ function$）又称期望损失（$expected\ loss$），是损失函数的期望，度量平均意义下模型预测的好坏. 风险函数一般为下面的形式：

$$\varphi(x) = E[c(X,x)] = \sum_{x\in\mathcal{X}} c(x,\hat{x})\, p_X(x) \tag{2}$$

让风险函数取最小值的估计值为最优估计值. 即：

$$\hat{x}^* = \arg\min_x \varphi(x) \tag{3}$$

2 基于代价函数的决策推断

2.1 最小概率误差估计（MPE）（也称MAP 估计）

我们有：

$$c(x,\hat{x}) = \begin{cases} 0\,, x = \hat{x} \\ 1\,, otherwise \end{cases} \tag{4}$$

上式被称为0-1损失函数. 可以看出，该损失函数的意义就是，当预测错误时，损失函数值为1，预测正确时，损失函数值为0. 该损失函数不考虑预测值和真实值的误差程度，也就是只要预测错误，预测错误差一点和差很多是一样的. 它实际上就是**将风险函数写成二进制的结果**，即非1即0，在此基础上得出了风险函数与概率之间的关系（$=1-p_x$），所以有风险函数最小值=概率最大值，也就是在数据中出现最多的数据，也就是我们规定的“众数”.

$$\varphi(x) = \mathbf{E}[c(X,\hat{x})] = \sum_{x:x\neq\hat{x}} p_X(x) = P(X\neq\hat{x}) = 1 - P(X=\hat{x}) = 1 - p_X(\hat{x}) \tag{5}$$

这个式子引入了分布列，其中$\varphi(\hat{x}) \to 1 - p_X(\hat{x})$. 在这个时候，我们有

$$\hat{x}^*_{\mathrm{MPE}} = \arg\max_x p_X(x) \tag{6}$$

上式基于$p_X(x)$分布.

例1 抛掷一枚不均匀（*biased*）的硬币. 假设各个面出现的概率如下：

$$p_X(x) = \begin{cases} 2/3, x = H \\ 1/3, x = T \end{cases}$$

在这个时候，我们有

$$\hat{x}^*_{\mathrm{MPE}} = \arg\max_x p_X(x) = H$$

如果硬币是均匀的（*unbiased*），则

$$p_X(x) = \begin{cases} 1/2, x = H \\ 1/2, x = T \end{cases}$$

于是，我们有

$$\hat{x}^*_{\mathrm{MPE}} = \arg\max_x p_X(x) = H = T$$

同香农熵一样，MPE估计的优势在于只考虑概率而不考虑变量类型是否为数值（对符号变量和数字变量均适用）.

对于多个变量$x_1, x_2, ..., x_n \in \mathcal{X}$，我们有

$$\hat{x}_1, \hat{x}_2, ..., \hat{x}_n = \arg \max_{x_1, x_2, ..., x_n} p_{x_1^n}(x_1^n).$$

例2 抛两枚六面骰子，每次实验的总点数认为是随机变量X的一个取值. 请写出X的PMF,以及MPE、MMSE估计时的取值（有计算过程）. X的PMF为：

$$p_X(x)=\begin{cases}\frac{1}{36}, & x=2\\ \frac{2}{36}, & x=3\\ \frac{3}{36}, & x=4\\ \frac{4}{36}, & x=5\\ \frac{5}{36}, & x=6\\ \frac{6}{36}, & x=7\\ \frac{5}{36}, & x=8\\ \frac{4}{36}, & x=9\\ \frac{3}{36}, & x=10\\ \frac{2}{36}, & x=11\\ \frac{1}{36}, & x=12\\ 0, & \text{otherwise}\end{cases}$$

MPE估计是使概率$p_X(x)$最大的x.由上面的PMF知，X的MPE估计为：

$$\hat{x}_{\text{MPE}}=\arg\max_x p_X(x)=7$$

MMSE估计（最小均方误差估计）是在所有可能的估计中，均方误差（MSE）最小的估计。因为X是一个离散随机变量，所以可以直接根据$\hat{x}^*_{\text{MMSE}}=\mathbf{E}[x]$计算MMSE估计：

$$\hat{x}^*_{\text{MMSE}}=\mathbf{E}[x]=2\times\frac{1}{36}+3\times\frac{2}{36}+4\times\frac{3}{36}+5\times\frac{4}{36}+6\times\frac{5}{36}+7\times\frac{6}{36}+8\times\frac{5}{36}+9\times\frac{4}{36}+10\times\frac{3}{36}+11\times\frac{2}{36}+12\times\frac{1}{36}=7$$

2.2 最小均方误差估计（$MMSE$）

最小均方估计的代价函数为：

$$c(x,\hat{x}) = (x-\hat{x})^2 \tag{7}$$

其风险函数为：

$$\varphi(x) = \sum_{x\in\mathcal{X}} (x-\hat{x})^2 p_X(x) = \mathbf{D}(X) \tag{8}$$

对其进行求导，得

$$\varphi'(x) = 2\sum_x (\hat{x}-x)p_X(x) = 2\left(\sum_x \hat{x}p_X(x) - \sum_x xp_X(x)\right) = 2(\hat{x}-\mathbf{E}[x]) = 0$$

其中，

$$\hat{x}^*_{\mathrm{MMSE}} = \mathbf{E}[x] \tag{9}$$

于是，我们有$min\varphi(\hat{x}) \to \varphi'(\hat{x}) = 0$. 于是，

$$\varphi(\hat{x}) = \sum \left(x - E[x]^2\right) p_X(x) = D(x) \tag{10}$$

上式对于多个随机变量仍然有效：

$$(\hat{x}_1, \hat{x}_2, ..., \hat{x}_n)_{\mathrm{MMSE}} = (E[x_1], E[x_2], ..., E[x_n])$$

公式(10)揭示了期望和方差的一种关系.

2.3 边缘MPE估计

经过观测到的或者相关的数据可以提高估计的精准度，并且利于代入观测数据. 观测数据（$Obs.Y$）的引入使得x变为条件分布，利于消除不确定性.

我们可以用条件概率来代替先验概率（$prior$）：

$$\hat{x}_{\mathrm{MPE}} = \arg_{\hat{x}} \max p_{X|Y}(x|y) \tag{11}$$

其中，$p_{X|Y}(x|y)$称为后验分布（$posterior\ dist.$）.上式称为最大后验估计（$maximum\ a\ posterior\ estimation$）.

边缘MPE估计是一种贝叶斯参数估计方法. 它通过最大化边缘后验概率来估计参数. 设有未知参数x和数据y. 根据贝叶斯定理，x的后验概率可以表示为:

$$p_{Y|X}(y|x) = \frac{p_X(x)\,p_{Y|X}(y|x)}{p_Y(y)}$$

其中,$p(x)$是x的先验概率，$p(y|x)$是x下y的似然函数，$p(y)$是y的边缘概率. 在MPE估计中，我们寻找使$p(y|x)$最大化的x值，作为x的估计值. 而在边缘MPE估计中，我们直接最大化x的边缘后验概率来获得x的估计值，而不考虑$p(y)$这个常数项.

x的边缘MPE估计可以表示为:

$$\hat{x}^*_{\mathrm{MAP}} = \arg_{\hat{x}} \max \frac{p_X(\hat{x})\,p_{Y|X}(y|\hat{x})}{p_Y(y)} = \arg\max\left(p_X(\hat{x})\,p_{Y|X}(y|\hat{x})\right) \tag{12}$$

与普通MPE估计相比，边缘MPE估计在计算上更简单，因为它忽略了$p(y)$这个常数项. 但是，它的估计效率却并不低于MPE估计。在某些情况下，边缘MPE估计的性能甚至优于MPE估计.

注意：$\hat{x}$是$Obs.Y$的一个函数；$p_{X|Y}$关于X后验分布；根据贝叶斯规则，有公式(12).
此外，当有$Y=y$的观测值时，x的MMSE估计为：

$$\hat{x}_{\mathrm{MMSE}}=\mathbf{E}(X|Y=y)=\sum_{x}xp_{X|Y}(x|y) \tag{13}$$

此时的风险函数为：

$$\varphi(x)=\sum_{x}(x-\mathbf{E}[x|Y=y])^{2}p_{X|Y}(x|y)=D(x|Y=y) \tag{14}$$

上面两个式子也是期望和方差的一种关系.

例3 写出$p_{X_2,X_1}(x_2,x_1)$的边缘MPE估计.

