# 概率统计中篇

## 一：数理统计基本知识

- 概率论：随机变量，分布已知
- 梳理统计：随机变量，分布**未知**，**通过观察值，对分布推断**
- 一个总体对应于一个随机变量$X$，$X_1, X_2, ... X_n$是随机样本，与$X$独立同分布（分布函数为F），$x_1, x_2, ..x_n$是样本值（观察值）
  - 举例如下：$X$:100名男性的身高；$X_1, X_2, ... X_n$：每名男性的身高；$x_1, x_2, ..x_n$：每次的样本观测值，显然第一次取到的观测值$x_1, x_2, ..x_n$和下一次取到的$y_1, y_2, ..y_n$一般是不同的。
- 统计量：样本平均值：$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

## 二：最大似然估计（MLE)

- 总体$X$的分布函数已知，但是一个或多个参数未知，我们借助样本来估计总体未知的参数值。
  - 我们可以这样来理解这句话：假设$X$服从高斯分布$\mathcal{N}(\mu, \sigma^2)$，但是$\mu$和$\sigma^2$是未知的，因此，我们打算借助样本来估计这些参数。
- 最大似然估计的主要思想：对于$P(A|\theta)$，在$\theta$的可能的取值范围内尽量选取使得$P(A|\theta)$最大的$\hat{\theta}$。

### 2.1 离散情形下的最大似然估计

可理解为男性身高

具体的身高数值

- 总体 $X$，$X_1,\cdots,X_N$ 来自 $X$ 样本，独立同分布，相应的观测值为 $x_1,\cdots,x_N$，参数取值未知．利用已知观测值 $x_1,\cdots,x_N$（常数）对 $\theta$ 进行点估计．

mu和sigma未知

身高数值已经通过测量得到

- 离散情况：实际中假定一 $\theta$，利用离散联合分布率定义，$X$，$X_1,\cdots,X_N$ 取到观测值为 $x_1,\cdots,x_N$ 的概率为 $L(\theta)=\prod_{i=1}^{n}p(x_i;\theta)$．这一概率随着 $\theta$ 变化，称为**样本**似然函数．

连乘是因为"独立同分布"

分号表明theta是个参数，给定不同的theta,L(theta)就会发生变化。

- 由于已经确认取到观测值，可认为取到这一样本的概率 $L(\theta)$ 比较大．显然，肯定不会去找让那些不能使样本 $x_1,\cdots,x_N$ 出现的 $\theta$ 作为估计值，如果 $\Theta$ 里面有个能让 $L(\theta)$ 取到最大的 $\hat{\theta}$，自然认为 $\hat{\theta}$ 就是 $\theta$ 的估计值．

从本质上来讲，是找到一个$\theta$使得这些样本出现的概率最大，但是现在$x_1,x_2,..x_n$这组样本已经出现了，那么我们就找一个$\hat{\theta}$，使得$L(\theta)$的值最大。这个$\hat{\theta}$就是$\theta$的估计值。就是说，其它的任意一个$\theta\neq\hat{\theta}$,$x_1,x_2,..x_n$出现的概率都小于$L(\hat{\theta})$。

## 2.2 连续情形下的最大似然估计

- 连续情况下联合概率密度为 $\prod_{i=1}^{n}f(x_i;\theta)$, 由于连续变量在某一点概率为 0，考虑随机变量 $X$，$X_1,\cdots,X_N$ 落在点 $(x_1,x_2,\cdots,x_n)$ 周围一个很小区域内（一维下就是求面积）的概率近似为 $\prod_{i=1}^{n}f(x_i;\theta)\,dx_i$，因此类似我们选取 $\theta$ 让 $\prod_{i=1}^{n}f(x_i;\theta)\,dx_i$（$dx_i$ 是宽度）最大，但是由于 $\theta$ 和 $dx_i$ 没关系，因此只需要 $\theta$ 让似然函数 $\prod_{i=1}^{n}f(x_i;\theta)=L(\theta)$ 最大即可．

- 注意：由于 $(x_1,x_2,\cdots,x_n)$ 已知，$L(\theta)$ 只和 $\theta$ 有关，是个标准的函数，既不是概率，也不是条件概率密度．

运用最大似然的步骤：

1. 区分离散还是连续
2. 
   - 在离散情形下：$p(x_i|\theta)\rightarrow L(\theta)=\prod_{i=1}^{n}p(x_i|\theta)$
   - 在连续情形下：$f(x_i|\theta)\rightarrow L(\theta)=\prod_{i=1}^{n}f(x_i|\theta)$
3. 最大化$L(\theta)$，求出$\hat{\theta}=\arg\min_{\theta}L(\theta)$.

