白板推导系列(一)　频率派 vs 贝叶斯派

$\boxed{\text{Github: CBJerry993}}$

X: data $\longrightarrow X=(x_1\ x_2 \cdots x_N)^T_{N\times p}$

θ: parameter

$$= \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \\ x_{N1} & x_{N2} & \cdots & x_{Np} \end{pmatrix}$$

$x \sim p(x|\theta)$ x服从概率模型.

频率派：θ是未知常量. x是随机变量(r.v)

MLE: $\theta_{mle} = \underset{\theta}{\operatorname{argmax}} \log p(x|\theta)$
(最大似然估计)

贝叶斯派：θ是随机变量，服从概率分布 $p(\theta)$ ↳先验

把参数先验、后验用似然联系.

贝叶斯 → 概率图模型
↳求积分 → mcmc 蒙特卡洛

频率 → 统计机器学习
↳优化 loss func，梯度下降

似然　　先验 prior

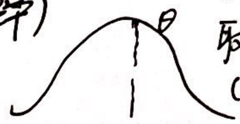
$$p(\theta|x) = \frac{p(x|\theta)\cdot p(\theta)}{p(x)} \propto p(x|\theta)\cdot p(\theta)$$

$p(x)$ ↳ $\int_\theta p(x|\theta)\cdot p(\theta)\, d\theta$

后验

posteriori [后验转化成似然和先验]

MAP: $\theta_{map} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|x) = \operatorname{argmax} p(x|\theta)\cdot p(\theta)$
(最大后验概率)



取θ最大的位置.
(认为θ服从概率分布 $p(\theta)$)

贝叶斯估计 $p(\theta|x) = \dfrac{p(x|\theta)\cdot p(\theta)}{\int_\theta p(x|\theta)\cdot p(\theta)\, d\theta}$

求 $p(\theta|x)$：实打实求解 (复杂)

贝叶斯预测，新样本 $\tilde{x}$

$$p(\tilde{x}|X) = \int_\theta p(\tilde{x},\theta|X)\, d\theta$$

$$= \int_\theta p(\tilde{x}|\theta)\cdot p(\theta|X)\, d\theta$$

后验

$X \underset{\theta}{\smile} \tilde{x}$

求 $p(\tilde{x}|X)$ 在其中引入θ，求 $p(\tilde{x}|\theta)$ 与 $p(\theta|X)$.

白板推导系列（2） 数学基础 高斯分布 → 极大似然

Github: CBJerry993

Data $X=(x_1\ x_2\ \cdots\ x_N)^T=\begin{pmatrix}x_1^T\\x_2^T\\\vdots\\x_N^T\end{pmatrix}_{N\times P}$

$x_i\in\mathbb{R}^P$

$x_i\overset{iid}{\sim}N(\mu,\Sigma)$ iid: 独立同分布

MLE: $\theta_{MLE}=\underset{\theta}{\arg\max}\,P(X|\theta)$

令 $p=1$，一维 $\theta=(\mu,\sigma^2)$

$p(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$

$\log P(X|\theta)=\log\prod_{i=1}^{N}p(x_i|\theta)=\sum_{i=1}^{N}\log p(x_i|\theta)$

$=\sum_{i=1}^{N}\log\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

$=\sum_{i=1}^{N}\left(\log\frac{1}{\sqrt{2\pi}}+\log\frac{1}{\sigma}-\frac{(x_i-\mu)^2}{2\sigma^2}\right)$

μ

$\mu_{MLE}=\underset{\mu}{\arg\max}\log P(X|\theta)$ 把与μ无关去掉

$=\underset{\mu}{\arg\max}\sum_{i=1}^{N}-\frac{(x_i-\mu)^2}{2\sigma^2}=\underset{\mu}{\arg\min}\sum_{i=1}^{N}(x_i-\mu)^2$

求偏导

$\frac{\partial\sum(x_i-\mu)^2}{\partial\mu}=\sum_{i=1}^{N}2(x_i-\mu)\cdot(-1)=0$

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

$\Rightarrow\mu_{MLE}=\frac{1}{N}\sum_{i=1}^{N}x_i$ （均值）

无偏估计

$E[\mu_{mle}]=\frac{1}{N}\sum_{i=1}^{N}E(x_i)=\frac{1}{N}\sum_{i=1}^{N}\mu=\frac{1}{N}\cdot N\cdot\mu=\mu$