$p_{X2,X1}(x_2,x_1)$	x_2=1	x_2=2
x_1=1	0.35	0.05
x_1=2	0.3	0.3

Fig. 2. 例3 图

由边缘MPE的定义可知：
x_1的边缘分布和边缘MPE估计为

$$p_{X_1}(x_1)\begin{matrix}0.4, x_1=1\\0.6, x_1=2\end{matrix}$$

$$\hat{x}_{1MPE}=2$$

x_2的边缘分布和边缘MPE估计为

$$p_{X_2}(x_2)\begin{matrix}0.65, x_2=1\\0.35, x_2=2\end{matrix}$$

$$\hat{x}_{2MPE}=1$$

2.4 最小绝对误差估计（MAE估计）

MAE估计（$Minimum\ Absolute-Error\ Estimation$）是最小绝对误差估计. 代价函数为：

$$c(x,\hat{x})=|x-\hat{x}| \tag{15}$$

基于观测值$Y=y$的风险函数为：

$$\varphi(\hat{x})=\sum|x-\hat{x}|p_{X,Y}(x,y) \tag{16}$$

令$x=a$，将其拆分为两部分：

$$\varphi(\hat{x})=\sum|x-x|p_{X|Y}(x|y)=\sum_{-\infty}^{a}(a-x)p_{X|Y}(x|y)+\sum_{a}^{+\infty}(x-a)p_{X|Y}(x|y)$$

其中$\sum_{-\infty}^{a}(a-x)p_{X|Y}(x|y)$是下界，$\sum_{a}^{+\infty}(x-a)p_{X|Y}(x|y)$是上界. 对其求导，得

$$\varphi(a)'=\sum|x-x|p_{X|Y}(x|y)=\sum_{-\infty}^{a}(a-x)p_{X|Y}(x|y)+\sum_{a}^{+\infty}(x-a)p_{X|Y}(x|y)=0$$

上式中上下界各占一半，可以使得导数等于零. 因此，$\hat{x}_{\mathrm{MAE}}(y)$就是$p_{X|Y}(x|y)$的中位数.

下面对于之前的一些略语做出解释：

Table 1. 略语表

简称	英文表示	中文表示	数学表示
r.v.	random variable	随机变量	-
arg	argument	自变量	-
dist.	Distribution	分布	-
Obs.	Observation	观察资料（数据）	-
Est.	Estimation	估计	-
Prob.	Probability	概率	-
Ex.	Example	例子	-
w.r.t.	with respect to	关于、谈到、涉及	-
i.e.	id est	即，换句话说	-
MPE	Minimum Probability of Error Estimation	最小概率误差估计	$c(x,\hat{x})$,$\varphi(x)$
MMPE	Minimum Mean-Square Error Estimation	最小均方误差估计	$\hat{x}^*_{\mathrm{MMSE}}=\mathbf{E}[x]$
MSE	Marginal Posteriori Estimation	边际后验估计	-
MPC	-	最大可能序列	$\arg\max_{x_1,\ldots,x_n}(p(x_1^n,\ldots,x_n^n))$
MAP	Maximum a Posterior Estimation	最大后验估计	$\arg(p_X(\hat{x})p_{Y\|X}(y\|\hat{x}))$
MAE	Minimum Absolute-Error Estimation	最小绝对误差估计	-

在前面广为提到的“$\arg\max/\min f(x)$”，意思就是“当函数取得最值的时候自变量的取值”.

我们再来看一下两个似然函数（*Likelihood function*）与观测模型. 观测数据由隐随变量产生.
$p_{Y|X}(y|\cdot)$是一个特殊的/特定的y的观测数据（$aparticular obs.y$）的一个似然函数.
$p_{Y|X}(\cdot|x)$是特定隐随变量的观测模型X（Observation model X for a particular latent r.v.）.

3 贝叶斯决策估计

针对Y的观测值$y_1,y_2,\ldots,y_n$，我们一个一个把它们输入到决策估计里，每一步的后验随着新的观测值的引入而变成先验，通过对数据的输入减少了信息的不确定性. 通过输入了观测值（数据），经过了算法的运算之后，得到了数据x的特征，即隐随机变量，是所谓“信念修正”（*Belief revision*）：

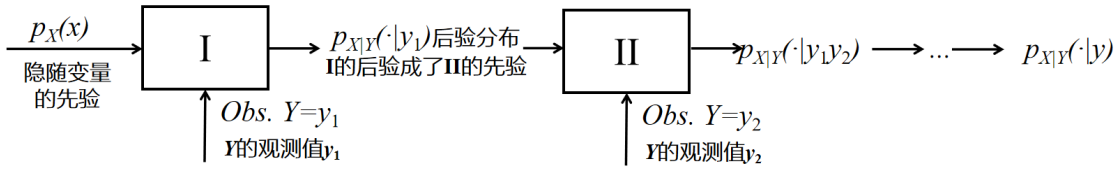

Fig. 3. “信念修正”图示

输入的数据包含了Y的观测值、x的先验、$p_{X|Y}(\cdot|y)$的观测模型和代价方程$c(x,\hat{x})$.

为了一劳永逸，我们希望估计的结果不是一个孤零零的数字（这样的决策推断被称为*hard decision*），而是一个分布（这样的决策推断被称为*soft decision*）.

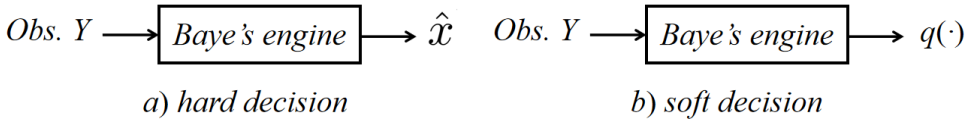

Fig. 4. *hard decision*和*soft decision*的图示.其中，$q(\cdot)$是$\hat{x}(y)$的分布列.

3.1 二次代价与对数代价

有一个与代价函数相类似但取值截然不同的函数，我们用a来代替$\hat{x}$:

$$q(a)=\begin{cases}1,x=a\\0,otherwise\end{cases} \tag{17}$$

这是指示函数，把它记为$\mathbb{I}(a)$吧。

我们将代价函数变一下，引入代价准则（*cost criterion*）：

$$c(x,q(\cdot)) \tag{18}$$

上式用分布$q(\cdot)$代替了$\hat{x}$.

有二次代价准则：

$$c(x,q(\cdot))=A\sum(q(a)-\mathbb{I}(a))^2+B(x) \tag{19}$$

上面的$\mathbb{I}(a)$是公式(17). 其中A=0，B是任意的.

有对数代价准则：

$$c(x,q(\cdot))=-A\log q(x)+B(x) \tag{20}$$

上面的$q(x)$是x的分布. 若令A=1、$B(x)$=0，则$c(x,q)=-\log q(x)$. 在没有观测数据、$q(x)$先验分布的情况下，有$q_x(\cdot)=p_X(\cdot)$.

3.2 先验代价与后验代价

$$\mathbf{E}[c(X,q)]=\mathbf{E}[c(X,p_X(\cdot))]=-\mathbf{E}[\log p_X(x)]=-\sum_a p_X(a)\log p_X(a)=H(X) \tag{21}$$

我们看见老朋友了，香农熵，又见面了. 这一部分由于没有观测数据，被称为先验代价.

若有了观测数据（有$Y=y$的观测数据的$q(\cdot)=p_{X|Y}(\cdot|y)$）时有

$$\mathbf{E}[c(X,q)|Y=y]=\mathbf{E}[c(X,p_{X|Y}(\cdot|y)|Y=y]=-\sum_a p_{X|Y}(a|y)\log p_{X|Y}(a|y)=H(X|Y=y) \tag{22}$$

其中$H(X|Y=y)$是香农熵的中间形式.

$$\mathbf{E}[c(X,p_{X|Y})]=\mathbf{E}_{p_{Y(\cdot)}}[\mathbf{E}[c(x,p_{X|Y})|Y=y]]=H(X|Y) \tag{23}$$

最后我们发现得到了条件熵. 上式称为后验代价.

我们知道**观测数据可以消除不确定性.** 由上可知，两个代价的差值为

$$\Delta\mathbf{E}[c(x,q)] = H(X) - H(X|Y) = I(X;Y) \tag{24}$$

于是又和互信息量扯上关系了。

此外，有$0 \leq H(X|Y) \leq H(X)$；MI = cost reduction.

3.3 不完全推理与信息分歧（*Imperfect Inference and Information Diverqence*）

we know true belief $p_{X|Y} = p_x(\cdot)$; a dist $q \to p_{X|Y}\, or\, p_X(\cdot)$.