## 2.3 最大似然估计举例

### 2.3.1 一元高斯分布

- 设 $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$，但 $\mu, \sigma^2$ 未知．$x_1 \cdots x_n$ 来自 $X$ 的一个样本值，求 $\mu, \sigma^2$ 最大似然估计
- 已知 $f\left(x; u, \sigma^2\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x-u)^2\right]$，似然函数

$$L\left(u, \sigma^2\right) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2\sigma^2}(x_i-u)^2\right]$$

- 联立求解 $\frac{\partial lnL}{\partial\mu} = 0$，和 $\frac{\partial lnL}{\partial\sigma^2} = 0$，可得

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

$$\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2$$

$$\ln L = n\ln\frac{1}{\sqrt{2\pi}\sigma} - \sum_{i=1}^{n}\frac{1}{2\sigma^2}(x_i-\mu)^2$$

$$\frac{\partial \ln L}{\partial \mu} = -\sum_{i=1}^{n}\frac{1}{\sigma^2}(x_i-\mu)(-1) = 0$$

$$\Rightarrow \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n}(x_i-\mu) = 0$$

$$\Rightarrow \sum_{i=1}^{n}x_i - n\mu = 0$$

$$\Rightarrow \hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \bar{x}$$

∴ 总体均值的估计值为样本均值.

令 $\sigma^2 = t$，∴ $\ln L = n\ln\frac{1}{\sqrt{2\pi t}} - \sum_{i=1}^{n}\frac{1}{2t}(x_i-\mu)^2$

$$= -\frac{n}{2}\ln(2\pi t) - \frac{1}{2t}\sum_{i=1}^{n}(x_i-\mu)^2$$

$$\therefore \frac{\partial \ln L}{\partial t} = -\frac{n}{2}\cdot\frac{2\pi}{2\pi t} + \frac{1}{2t^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$$

化简得：$-n + \frac{1}{t}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$

$$\Rightarrow t = \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2$$

∴ 总体方差的估值为样本方差.

### 2.3.2 多元高斯分布

$$f_{\mathbf{x}}\left(x_1, \ldots, x_k\right)=\frac{1}{\sqrt{(2 \pi)^k|\Sigma|}} \exp \left(-\frac{1}{2}(\mathrm{x}-\mu)^{\mathrm{T}} \Sigma^{-1}(\mathrm{x}-\mu)\right)$$

似然函数：

$$L=\prod_{i=1}^n f_X=\prod_{i=1}^n \frac{1}{\sqrt{(2 \pi)^k|\Sigma|}} \exp \left(-\frac{1}{2}\left(\mathrm{x}^{(\mathrm{i})}-\mu\right)^{\mathrm{T}} \Sigma^{-1}\left(\mathrm{x}^{(\mathrm{i})}-\mu\right)\right)$$

对数似然函数：

$$\begin{aligned} lnL &= -\frac{n}{2} ln\left[(2 \pi)^k|\Sigma|\right]-\frac{1}{2} \sum_{i=1}^n\left(\mathrm{x}^{(\mathrm{i})}-\mu\right)^{\mathrm{T}} \Sigma^{-1}\left(\mathrm{x}^{(\mathrm{i})}-\mu\right) \\ &= -\frac{nk}{2} ln(2 \pi)-\frac{n}{2} ln|\Sigma|-\frac{1}{2} \sum_{i=1}^n\left(\mathrm{x}^{(\mathrm{i})}-\mu\right)^{\mathrm{T}} \Sigma^{-1}\left(\mathrm{x}^{(\mathrm{i})}-\mu\right) \end{aligned}$$

仿照一元函数：可以得到如下估计值：

$$\begin{aligned} \mu_{ML} &= \frac{1}{N} \sum_{i=1}^N \mathbf{x}_i \\ \Sigma_{ML} &= \frac{1}{N} \sum_{i=1}^N\left(\mathbf{x}_i-\boldsymbol{\mu}_{ML}\right)\left(\mathbf{x}_i-\boldsymbol{\mu}_{ML}\right)^T \end{aligned}$$

