MLE, MAP和Bayesian Estimation 概述

Created by **Qi Yang**

1. 贝叶斯公式

需要先从贝叶斯公式谈起了： 数据集（观测集）的误差和局限导致了不确定性，为了量化和估计，就产生了概率论的研究。

$$P = \frac{\text{发生数}}{\text{总观察数}}$$

概念

> - 加和规则 sum -> 边缘概率 -> $p(x) = \sum_{Y}{p(x|y)}$
> - 乘积规则 product -> 联合概率 -> $p(x,y) = p(x|y)p(y) = p(y|x)p(x)$

所以得到bayes公式：$p(x|y) = \frac{p(y|x)p(x)}{p(y)}$

$n_{1,1}$	$n_{1,2}$	...	$n_{1,j}$	...	$n_{1,n}$
$n_{2,1}$	$n_{2,2}$	...	...	...	$n_{2,n}$
...	...	...	...	...	...
$n_{i,1}$	...	...	$n_{i,j}$	...	$n_{i,n}$
...	...	...	...	...	...
$n_{m,1}$	$n_{m,2}$	...	$n_{m,j}$	...	$n_{m,n}$

如上图，推导概率的加和规则和乘积规则：如西瓜书（周志华著）中的选西瓜案例，行m可以表示西瓜的好坏（结果，观察值），列n可以表示西瓜的特征，其他实例同理，比如病症和病人体检结果等。 考虑两个变量X，取值${x_i}$，其中$i=1,...,m$（第i行），和变量Y，取值${y_j}$，其中$j=1,...,n$（第j列）。注意，这里我用x表示行，用y表示列，这和PRML原文中略有出入，但不影响推论和结果。

显然，**边缘概率** $p(X=i) = \frac{c_j}{N_{total}}$，其中$c_j = \sum_{k=1}^{n} n_{i,k}$ 也即**加和规则**。

$$p(x=i) = \sum_{k=1}^{n} \frac{n_{i,k}}{N_{total}}$$

且 $p(Y=j)$ 同理，由其他变量的边缘化或者加和求得。

条件概率，我们可以考虑在$x=i$的势力中，$y=j$的概率（比例），可以写作条件概率$p(y=j|x=i$,计算方式可以为为单元格i,j的值与列i总数的比例，即：

$$p(y=j|x=i) = \frac{n_{i,j}}{c_j}$$

进一步，**全概率**：$p(x=i,y=j) = \frac{n_{i,j}}{N_{total}}$ 所以可以得到**乘法规则**：

$$p(x=i,y=j) = \frac{n_{i,j}}{N_{total}} = \frac{n_{i,j}}{c_j} \cdot \frac{c_j}{N_{total}} = p(y=j|x=i)p(X=i)$$

先验知识 prior: $p(\theta)$
后验概率 posterior: $p(\theta|x)$

- $p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)}$
- $posterior = \frac{likehood \cdot prior}{evidence}$

频率学派（Frequentist） 和 贝叶斯学派（Bayesian）的争论：
频率学家认为参数是固定值，所以应该通过“估计”来确定 贝叶斯学派的观点是在仅有实际的观察数据集的时候，参数的不确定性通

过概率分布来表示，需要应用先验知识。他们将概率解释成信念（belief）的度量，或者一个事件发生的信心（confidence）。

举例，事件A发生：

> - P(A): 抛硬币，猜正反面。P(A|X): 庄家的口碑，表示该信息为X，比如是韦小宝，认定他作弊的概率
> - P(A): 代码可能有一个bug。P(A|X): 写代码的是个新手，或者是个老鸟

一名贝叶斯主义者在看到证据后的信念的更新,表示为P(A|X)，解释为在给定证据X后的A事件发生的概率(后验概率）。

---

2. 最大似然估计

到这里就可以继续介绍最大似然了，简而言之，是给定数据集的情况下最大化概率的参数：给定一堆数据，假如我们知道它是从某一种分布中随机取出来的，可是我们并不知道这个分布具体的参，即**"模型已定，参数未知"**。例如，我们知道这个分布是正态分布，但是不知道均值和方差；或者是二项分布，但是不知道均值。 最大似然估计（MLE，Maximum Likelihood Estimation）就可以用来估计模型的参数。MLE的目标是找出一组参数，使得模型产生出观测数据的概率最大：

$$L(\theta)=L(x_1,x_2,\dots,x_n|\theta)=\prod_{i=1}^{n}p(x_i|\theta),\theta\in\Theta$$

最大似然： $\hat{\theta} = \mathrm{argmax}{\hat{l}(\theta)}$

$$H(\theta)=\ln(L(\theta))=\ln\prod_{i=1}^{n}p(x_i|\theta)=\sum_{i=1}^{n}\ln p(x_i|\theta)$$