我们考虑没有$p_{X|Y} = p_X(\cdot)$的数据的情况；这可以用于衡量近似损失（*approximate loss*）.

$$c(x,q) = \log\frac{p_X(x)}{q_X(x)} \tag{25}$$

$$\mathbf{E}[c(x,q)] = -\mathbf{E}p_X[\log q(x)] + \mathbf{E}[\log p_X(x)] = \sum p_X(x)\log\frac{p_X(x)}{q_X(x)} = D(p_X||q_X) \tag{26}$$

我们惊喜地发现结果是KL散度.需要额外的比特来表示X（additional bits is required to represent X）；Q相对于P的相对熵（纠缠度）（relative entropy of Q with respect to P ）.

让我们考虑有观测数据$Y = y$的情况：

$$\mathbf{D}(p_{X|Y}(\cdot|y)||q_{X|Y}(\cdot|y)) \overset{x=a}{=\!=\!=} \sum_a p_{X|Y}(a|y)\log\frac{p_{X|Y}(a|y)}{q_{X|Y}(a|y)} \tag{27}$$

$$\Delta\mathbf{E}_{p_Y(\cdot)}[c(x,q)] = \mathbf{E}[D(p_{X|Y}(\cdot|y)||q_{X|Y} \times p_y)] \tag{28}$$

也就是说，当我们有观测值时，直接对观测值即数据进行分析；当我们没有观测值时，需要观察数据x的分布，从而进行一个最佳决策.

至此，本章内容就告一段落，接下来我们进入参数估计的学习.

Parameter Estimation
参数估计

1 参数估计和模型类

例1 抛掷一枚不均匀（*biased*）的硬币. 假设各个面出现的概率如下：

$$p_X(x)=\begin{cases} q, x=H \\ 1-q, x=T \end{cases}$$

由于硬币的均匀程度未知，q也未知. 这里有$0\le q\le 1$. q的不同值与不同的可能的分布相对应.通过对参数q的估计，我们可以得到不同的模型（*model*）.

在估计中我们有多个模型可供于推测数据. 如线性（一次）模型、二次模型、对数模型等.

模型类（model class）是模型的一个集合，或者是“参数分布的集合”. 采用集合分布的形式，通过参数θ 进行索引.

第一次模型转化
基于数据，把估计转化为参数模型

参数空间的一般写法为：

$$\{p_X(\cdot;\theta):\theta\in\Theta\} \tag{1}$$

如例1的参数空间可以写为：

$$\{p_X(\cdot;q):q\in[0,1]\}$$

样本空间和参数空间的异同点

异点:
1. 概念不同：样本空间是随机试验的所有可能结果的集合,参数空间是模型的参数取值范围.
2. 性质不同：样本空间是确定的,参数空间需要估计求取.

同点:
1. 都是集合：样本空间和参数空间都是包含元素的集合,元素可以是实数、向量等.
2. 都与模型或随机试验相关:样本空间和参数空间都是与概率模型或随机试验密切相关的空间.
3. 影响分布或密度：样本空间和参数空间会影响相关随机变量的分布或密度. 样本空间影响观察数据的分布,参数空间影响假设模型下的分布.
4. 都需要确定：无论是模型假设还是参数估计,都需要确定样本空间和参数空间. 只有当空间确定,模型和参数才具有意义.

综上,样本空间和参数空间虽然在概念、性质和表示上有所不同,但本质上都是与概率模型和随机试验相关的集合,并且都需要确定与估计. 样本空间影响观察数据的统计变化,参数空间影响理论模型下的统计变化,两者密切相关.

我们需要注意四点：
1.上述模型的θ可以是离散的或连续的.
2.离散的X有$x\in\mathcal{X}$.
3.参数θ起到索引模型类别的作用，不索引随机变量.
4.选择合适的模型类别就是基于数据选择合适的参数θ.

一个参数θ对应一个模型，因此模型估计就是选择最好的参数θ以选择最好的模型.

现在让我们回到抛硬币的问题. 当我们估计q的偏差（bias）的时候，当n的输出已知，我们有$\hat{q}(x_1,...,x_n)=\frac{N_H}{n}$.其中$N_H$是正面出现的次数；$\frac{N_H}{n}$是正面出现的频率. 其中，$\hat{q}(\cdot)$是**估计算子**（*estimator*）. 算子提供了一种估计，θ的估计值$\hat{\theta}$是基于每一个x的观测值的.

当真实的参数（*true parameter*）为θ_0时，我们可以通过引入代价函数$c(\theta,\theta_0)$来衡量参数估计的质量. 同样地，我们追求代价函数的最小值：$\arg\min_{\theta\in\Theta},c(\theta,0)=0.$

除了代价函数，我们还有风险函数：

$$\varphi(\hat{\theta},\theta_0)=\mathbf{E}[c(\theta_0,\hat{\theta}(x);\theta_0)]=\sum c(\theta_0,\hat{\theta}(x)p_x(x;\theta_0)) \tag{2}$$

当风险函数最小值时，$\hat{\theta}=\theta_0$. 我们希望风险函数最小，就是要寻找一个特定的θ_0. 在$Y\to X_n$的推断过程里，特定的X_n由特定的θ_0决定.

同样地，我们希望得到一个分布尽可能地接近所有候选模型. 接近指的是散度意义上的接近，距离就是KL散度.

第二次模型转化

选参数，尽可能地接近所有分布

例2 设$X_n\in\{0,1\}\sim B(\theta),\theta\in\{0,1\}$，求$p_X(\cdot;0)$ 和$p_X(\cdot;1)$之间的K-L散度.

由题意知，X的分布为

$$p_x(X_n,\theta)=\begin{cases}\theta,x=1\\1-\theta,x=0\end{cases}$$

因此，有

$$p_X(x;\theta)=\theta^x(1-\theta)^{1-x}$$

代入1和0，可知

$$p_X(0;0)=0^0(1-0)^{1-0}=1$$
$$p_X(1;0)=0^1(1-0)^{1-1}=0$$
$$p_X(0;1)=1^0(1-1)^{1-0}=0$$
$$p_X(1;1)=1^1(1-1)^{1-1}=1$$

代入KL散度的公式，知

$$D(p_X(\cdot;0)||p_X(\cdot;1))=\sum_{x\in\{0,1\}}p_X(x;0)log\frac{p_X(x;0)}{p_X(x;1)}=p_X(0;0)log\frac{p_X(0;0)}{p_X(0;1)}+p_X(1;0)log\frac{p_X(1;0)}{p_X(1:1)}=1\times log\frac{1}{0}+0\times log\frac{0}{1}$$

即

$$D(p_X(x;0)||p_X(x;1))=log\frac{1}{0}$$

在这里我们运用极限的思想. 我们可以把$\frac{1}{0}$看做一个很大的数，如图

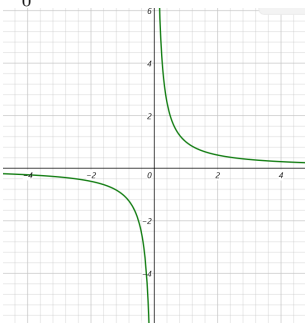

Fig. 1. 图

在这里我们默认log的底数为2（或者e）. 因此我们可以得出

$$D(p_X(\cdot;0)||p_X(\cdot;1))=log\frac{1}{0}=+\infty$$

只有让距离最小，我们才能得到最小的风险函数，才能得到最好的参数估计.

我们引入“大混合分布”（big mixture distribution）. 以例2为例，$q_X=\frac{1}{2}p_X(x;0)+\frac{1}{2}p_X(x;1)$，有$D(p_X(\cdot;0)||q_X)=D(p_X(\cdot;1)||q_X)=1\,bit$. 也就是说这在潜在分布中与两个模型都是接近的.

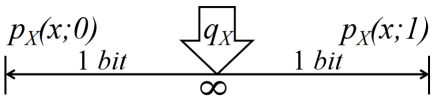

Fig. 2. 图

由此可见，选参数的目的是让参数尽可能地都接近于所有分布. 最好的参数是θ_0，我们需要让θ 尽可能地接近θ_0.

2 常见的几种参数估计

2.1 MMPE（最小均方误差）参数估计

开幕雷击，上来就放一个式子：

$$\varphi_{\mathrm{MSE}}=\mathbf{E}[(\hat{\theta}(x)-\theta_0)^2;\theta_0]=\sum_x(\hat{\theta}(x)-\theta_0)^2p_X(x;\theta_0) \tag{3}$$

上式的意义就是在θ_0条件下的期望. 如果将θ_0换成θ_1，这个式子就不适用了.

单纯的MMSE估计得到的是一个估计值，是$hard-descrion$；为了简化运算，我们将上式分解，转化为几个式子的和. 我们既然要求它的最小值，那么只需要求组成它的几个式子的分别的最小值就可以了.