### 2.3.3 (0-1)分布

$$\begin{aligned} & p(x=1|\mu)=\mu \\ & p(x=0|\mu)=1-\mu \\ & \operatorname{Bern}(x|\mu)=u^x(1-\mu)^{1-x} \\ & E(x)=\mu \\ & \operatorname{var}[x]=\mu(1-\mu) \end{aligned}$$

观测到一个数据集$\mathcal{D}=\{x_1, \ldots, x_N\}$，则似然函数为：

$$p(\mathcal{D}|\mu)=\prod_{n=1}^N p\left(x_n|\mu\right)=\prod_{n=1}^N u^{x_n}(1-\mu)^{1-x_n}$$

对数似然函数为：

$$lnp(\mathcal{D}|\mu)=\sum_{n=1}^N\left[x_n ln\mu+\left(1-x_n\right) ln(1-\mu)\right]$$

对$\mu$求导可得：

$$\frac{\partial lnp(\mathcal{D}|\mu)}{\partial\mu}=\frac{1}{\mu}\sum_{n=1}^{N}x_n-\frac{1}{1-\mu}(N-\sum_{n=1}^{N}x_n)=0$$

$$\hat{\mu}=\frac{1}{N}\sum_{n=1}^{N}x_n,$$

因为$x_n$只能取0或1，所以：

$$\hat{\mu}=\frac{m}{N},m为x_n=1的次数,N为实验总次数。$$

# 三：从MLE角度看线性回归与逻辑回归

## 3.1 线性回归

- 假设有n个数据点作为训练集，我们希望得到这样一个模型：给定一个新的输入$\hat{x}$,预测它对应的输出$\hat{y}$.
- 模型：$y^i=\theta^{\mathbf{T}}\mathbf{x}^i+\epsilon^i$，最后一项为误差项。
- 误差项$\epsilon^i\sim\mathcal{N}(0,\sigma^2)$,独立同分布，又因为$\theta^T\mathbf{x}^i$在给定某个样本的情况下是固定值，因此有：$y^i\sim\mathcal{N}(\theta^{\mathbf{T}}\mathbf{x}^i,\sigma^2)$，(可理解为高斯分布发生了偏移)
- 所以：
- $$f(y^i|\mathbf{x^i},\theta)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2}{2\sigma^2}}$$
- 似然函数：

$$L(\theta)=\prod_{i=1}^{n}\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2}{2\sigma^2}}$$

- 对数似然函数：

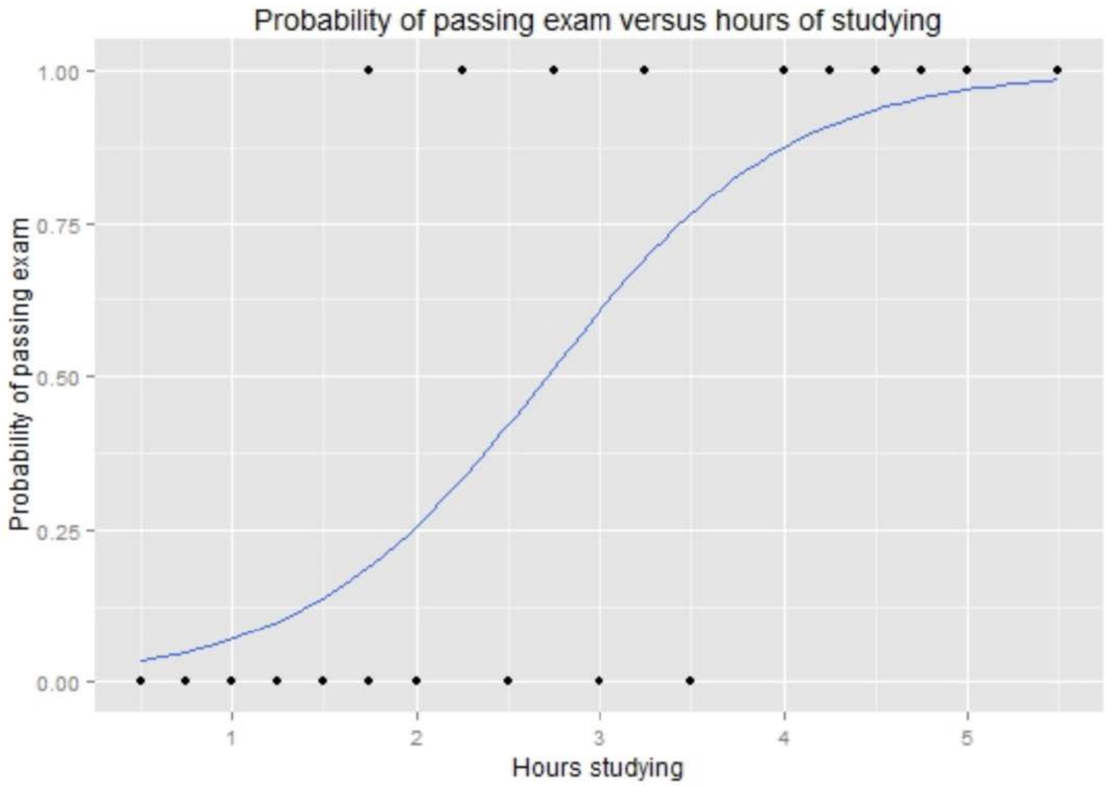
$$lnL(\theta)=nln(\frac{1}{\sigma\sqrt{2\pi}})-\sum_{i=1}^{n}\frac{(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2}{2\sigma^2}$$
$$=\underbrace{nln(\frac{1}{\sigma\sqrt{2\pi}})}_{常数}-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2$$