σ^2

$\sigma^2_{MLE}=\underset{\sigma}{\arg\max}\,P(X|\theta)$

$=\underset{\sigma}{\arg\max}\sum_{i=1}^{N}\underbrace{\left(-\log\sigma-\frac{1}{2\sigma^2}(x_i-\mu)^2\right)}_{令\ L(\sigma)}$

求偏导

$\frac{\partial L(\sigma)}{\partial\sigma}=\sum_{i=1}^{N}\left[-\frac{1}{\sigma}+\frac{1}{\cancel{2}}(x_i-\mu)^2\cdot\sigma^{-3}\right]=0$

$=\sum_{i=1}^{N}\left[-\sigma^2+(x_i-\mu)^2\right]=0$

$-\sum_{i=1}^{N}\sigma^2+\sum_{i=1}^{N}(x_i-\mu)^2=0$

$\sum_{i=1}^{N}\sigma^2=\sum_{i=1}^{N}(x_i-\mu)^2$

$\sigma^2_{mle}=\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu_{mle})^2$ 有偏估计

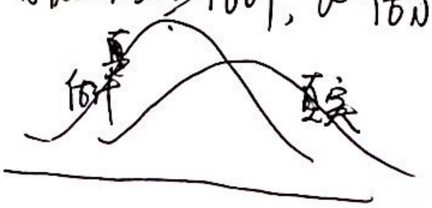
$E[\sigma^2_{mle}]=\frac{N-1}{N}\sigma^2$ (比原来小)

无偏 $\hat{\sigma}=\frac{1}{N-1}\sum_{i=1}^{N}(x_i-\mu_{mle})^2$

$E[\sigma^2_{mle}]=\sigma^2-\frac{1}{N}\sigma^2=\frac{N-1}{N}\sigma^2$

下节推导

使用极大似然估计，σ^2 估小，



白板推导系列(2) 数学基础 高斯分布-极大似然估计

- 有偏 vs. 无偏

Github: CBJerry993

$$E[\mu_{MLE}] = E\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N}\sum_{i=1}^{N} E(x_i) = \frac{1}{N}\sum_{i=1}^{N} \mu = \mu \Rightarrow \text{无偏}$$

$$E[\sigma_{mle}^2] \overset{?}{=} \sigma^2$$

$$\sigma_{mle}^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu_{MLE})^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i^2 - 2x_i\mu_{mle} + \mu_{mle}^2)$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N}\sum_{i=1}^{N} 2x_i\mu_{mle} + \frac{1}{N}\sum_{i=1}^{N}\mu_{mle}^2$$

$$\frac{1}{N}\sum x_i = \mu_{mle}$$

$$= \frac{1}{N}\sum_{i=1}^{N} x_i^2 - 2\mu_{mle}^2 + \mu_{mle}^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{mle}^2$$

同时 $\pm\mu^2$

那么 $E(\sigma_{mle}^2) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu_{mle}^2\right) = E\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right) - E(\mu_{mle}^2 - \mu^2)$

$$E\left(\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right) = E\left(\frac{1}{N}\sum_{i=1}^{N}(x_i^2 - \mu^2)\right) = \frac{1}{N}\sum_{i=1}^{N} E(x_i^2 - \mu^2) = \frac{1}{N}\sum_{i=1}^{N}\underbrace{(E(x_i^2) - \mu^2)}_{\sigma^2} = \frac{1}{N}\sum_{i=1}^{N}\sigma^2 = \sigma^2$$

$$D(x) = E[(x - E(x))^2] = E(x^2) - [E(x)]^2 = \sigma^2$$

$$E(\mu_{mle}^2) - E(\mu^2) = E(\mu_{mle}^2) - \mu^2 = E(\mu_{mle}^2) - E^2(\mu_{mle}) = Var(\mu_{mle}) = \frac{1}{N}\sigma^2$$

$$= \sigma^2 - \frac{1}{N}\sigma^2$$

$$Var[\mu_{mle}] = Var\left[\frac{1}{N}\sum_{i=1}^{N} x_i\right] = \frac{1}{N^2}\sum_{i=1}^{N} Var(x_i) = \frac{1}{N}\sigma^2$$