第三次模型转化

将风险函数转化为几个代数式之和

在转化之前，我们引入下面两组概念：

偏差（$bias$）

这里的偏差指的是经过多次估计后偏差的均值. 简写为$b[\ ]$.

$$u(\hat{\theta},\theta_0)=E[\hat{\theta}(x);\theta_0]$$

$$b[\hat{\theta},\theta_0]=u[\hat{\theta},\theta_0]-\theta_0$$

算子估计的方差

简写为$var[]$.

$$var[\hat{\theta};\theta_0]=\mathbf{E}[\hat{\theta}(x)-u(\theta,\theta_0)^2;\theta_0]=\sum_x(\hat{\theta}(x)-u(\hat{\theta},\theta_0)^2\times p_X(x,\theta_0)$$

其中分号后面的θ_0是“基于θ_0条件下”的意思.

我们对式子(3)进行展开，得到

$$\mathbf{E}[(\hat{\theta}(x)-\mathbf{E}[\hat{\theta}(x);\theta_0]+\mathbf{E}[\hat{\theta}(x);\theta_0]-\theta_0)^2;\theta_0]=\mathbf{E}[(\hat{\theta}(x)-\mathbf{E}[\hat{\theta}(x);\theta_0])^2;\theta_0]+\mathbf{E}[(\mathbf{E}[\hat{\theta}(x);\theta_0]-\theta_0)^2;\theta_0]+2AB \tag{4}$$

其中，我们设$\hat{\theta}(x)-E[\hat{\theta}(x);\theta_0]$为$A$，$E[\hat{\theta}(x);\theta_0]-\theta_0)^2$为$B$.

于是，我们可以得到：

$$\varphi_{\mathrm{MSE}}=\mathbf{E}[(\hat{\theta}(x)-\theta_0)^2;\theta_0]=var[\theta;\theta_0]+B^2[\theta,\theta_0]+0=[var]_{min}+B^2{}_{min} \tag{5}$$

（$2AB$=0，证明过程略）.

对于任何的θ_0，假如我们让$b>0$，就得到了minimum variance unbiased estimator（最小无偏差方差算子，MVU算子）.

让我们回到一开始的那个扔硬币的例子. 扔到正面（H）的次数N_H服从二项分布：
$P(X=N_H)=C_n^{N_H}p^{N_H}(1-p)^{1-N_H}$，$\mathbf{E}[X]=np$，$\mathbf{D}[X]=np(1-p)$，其中$p$是扔到正面的概率. 我们有

$$b[\hat{p},p_0]=\mathbf{E}[\hat{p};p_0]-p_0=\frac{1}{n}\mathbf{E}[N_H,p_0]-p_0=\frac{1}{n}np_0-p_0=0$$

$$var[\hat{p},p]=var[\frac{N_H}{n};p_0]=\frac{1}{n^2}var[N_H;p_0]=\frac{1}{n}p_0(1-p_0)$$

$$\varphi_{\mathrm{MSE}}(\hat{p},p_0)=\frac{1}{n}p_0(1-p_0) \tag{6}$$

其中$p(\cdot)$是MVU算子.

2.2 ML（最大似然参数）估计

估计方法的两种分类标准

1.有无观测数据（$Obs.$）
2.是$hard-decision$还是$soft-desicion$

我们之前讲过似然函数$p_X(x;\cdot)$，其中x是观测数据. 当知道了x和θ之后，我们就可以求$p_X(x;\theta)$（概率），$p_X(x;\theta)=L(\cdot;x)$.

注意：似然函数在参数估计中起到重要作用甚至中心作用. $p_X(x;\theta)$是源于模型$p_X(x;\cdot)$的简化的x的观测值的概率.

对于**ML（最大似然参数）估计**而言，进行估计的前提是有观测数据.

当有观测数据（$data\ x$）时

$$\hat{\theta}_{\mathrm{ML}}(x)=arg\max_{\theta\in\Theta}L(\theta;x)=\arg\max_{\theta\in\Theta}p_X(x;\theta) \tag{7}$$

即在最大似然参数估计中，我们要选择一个模型来让观测数据的概率尽可能地大. 换句话说，就是“最可能出现”.

我们以扔一个*不均匀*的骰子的例子来说明之：考察扔出6点的概率：

模型1（θ_1）：

$$p_X(x;\theta_1)=\begin{cases}1/6,x=1\\ \cdots\\ 1/6,x=6\end{cases}$$

模型2（θ_2）：

$$p_X(x;\theta_2)=\begin{cases}1/21,x=1\\ \cdots\\ 4/5,x=6\end{cases}$$

上述两个模型中，模型2更胜一筹，因此我们可以说$\hat{\theta}(x)=\theta_2$.

我们对最大似然函数取对数：

$$\ell(\theta;x)=\log p_X(x;\theta) \tag{8}$$

$$\theta_{ML}=\arg\max_{\theta\in\Theta}\ell(\theta;x) \tag{9}$$

我们始终需要注意，参数估计是基于已经观测到的数据. 如果我们重复多次扔硬币的估计（$x_1^n=(x_1,x_2,...,x_n)$），就有：

$$\frac{1}{n}\ell(q;x_1^n)=\frac{1}{n}\log p_{X_1^n}(x_1^n;q)=\frac{1}{n}\prod_{i=1}^{n}\log p_X(x_i,q)=\frac{1}{n}\sum_{i=1}^{n}\log p_X(x_i;q)$$

$$=\frac{1}{n}\sum_{i=1}^{n}\log q^{I(x_i=H)}(1-q)^{I(x_i=T)}=\hat{p}(H;x_1^n)\log q+\hat{p}(T;x_1^n)\log(1-q)=\frac{N_H}{n}\log q+\frac{N_T}{n}\log(1-q) \tag{10}$$

上面的$I(x)$叫做**指示函数**. 它的含义是：当输入为True的时候，输出为1；输入为False的时候，输出为0. 也就是说满足条件的时候是1，反之为0.

回到$\hat{q}_{ML}$. 有$\hat{q}_{ML}\to\frac{\partial}{\partial q}\cdot\frac{1}{n}\ell(q;x_1^n)=\frac{1}{q}\cdot\frac{N_H}{n}-\frac{1}{1-q}(1-\frac{N_H}{n})=0$，$\hat{q}_{ML}=\frac{N_H}{n}$.

2.3 经验分布（Empirical Dist.）

有一种分布叫做**经验分布（Empirical Dist.）**. 所谓经验分布，就是对于X而言，其分布可以通过观测值来获得，与观测值（数据）出现的总次数，目标次数有关。

ML参数估计与经验分布的关系——经验分布在ML参数估计中起到关键作用；不适用于两点分布（如抛硬币），但适用于一切独立同分布.

接下来我们看几个式子：

$$p_{x_i}(x_1^n) = \prod_{i=1}^{n} p_X(x_i, \theta) \tag{11}$$

由于各个样本点是独立同分布，因此样本顺序不影响其出现的概率；$\mathcal{X}$中每个符号出现的相对频率.

$$\hat{p}_x(a; x_1^n) = \frac{1}{n}\sum_{i=1}^{n} I(x_i = a)(a \in \mathcal{X}) \tag{12}$$

这里的x_1^n是观测数据；此式子的计算必须在观测数据已知的情况下进行；计算结果（p的估计值）由观测数据决定；此信息是通过经验分布获得的.

$$\begin{aligned}
p_x(x_i;\theta) &= \prod_{a\in\mathcal{X}} p_X(a;\theta)^{I(x=a)} \\
&= \sum_{i=1}^{n}\sum_{a\in\mathcal{X}} I(x_i = a)\log p_X(a;\theta) \\
&= n\sum_{a\in\mathcal{X}} \log p_X(a;\theta)\frac{1}{n}\sum_{i=1}^{n} I(x_i = a) \\
&= n\sum_{a\in\mathcal{X}} \hat{p}_X(a;x_1^n)\log p_X(a;\theta) \\
&= n\sum_{a\in\mathcal{X}} [\hat{p}_X(a;x_1^n)\log \frac{\hat{p}_X(a;x_1^n)}{\frac{\hat{p}_X(a;x_1^n)}{p_X(a;\theta)}}] \\
&= n\sum_{a\in\mathcal{X}} [\hat{p}_X(a;x_1^n)\log \hat{p}_X(a;x_1^n) - \hat{p}_X(a;x_1^n)\log \frac{\hat{p}_X(a;x_1^n)}{p_X(a;\theta)}] \\
&= -n[\hat{H}(X) + D(\hat{p}_X(a;x_1^n)||p_X(\cdot;\theta))]
\end{aligned} \tag{13}$$

其中$\hat{H}(X)$是经验熵；$D(\hat{p}_X(a;x_1^n)||p_X(\cdot;\theta))$唯一取决于$\theta$.