所以，若想使得似然函数最大，只能使得$\sum_{i=1}^{n}(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2$最小。即：

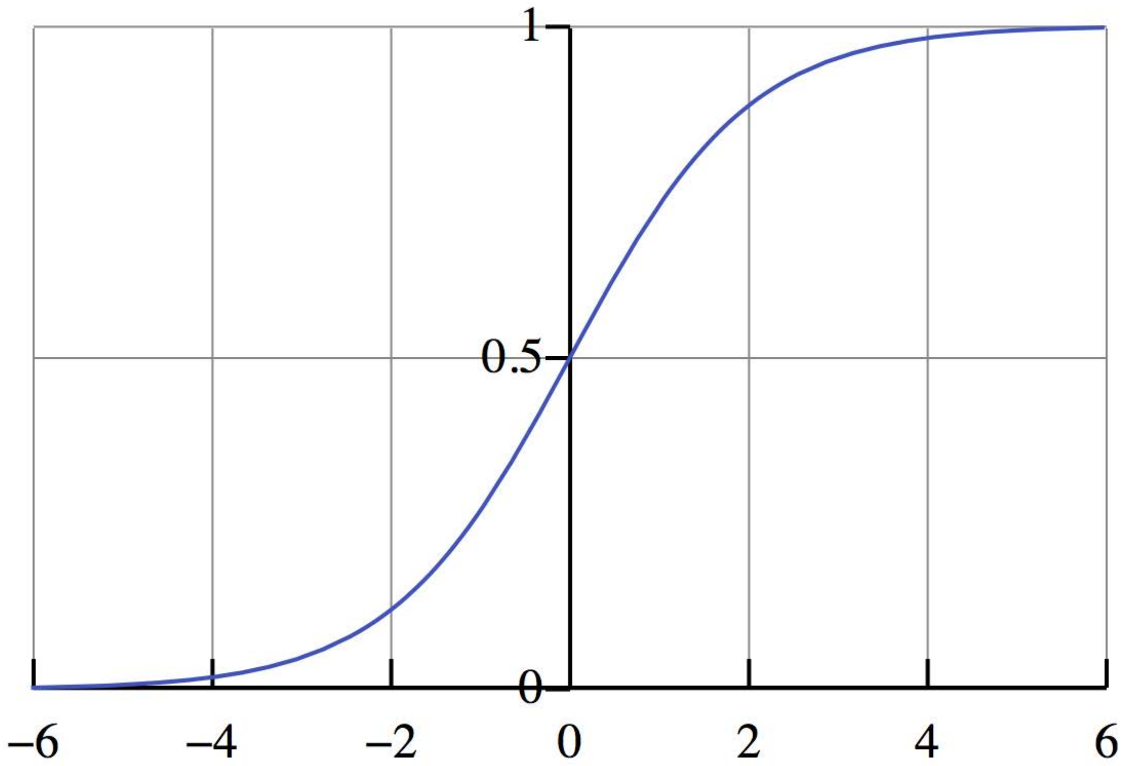
$$maxL(\theta)\Rightarrow min\sum_{i=1}^{n}(y^i-\theta^{\mathbf{T}}\mathbf{x}^i)^2\Rightarrow min(\mathbf{y}-\theta^{\mathbf{T}}\mathbf{x})^T(\mathbf{y}-\theta^{\mathbf{T}}\mathbf{x})\Rightarrow min||\mathbf{y}-\theta^{\mathbf{T}}\mathbf{x}||_2^2$$