于是我们有

$$\theta_{\mathrm{ML}}(x_1^n) = \arg\max_{\theta\in\Theta} p_{x_1^n}(x_1^n;\theta) = \arg\min_{\theta\in\Theta} D(\hat{p}_x(\cdot;x_1^n)||p_x(\cdot;\theta)) \tag{14}$$

上式告诉我们，**我们欲在参数里找一个最优的参数，使其对应的分布于观测数据生成的经验分布之间是最匹配的（散度最小）.**而通过经验分布，我们也知道了一切我们需要在数据中需要知道的的信息。

2.4 经验熵

在前面的讨论里，我们引入了经验熵。经验熵的表达式为：

$$\hat{H}(X) = -\sum_{a\in\mathcal{X}} \hat{p}_X(a;x_1^n)\log \hat{p}_X(a;x_1^n) \tag{15}$$

我们需要选择参数使得$p_X(\cdot;\theta)$尽可能地接近$p_X(\cdot;x_1^n)$. 存在分歧（$Divergence$）.

2.5 经验分布匹配的复杂度

提及匹配（$match$），就必然有一个选择的问题. 这个选择，也有复杂度. 不但样本空间有，参数空间也有.

设Θ为参数空间（$\theta \in \Theta$），$|\Theta|$（参数空间里参数的数量，即模型数量）记为$\sharp$. 对于最大似然参数估计（ML估计），其复杂度为

$$O(|\Theta|) = O(\sharp) \tag{16}$$

即有多少个参数就要匹配多少次（线性方程关系）.

好的算子比差的算子需要的观测数据更少；复杂度取决于算子.

独立同分布可以显著降低计算的复杂度.

数据在进行模型的选择过程中发挥着关键的作用，通过对参数空间、参数估计、经验分布、模型复杂度的影响来影响到最终最佳模型的选择。

其实，以下三个词语是一个意思：

Model Selection / Parameter Estimation / Model Order Selection

我们把大数据集分为两部分－－一部分用于训练模型，一部分用于测试模型. 在训练模型的时候，可能发生欠拟合（$underfitting$）和过拟合（$overfitting$）的问题.

欠拟合就是模型没有很好地捕捉到数据特征，不能够很好地拟合数据；过拟合就是模型把数据学习的太彻底，学的东西中有干扰数据，存在用个性代替共性的问题，这样就会导致在后期测试的时候不能够很好地识别数据、得出正确的结果。

Fig. 3. 欠拟合和过拟合示意图

2.6 奥卡姆剃刀（Occam's Razor）

公元14 世纪，来自奥卡姆的威廉(William of Ockham)认为应该只承认确实存在的东西，认为那些空洞无物的普遍性要领都是无用的累赘，应当被无情地“剃除”。概括起来就是“如无必要，勿增实体”. 奥卡姆剃刀又被称为又称简单性原理，是科学界常用的一个准则，具体表述为，**如果多个理论同时都能解释某一现象，那么我们优先取利用假设最少的理论，这个理论被认为是最好的.**

在推断与估计上，如果有多个模型可以拟合数据，我们要**选择参数最少的模型、假设最少的模型**，或者给出可遵循的法则的模型（但是奥卡姆剃刀模型不会告诉我们如何实现法则）.

2.7 最小描述长度（Mininum Description Length）

我们需要选择一个提供最紧凑的数据集的模型；“考虑到模型（taking into account the model）”.

三个式子：观测值、概率值和似然函数

$p_X(\cdot;\theta)$是观测值；
$p_X(x;\theta)$是概率值（参数θ条件下的x的概率）
$p_X(x;\cdot)$（L）是似然函数. **所谓“似然”，就是“可能性”**.

Exponential Family
指数簇

在参数估计部分，我们知道了模型类的写法是$\{p_X(\cdot;\theta):\theta\in\Theta\}$，实际上就是参数的概率分布.

为了阅读方便，在下面的讨论中，我们用exp[x]代替e^x.
本文使用了大量缩写，一般是英文首字母连写.

1 指数簇

指数簇（*Exponential Family*，EF）在推断中起到关键作用；很多众所周知的参数概率分布实际上就是指数簇.

2 单一参数的指数簇

$$\{p_X(\cdot;x);x\in\mathcal{X}\} \tag{1}$$

上式在字母表y上是一个单一参数的指数簇. 其中$p_X(\cdot;x)$是数据，$x\in\mathcal{X}$是参数.
在接下来的讨论中，y是数据，x是参数.

$$p_Y(y;x)=\exp[\lambda(x)+t(y)-\alpha(x)+\beta(y)] \tag{2}$$

上式右边指数部分共有四项：$\lambda(x)$是自然参数（natural parameter）；$t(y)$是自然统计（natural statistic）；$\alpha(x)$是基于对数的函数（log-based function）；$\beta(y)$是归一化每个分布（normalize each distribution）. 因此

$$p_Y(y;x)=\frac{1}{z(x)}\exp[\lambda(x)+t(y)+\beta(y)] \tag{3}$$

其中$z(x)=e^{a(x)}=\sum_y\exp[\lambda(x)+t(y)+\beta(y)]$.

$$y\sim\varepsilon\{\mathcal{X},y,\lambda(\cdot),t(\cdot),\beta(\cdot)\}\sim\{\mathcal{X},y,\lambda(\cdot),t(\cdot)-c_1,\beta(\cdot)-c_2\} \tag{4}$$

上式中c_1、c_2是常数.

$$y\sim\varepsilon\{\mathcal{X},y,\lambda(\cdot),t(\cdot),\beta(\cdot)\}\sim\{\mathcal{X},y,\lambda(\cdot),t(\cdot)-c_1,\beta(\cdot)-c_2\} \tag{5}$$

由上可知，有$\ln p_Y(y;x)=\lambda(x)t(y)-(\alpha(x)+c_1\lambda(x)+c_2)+\beta(y)$. 其中$\alpha(x)+c_1\lambda(x)+c_2$是$\alpha(x)$.
x是一个真实的数（标量），y可以是标量或矢量（向量）.

例1（伯努利分布） 伯努利分布如下所示：

$$p_Y(y;x)=\begin{cases}x,y=1\\1-x,y=0\end{cases}$$

也就是$p_Y(y;x)=x^y(1-x)^{1-y}$.
两边取对数，得

$$\ln(p_Y(y;x))=\ln x^y+ln(1-x)^{1-y}=y\ln x+(1-y)ln(1-x)=y\ln\frac{x}{1-x}+\ln(1-x)$$

$$p_Y(y;x)=\exp[y\cdot\ln\frac{x}{1-x}+\ln(1-x)]$$

其中$\ln\frac{x}{1-x}$是$\lambda(x)$；y是$t(y)$；$\ln(1-x)$是$\alpha(x)$；$\beta(y)$=0.

例2（指数分布） 指数分布的表达式为：$p_y(y;x)=\frac{1}{x}e^{-\frac{y}{x}}(y\geq 0)$. 两边取对数，得

$$\ln(p_Y(y;x))=-\frac{y}{x}+\ln\frac{1}{x}$$

$$p_Y(y;x)=\exp[-\frac{1}{x}\cdot y-\ln x]$$

其中$-\frac{1}{x}$是$\lambda(x)$；y是$t(y)$；$-\ln x$是$\alpha(x)$；$\beta(y)$=0.

例3（高斯分布） 高斯分布的表达式为：

$$p_Y(y;x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(y-x)^2}$$

两边取对数，得

$$\ln(p_Y(y;x))=-\frac{1}{2}(y-x)^2-\frac{1}{2}ln(2\pi)=p_y(y;x)=xy-[\frac{1}{2}x^2+\frac{1}{2}\ln(2\pi)]-\frac{1}{2}y^2$$

其中x是$\lambda(x)$；y是$t(y)$；$\frac{1}{2}x^2+\frac{1}{2}\ln(2\pi)$是$\alpha(x)$；$-\frac{1}{2}y^2$是$\beta(y)$.

例4（几何分布） 高斯分布的表达式为：$p_Y(y;x)=(1-x)x^y(y=0,1,2,...)$
两边取对数，得

$$\ln(p_Y(y;x))=(\ln x)y+ln(1-x)$$

其中$lnx)$是$\lambda(x)$；y是$t(y)$；$\ln(1-x)$是$\alpha(x)$；$\beta(y)$=0.
例3和例4省略了最后一步脱掉对数的过程.

对于离散的y，它的维度（*dimension*，*dim.*）是$|y|$. 我们在立体图中表示y：$y=\{a,b,c\}$，$\mathcal{D}=\{p_Y(a),p_Y(b),p_Y(c)\}$. 我们设置a、b、c的横、纵、竖坐标为1，由abc三个点连接起来的斜面为“概率纯形”（*Probability Simplex*，*PS*），EF在PS上.

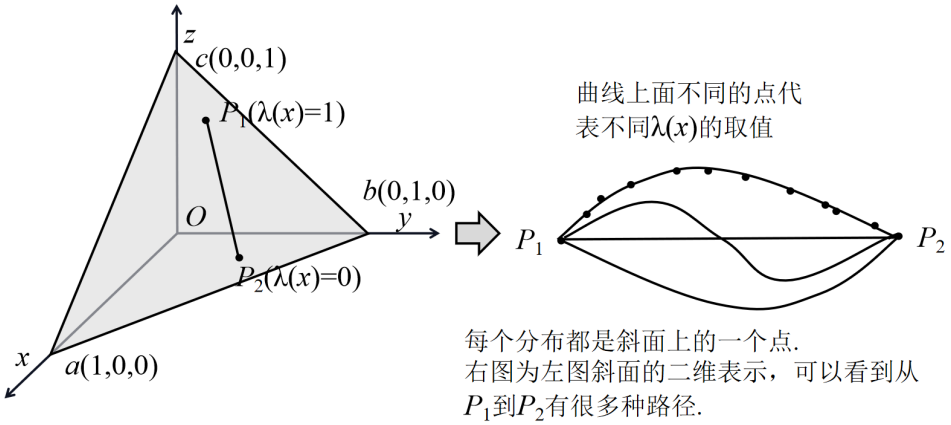

Fig. 1. 关于“概率纯形”的图示

$p_1(y)$、$p_2(y)$是PMF（$0<x<1$）. 我们有

$$p_Y(y;x)=\frac{(P_1(y))^{\lambda(x)}(P_2(y))^{1-\lambda(x)}}{z(\lambda(x))}$$

两边取对数，有

$$\ln(p_Y(y;x))=\lambda(x)\ln\frac{p_1(y)}{p_2(y)}-\ln(z(\lambda(x)))+\ln p_2(y)$$

其中$\lambda(x)$是$\lambda(x)$；$\ln\frac{p_1(y)}{p_2(y)}$是$t(y)$；$\ln(z(\lambda(x)))$是$\alpha(x)$；$\ln p_2(y)$是$\beta(y)$.

其实，指数簇是自然存在的，只是它的“形式”恰恰是指数簇罢了.

3 充分统计量和无损预处理（*Sufficient Statistic and lossless pre-processing*）

我们用数据的特征来表示数据——或者说，特征本身就是数据的一个描述. 对于推断的观测数据（*inference obs. data*），有许多大的维度（*large dimension*）. 那么，我们如何解决原始数据维度过多的弊端？答案是对数据进行预处理（*pre-processing*）；此外，我们需要获取更紧凑的数据集（*more compant set of data*）.

如果$p_{Y|t}(\cdot|\cdot;x)$不是所有$x\in\mathcal{X}$的x的函数，则统计量$\overrightarrow{t}(\cdot)$相对于分布$p_Y(\cdot;x)$是充分统计量. 换句话说，基于$y$的推断实际上就是基于$t(y)$的推断.

$$p_Y(y;x)\propto p_t(t(y);x)$$

但很不幸，这个对于“充分”的定义没有给我们实现“充分”的方法.

我们希望无损地预处理数据，以得到一个可以代替源数据集的紧凑的数据集（*compant set of data*）.

为了找到实现“充分统计量”的方法，我们引入Neyman因子分解（*Factorization*）：
一个统计量$t(\cdot)$若可以分解为两个式子$a(\cdot;\cdot)$和$b(\cdot;\cdot)$的乘积，即

$$p_Y(\overrightarrow{y},\overrightarrow{x})=a(t(\overrightarrow{y},\overrightarrow{x}))b(\overrightarrow{y})$$

其中$x\in\mathcal{X}$，$y\in Y$，二者分别是参数空间和样本空间.

例1 令$y=[y_1,y_2]^T$是一个组分之间相互独立的二维向量. 让二者是高斯分布，均值是x，方差为1：

$$p_Y(y;x)=\frac{1}{\sqrt{2\pi}}\exp[-\frac{(y_1-x)^2+(y_2-x)^2}{2}]=\frac{1}{\sqrt{2\pi}}\exp[-\frac{y_1^2+y_2^2}{2}]\exp[x(y_1+y_2)-x^2]$$

$b(y)=\exp[-\frac{y_1^2+y_2^2}{2}]$；$a(t(y),x)=\exp[x(y_1+y_2)-x^2]$，所以$t(y)=y_1+y_2$，是“充分统计量”.

上个实例告诉我们，（凡是均值为x、方差为1的）自然统计量$t(\cdot)$，若可以分解成“观测值乘以参数”的形式，就可以认为它是充分估计量.

例2 设$p_Y(\cdot|x)$是指数簇的一部分，$t(\cdot)$是充分统计量. 由于其满足有Neyman因子分解的要求，故$a(t,x)=\exp[\lambda(x)t(y)-\alpha(x)]$，$b(y)=\exp[\beta(y)]$.

4 指数簇与参数估计

在上一张我们对比了“方差”（*var.*）和“偏差”（*bias*）. 二者都是表示“偏离程度”的估计量.

我们在进行最小均方误差估计的时候提到过：

$$b(\hat{\theta};\theta_0)=u(\hat{\theta};\theta_0)-\theta_0$$

当MVU算子满足$b=0$时，我们称之为“无偏”（*unbiased*）算子. 此时$var[\hat{\theta};\theta_0]\leq var[\hat{\theta}_1;\theta_0]$.

我们用图示来说明算子的质量：

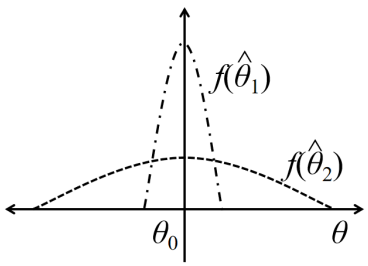

Fig. 2. 通过偏离程度衡量算子质量的图示

上图中θ_0是样本中心点，由图可知$\hat{\theta}_1$在横轴（θ）上的偏离更小，方差更小，因此$\hat{\theta}_1$相对于$\hat{\theta}_2$来说更有效. 图线越高窄，参数估计能力越强，信息量越大.

说一点题外话：在人文社科的研究中，经常牵扯到“信度”和“效度”的问题. 简言之，“信度”可以用方差来衡量；“效度”可以用偏差衡量. 我们要追求方差和偏差都较小.

接下来我们讨论“有效率的”（*efficient*）和指数簇之间的关系. 首先，“有效率的估计”（*efficient estimation*）和“无偏估计”（*unbiased estimation*）是一个含义，我们用$\hat{x}_{\text{eff}}(\cdot)$来表示；$\hat{x}_{\text{eff}}(\cdot)$很接近EF.

若模型$p_Y(\cdot;x)$是EF的一部分，$\hat{x}_{eff}$存在于非随机参数的灭绝（*exists for extimating a nonrandom parameter*）.

$$x_{\text{eff}} = c \cdot t(y)$$

它不依赖于x，一定是一个“有效的”（$efficient$）算子. 有效算子更加“接近”EF/真实值.

5 费希尔信息（*Fisher Information*，FI）

FI测量观测数据的向量关于参数x的信息量（*FI measures how informative a vector of a observe data is about a parameter* x）（$\vec{x}$），即其衡量的是一个数据的矢量能对x 提供多少信息. 费希尔信息的引入可以估计MLE方程的方差，可以收集更多的数据，因此可以获得更多的有效信息；其是在参数空间的KL散度的本地化版本，且在真实参数附近，可以衡量一个分布进行估算时的最小误差以达到减小误差的目的.

我们之前引入了似然函数：$L(x;y) = p_Y(y;x)$；对其取对数，有$\ell(x;y) = \log p_Y(y;x)$.

在上面的基础上，我们引入$scorefunction$：

$$S(x;y) = \frac{\partial}{\partial x} \log p_Y(y;x) = \frac{\partial}{\partial x} \ell(x;y) \tag{6}$$

FI被降级为$J(x)$：

$$J(x) = var[S] = \mathbf{E}[(\frac{\partial}{\partial x} \log p_Y(y;x))^2] \tag{7}$$

当且仅当$p_Y(y;x)$满足正则条件（*regularity condition*），$\mathbf{E}$[S]=0即$\mathbf{E}[\frac{\partial}{\partial x} \log p_Y(y;x)] = 0$，FI可以被压缩（*be companted*），否则不能压缩.