所以，最小二乘等价于最大似然估计，前提是误差服从高斯分布，在实际中，一般使用“预测值使用高斯分布”的条件。

## 3.2 逻辑回归




- $y = \mathbf{w}^\mathbf{T}\mathbf{x} + b$
- 采用非线性映射：$z = \frac{1}{1+e^{-y}}$

- 

- 逻辑回归一定选取sigmoid函数，其实就是把y的值从$(-\infty,+\infty)$压缩到$(0,1)$

**其实，逻辑回归本质上对应于(0-1)分布**,说明如下：

令$h(\mathbf{x})=g(\theta^{\mathbf{T}}\mathbf{x})=\frac{1}{1+e^{-\theta^{\mathbf{T}}\mathbf{x}}}$

则：$h(\mathbf{x})$**代表了结果为1的概率。即y取1的概率为h(x)，y取0的概率为1-h(x)**

$$P(y=1|\mathbf{x},\theta)=h(\mathbf{x})$$
$$P(y=0|\mathbf{x},\theta)=1-h(\mathbf{x})$$

于是：

$$P(y|\mathbf{x},\theta)=h(\mathbf{x})^y(1-h(\mathbf{x}))^{1-y}$$

似然函数：

$$L(\theta)=\prod_{i=1}^{n}h(\mathbf{x}^i)^{y^i}(1-h(\mathbf{x}^i))^{1-y^i}$$

对数似然函数：

$$lnL(\theta)=\sum_{i=1}^{n}[y^i lnh(\mathbf{x}^i)+(1-y^i)ln(1-h(\mathbf{x}^i))]$$

其实，该函数是一个凹函数，也就是说$-lnL(\theta)$是一个凸函数，证明如下：

根据“凸函数的非负线性组合依旧是凸函数”的原则，我们只需要证明$-ln(h(\mathbf{x}))$和$-ln(1-h(\mathbf{x}))$是凸函数即可。

$$
\begin{aligned}
-ln(h(\mathbf{x})) &= ln(1+e^{-\theta^{\mathbf{T}}\mathbf{x}}) \\
\nabla_\theta ln(1+e^{-\theta^{\mathbf{T}}\mathbf{x}}) &= \frac{e^{-\theta^{\mathbf{T}}\mathbf{x}}}{1+e^{-\theta^{\mathbf{T}}\mathbf{x}}}(-\mathbf{x}) = (h(\mathbf{x})-1)\mathbf{x} \\
\nabla_\theta^2 ln(1+e^{-\theta^{\mathbf{T}}\mathbf{x}}) &= h(\mathbf{x})(1-h(\mathbf{x}))\mathbf{x}\mathbf{x}^{\mathbf{T}}
\end{aligned}
$$

注意：在求Hessian矩阵时，需要对$\mathbf{x}$做转置。

对于任意的向量$\mathbf{z}$，

$$
\mathbf{z}^{\mathbf{T}}h(\mathbf{x})(1-h(\mathbf{x}))\mathbf{x}\mathbf{x}^{\mathbf{T}}\mathbf{z} = \underbrace{h(\mathbf{x})(1-h(\mathbf{x}))}_{\text{常数}}\underbrace{(\mathbf{x}^{\mathbf{T}}\mathbf{z})^2}_{\text{常数}} \geq 0
$$

所以hessian矩阵是半正定矩阵，所以该函数是凸函数。

同理，可以证明$-ln(1-h(\mathbf{x}))$是凸函数。

这样，我们接下来就可以用梯度下降法来求解最优的$\theta$值了。

所以我们也可以得出这样的结论：逻辑回归的损失函数就是对数似然函数的负值。

为什么逻辑回归的损失函数不采用最小二乘呢?

- 原因1：逻辑回归本质上是从(0-1)分布而来，而线性回归本质上是从高斯分布而来，二者就不应该混用 。
- 原因2：假设使用最下二乘，那么损失函数为$\sum_{i=1}^{n}[y^i-g(\theta^{\mathbf{T}}\mathbf{x}^i)]^2$，但是$y^i$的取值只有0和1，而g函数的取值为$[0,1]$，两者都不对应，误差肯定很大，这个函数也不是凸函数，有许多局部极小值。