在这里讨论一点FI参数空间散度问题. 由于篇幅限制，不再详细展开. 我们用一个例子来说明之：

例 有高斯分布$Y \sim N(\mu, \sigma^2)$，

$$p_Y(y;x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp[-\frac{(y-\mu)^2}{2\sigma^2}]$$

$$\ell(x,y) = \ln p_Y(y;x) = \frac{1}{2} \ln(2\pi\sigma^2) - \frac{(y-\mu)^2}{2\sigma^2}$$

现在，$x = \mu$：

$$J(\mu) = \mathbf{E}[(\frac{\partial \ell}{\partial x})^2] = -\mathbf{E}[\frac{\partial^2 \ell}{\partial x}] = -\mathbf{E}[\frac{\partial S}{\partial x}] = \frac{1}{\sigma^2}$$

如果用Fig.2所示的图线衡量，这个图线是“高”“窄”的. 这意味着参数估计能力强，信息量大.

若Y_1、Y_2两个分布的互相独立的，而且有同一个参数：

$$p_{Y_1 Y_2}(y_1 y_2; x) = p_{Y_1}(y_1; x) p_{Y_2}(y_2; x)$$

$$J_{Y_1 Y_2}(x) = J_{Y_1}(x) + J_{Y_2}(x)$$

对于独立同分布的观测数据，有：

$$J_{Y_1^n}(x) = \sum_{i=1}^{n} J_{Y_i}(x) = n J_{Y_i}$$

每一个参数对应每一种分布. 估计到的参数对应的分布与所对应的真实分布之间难免存在误差.

我们接下来讨论一点FI与香农信息（*Shannon Information*）. 香农把信息分为**自信息**（*self Info.*）和**互信息**（*mutual Info.*）. 其中自信息就是**熵**. FI 则衡量了分布$D(\cdot)$参数估计的最小误差（*the minimum error in estimating a parameter of a distribution* $D(\cdot)$）.

无论是FI还是香农信息，二者都是**从观测数据中找到信息**. 不同的是，**香农信息注重获取数据的潜在规律和特征，FI重在获取有关参数的信息**（模型选择）. FI实际衡量的是**参数估计在散度意义上的误差**.

最后我们再来讨论一点指数簇和FI的结合.

$$J(x)=\lambda'(x)^2 var[t(y)]=\alpha''(x)-\lambda(x)E[t(y)]$$

其中，

$$\alpha''(x)=\frac{d}{dx}\alpha'(x)=\frac{d}{dx}[\lambda'(x)\sum t(y)\exp[\lambda(x)+t(y)+\beta(y)-\alpha(x)]]=\lambda'(x)\mathbf{E}[t(y)]+\lambda'(x)^2 var[t(y)]$$

Information Geometry
信息几何

我们已经知道K-L散度：

$$D(p||q)=\sum p(y)\log\frac{p(y)}{q(y)}>0$$

我们设$p(\cdot)$是观测值y中的一个分布，我们再次看这个“概率纯形”的图示：

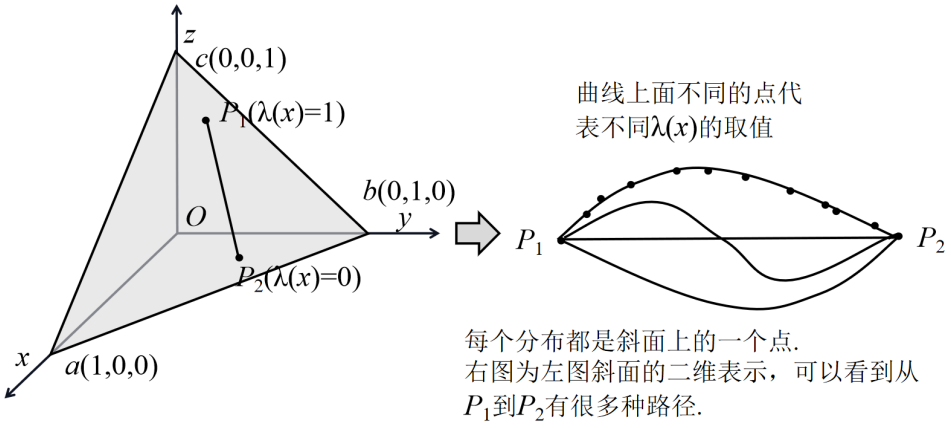

Fig. 1. 关于“概率纯形”的图示

我们知道，上述过程实际上就是一个“降维”的过程。三维降到二维，二维降到一维.

例1 已知$p_1p_2\in\mathcal{P}^y$（观测值y）. 考虑到分布的集合（模型类）：

$$p_Y(y;x)=\frac{p_1(y)^x p_2(y)^{1-x}}{z(x)}\in\mathcal{P}^y,\ x\in[0,1],\ x\in\mathbf{R}$$

$$\mathcal{P}=\{p\in\mathcal{P}^y,\ p(y)=p_Y(y;x)=\frac{p_1(y)^x p_2(y)^{1-x}}{z(x)},\ x\in[0,1]\}$$

例2 已知$p_1p_2\in\mathcal{P}^y$（观测值y）. 考虑到分布的凸集（*convex set*）：

$$p_Y(y;x)=xp_1(y)+(1-x)p_2(y)$$

上式说明了这三者之间是线性关系.

例1是一种几何平均，例2是一种算术平均.

我们都知道大名鼎鼎的**勾股定理**，洋人把这玩意儿叫做毕达哥拉斯定理（*Pythagorean theorem*）. 现在我们把这个东西引入信息几何学上.

设$\mathcal{P}$是$\mathcal{P}^y$的一个子集. q是任一分布. 让q在$\mathcal{P}$中作正交投影，这两个正交投影所对应的分布与q的散度较小. 即：

$$p^*=\arg\min_{p\in\mathcal{P}}D(p||q)\tag{1}$$

其中，p^*被称为“信息投影”（*information projection*）；p^*存在，因为$D(\cdot||\cdot)$ 是非负且连续的；p^*不一定是独特的，取决于$\mathcal{P}$的性质（whom $\mathcal{P}$ convex set）；$D(\cdot||\cdot)$表现为欧几里得距离的平方；$p_\lambda=(1-\lambda)p^*+\lambda p_{\in\mathcal{P}}\ \in\mathcal{P}\ (0\le\lambda\le1)$.

p^*是q在自己中的一个样本点.设q是任意的分布，p^*是信息投影. 考虑$\mathcal{P}$的子集，对于任意的$p\in\mathcal{P}$：

$$D(p||q)\geq D(p||p^*)+D(p^*||q) \tag{2}$$

证明 由于$D(p^*||q)$是$D(p_\lambda||q)$的最小值，因此仅是参数$p^*\to p$. 我们注意到$\lambda\in[0,1]$：

$$\frac{d}{d\lambda}D(p_\lambda||q)=\frac{d}{d\lambda}\sum p_\lambda(y)\log\frac{p_\lambda(y)}{q(y)} \tag{3}$$

$$=\sum_y\frac{dp_\lambda(y)}{d\lambda}\log p_\lambda(y)+p_\lambda(y)\frac{1}{p_Y(y)}\frac{d}{d\lambda}p_Y(y)-\frac{d}{d\lambda}p_\lambda\log q(y) \tag{4}$$

$$=\sum_y(p-p^*)\log\frac{p_\lambda(y)}{q(y)} \tag{5}$$

当λ=0时：

$$\frac{d}{dx}D(p_\lambda||q)|_{\lambda=0}=\sum p(y)\log\frac{p_\lambda^*(y)}{q(y)}-\sum p^*(y)\log\frac{p^*(y)}{q(y)} \tag{6}$$

$$=\sum p(y)\log[\frac{p(y)}{q(y)}-\frac{p^*(y)}{p(y)}]-\sum p^*(y)\log\frac{p^*(y)}{q(y)} \tag{7}$$

$$=D(p||q)-D(p(y)||p^*(y))-D(p^*||q)\ \geq 0 \tag{8}$$

Prior
先验

关于“先验”的内容想必我们在之前的章节里已经了解很多. 在本章，让我们一起来回顾一下“先验”，同时引入一些新的内容.

参数估计的目标

确定未知的参数x.

模型的目标

混合模型：用混合模型表示模型类中的每一个模型，损失最小.

$q_w(y)=\sum_{x\in\mathcal{X}}w(x)p_Y(y;x)$；其中$y$是$q(\cdot)\to p(\cdot)$ 的数据。y是参数；$w(x)$是非负的；$\sum w(x)=1$；$w(x)$是一个先验.

模型选择

尽可能地和各个数据都接近——平均意义上的接近.

例1 设$y_n\in 0,1$服从二项分布$B(x)$，$x\in 0,1$，有

$$D(p_Y(\cdot;0)||p_Y(\cdot;1))=D(p_Y(\cdot;1)||p_Y(\cdot;0))=\infty$$

$$q(y)=\frac{1}{2}p_Y(y;0)+\frac{1}{2}p_Y(y;1)$$

$$D(p_Y(\cdot;0)||q(\cdot))=D(p_Y(\cdot;1)||q(\cdot))=1bit$$

用混合模型表示模型类中的每一个模型，损失最小. 没有一个模型能在推断上比混合模型效果好.

1 信息最少的先验（*Least Informative Prior*，LIP）

首先需要说明：LIP依赖观测数据.

我们引入一个新概念：容量（*capacity*）.

- 当c比较小时，复杂度会降低，有$w(x)=p_X(\cdot)$
- 当c比较大时，包括相应的模型：$q_w(y)=p_Y(y)=\sum p_X(x)p_{Y|X}(y|x)$

$$C=\max_{p_X}\sum_x p_X(x)D(p_{Y|X}(\cdot|x)||p_Y(\cdot))=\max_{p_X}D(p_{XY}(\cdot;\cdot)||p_X(\cdot)p_Y(\cdot))=\max(H(X)-H(X|Y))=\max I(X;Y) \quad (1)$$

我们在“数据的离散”一章里讲到了观测数据可以降低隐随变量的不确定性，但这是在一般情况之下. 对于LIP，我们有

$$p_x^{LIP}(x)=\arg\max_{p_X(\cdot)}I(X;Y) \quad (2)$$

此外，c（容量）是有范围的，与互信息有关. 有$0\le c\le\log|\mathcal{X}|$. 原本$x$是随机变量，但x 确定之后就不再具有随机性.

倘若我们可以从观测数据$Y=y$中得到有关x的信息，则$\max I(X;Y)=\max(H(X)-H(X|Y)=\max H(X)$，这里$H(X|Y)$=0. 相反地，若我们没有从观测数据$Y=y$中得到有关$x$的信息，则$\max I(X;Y)=0$，这里$H(X|Y)=H(X)$.

2 熵值最大的先验（*Maximum Entropy Prior*，MEP）

在经过观测后，有观测数据$Y=y_1,y_2,...,y_n$：

$$C=\max I(X;Y)=H(X)-H(X|y_1,...,y_n)\approx 0 \tag{3}$$

$$p_x^{ME}=\arg\max_{p_X}(\cdot)H(X) \tag{4}$$

在这里我们需要注意：它不与观测的维度（n）有关；它是熵值最大的先验，也是不确定性最大想先验，也是最最无知的先验.

接下来我们了解一下最大熵原理. 均匀分布的时候熵值最大. 对于均分分布q，我们有

$$D(p||q)=D(p||u)=\sum p(y)\log p(y)+\log(y)=\log(y)-H(p) \tag{5}$$

上式说明，熵值越大的时候散度越小.

$$\arg\max H(\cdot;p)=\arg\min D(p||u) \tag{6}$$

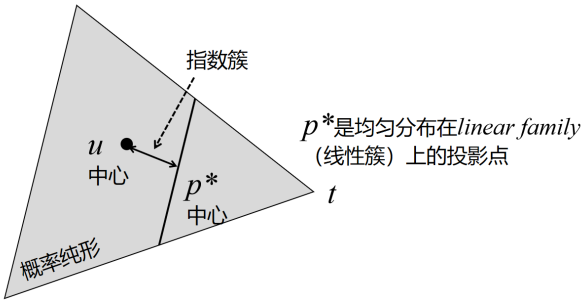

Fig. 1. p*的示意图

此外，我们有

$$\mathcal{L}_t(p)=\{p\in\mathcal{P}:\mathbf{E}[t_1(y)]=t_k,k=1,2,3,....n\}$$

3 共轭先验（*Conjugate Prior*）

下面的内容会基于排列组合的原理：
设$y=[y_1,y_2,...,y_n]^T$是有条件的独立同分布，有

$$p_{y_1^n|X}(y_1,y_2,...,y_n|X)=\prod_{i=1}^{N}p_{Y|X}(y_i|X)$$

其中y_1^n是随机变量的取值；

$$p_{Y_1,...,Y_n|X}(y_{n_1},...,y_{n_N}|X)=p_{Y_{n_1,...,1/n_N}}(y_{n_1},...,y_{n_N}|X)$$

其中$n_1,n_2,...,n_N$是N重排后的结果；随机变量序列（*Asequence*）$Y_1,...,Y_n$是可交换的.
如果对于所有的参数$n_1,...,n_N$，有

$$p_{Y_1,...,Y_n|X}(y_{n_1},...,y_{n_N}|X)=p_{Y_{n_1},...,Y_{n_N}}|X(y_{n_1},...,y_{n_N}|x)$$

任意的独立同分布序列都是可交换的；但并非全部可交换都是独立同分布.

我们聚焦于条件独立同分布模型：

$$p_{Y|X}(y_1,...,y_n|x)=\prod_{i=1}^{N}p_{Y|X}(y_i|x) \tag{7}$$

设Q是某个模型类. 参数$\theta\in\Theta$. $Q=\{q(\cdot;\theta),\theta\in\Theta\}$代表着一个簇. Q是一个共轭先验簇（对于(7)而言）.
如果对于所有的$y\in\mathcal{Y}$，我们有$p_{X|Y}(\cdot|y)\in Q$. 无论何时，$p_X(\cdot)\in Q$，$p_{X|Y}(\cdot|y)\in\theta(y,\theta_0)$.

4 信念修正（*Belief revision*）

当我们收到了$Y=y$后：

$$p_{X|Y}(x|y_1)=\frac{p_{Y_1X}(y_1|x)p_X(x)}{\sum_a p_{Y|X}(y|a)p_X(a)}=T_{y_1}^{(1)}[p_X(\cdot)]$$

在这里，Y和X都成了条件分布；这是第一次更新.

当我们收到了$Y=y$后：

$$p_{X|Y_1Y_2}(x|y_1y_2)=\frac{p_{1/2Y_1\cdot X}(y_2|y_1,x)p_{X|Y}(x|y_1)}{\sum_a p_{Y_2|Y_1;X}(y_2|y_1,a)p_{X|Y_1}(a|y_1)}=T_{y_1|y_2}^{(2|1)}[p_{X|Y_1}]$$

若Y_1和Y_2独立$|X$，则无需计算上述式子.

当一个观测数据来了后又来了一个新的观测数据，原先的后验就成了新的“先验”.

$Y=[Y_1,...,Y_n]$是条件独立同分布：

$$T_y^{(1)}[\cdot]=T_y^{(N)}[\cdot]=T_y[\cdot]$$

让我们令$p(\cdot)$是信念纠正：

$$q(\cdot;t_1|y_1)=T_y[p_X(\cdot)]$$

上述内容实际上就是一个*sufficient statistic.*

无论是先验还是在信念修正后的后验，都在Q内. 因此所有的参数必须在Θ内.

最后，我们用一张图来表示信念修正的具体原理：

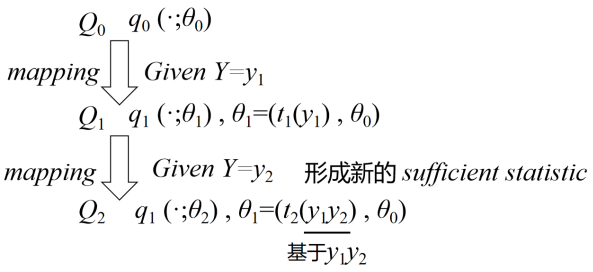

Fig. 2. 信念修正的具体原理. 每一次变化都是θ_0的迭代. Sufficient statistic的无损预处理的结果.

在这一章，我们介绍了三种构造先验的方法. LIP和MEP是有观测数据情况下使用的，共轭先验是在没有观测数据下使用的.

到这里，整个《大数据推断基础》的内容基本上结束了. 我们一起讨论了大数据的处理原理、香农信息论、决策估计、参数估计、指数簇、先验等内容，对推断的数学原理有了一些基本的了解. 这虽然是一门概论课，但是它给我们带来的思想启迪则是非常有价值的. 感谢这门课的讲授者——山东大学新闻传播学院于智源老师，感谢《大数据推断基础》！

Yin Tianyou
2023.6.8