怎么求一个函数的最值？当然是求导，然后让导数为0，那么解这个方程得到的θ就是了（当然，前提是函数L(θ)连续可微）。那如果θ是包含多个参数的向量那怎么处理啊？当然是求L(θ)对所有参数的偏导数，也就是梯度了，那么n个未知的参数，就有n个方程，方程组的解就是似然函数的极值点了，当然就得到这n个参数了。 即：

- 单个变量：$\frac{dH(\theta)}{d\theta}=\frac{dl\mathit{nl}(\theta)}{d_\theta}=0$
- 多个变量：θ 可以用向量表示

> - 未知参数：$\theta = [\theta_1,\theta_2,...,\theta_s]^T$
> - 梯度算子：$\nabla_{\theta}H(\theta)=\left[\frac{\partial}{\partial \theta_1},\frac{\partial}{\partial \theta_2},...\frac{\partial}{\partial \theta_s}\right]^T$
> - 解方程: $\nabla_{\theta}\mathit{H}(\theta) = \nabla_{\theta}\mathrm{ln}l(\theta)=\sum_{i=0}^{N}\nabla_\theta\mathrm{ln}P(x_i|\theta) = 0$

求最大似然函数估计值的一般步骤：

- （1）写出似然函数；
- （2）对似然函数取对数，并整理；
- （3）求导数，令导数为0，得到似然方程；
- （4）解似然方程，得到的参数即为所求；

以高斯分布为例，参数分别是μ 和 σ^2

- 似然函数的取对数：$\mathrm{ln}L(\mu,\sigma^2)=-\frac{n}{2}ln(2\pi)-\frac{n}{2}ln(\sigma^2)-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(x_i-\mu\right)^2$
- 求导后的方程组：$

  $$\left\{ \frac{\partial \ln L(\mu,\sigma^2)}{\partial\mu}=\frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)=0 \quad \frac{\partial \ln L(\mu,\sigma^2)}{\partial\sigma^2}=-\frac{n}{2\sigma^2}+\frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2=0 \right.$$

  $
- 求解的结果：$

  $$\left\{ \mu^\star=\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i \quad \sigma^{\star 2}=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2 \right.$$

  $
- $\left(\mu^\star,\sigma^{\star 2}\right)$

---

3. 极大后验估计
MAP与MLE最大区别是MAP中加入了模型参数本身的概率分布

MAP允许我们把先验知识加入到估计模型中，这在样本很少的时候是很有用的，因为样本很少的时候我们的观测结果很可能出现偏差，此时先验知识会把估计的结果“拉”向先验，实际的预估结果将会在先验结果的两侧形成一个顶峰。通过调节先验分布的参数，比如beta分布的，我们还可以调节把估计的结果“拉”向先验的幅度，越大，这个顶峰越尖锐。这样的参数，我们叫做预估模型的“超参数”。

我们以伯努利分布为例（抛硬币），假如我们的经验告诉我们，硬币一般都是匀称的，也就是$\mu = 0.5$ 的可能性最大， $\mu = 0.2$ 的可能性比较小，那么参数该怎么估计呢？这就是MAP要考虑的问题。 MAP优化的是一个后验概率，即给定了观测值后使概率最大：

$$\hat{\mu}\mathrm{MAP} = \mathrm{arg max}\mu{p(\mu|X)} = \mathrm{argmax}{\mu}\frac{{p(X|\mu)}{p(\mu)}}{p(X)} \propto \mathrm{argmax}{\mu}{p(X|\mu)}{p(\mu)}$$

我们可以看出第一项就是似然函数，第二项就是参数的先验知识。取log之后就是:

$$\mathrm{argmax}_{\mu}\mathrm{Pr}(\mu|X) = \mathrm{argmax}_{\mu} log\mathrm{Pr}(\mu|X) = \mathrm{argmax}_{\mu} log\prod_{x_i\in X}\mathrm{Pr}(x_i|\mu)\cdot \mathrm{P}r(\mu) = \mathrm{argmax}_{\mu} log\sum_{x_i\in X} log\mathrm{P}r(x_i|\mu) + log\mathrm{Pr}(\mu)$$

那么目标函数的导数即为：

$$\frac{\partial}{\partial\mu}\mathcal{L} = \sum_i \frac{\partial}{\partial\mu} log\mathrm{Bernoulli}(x_i|\mu) + \frac{\partial}{\partial\mu} log\mathrm{Beta}(\mu|\alpha,\beta)$$

令导数为0，即的所求。

---

4. Bayesian Estimation

MLE和MAP都属于频率学派，而贝叶斯推断为贝叶斯学派（显而易见）。
在贝叶斯公式中，

$$p(\theta|X) = \frac{p(X|\theta)p(\theta)}{p(X)} = \frac{p(X|\theta)p(\theta)}{\int p(X|\theta)p(\theta)d\theta}$$

分母为归一化常数，确保了后验概率分布的合理性，积分为1. 则：

$$p(X) = \int p(X|\theta)p(\theta)d\theta = \int \mathrm{likelihood}\cdot\mathrm{prior}$$

也被称为边缘似然度(Marginal likelihood)或证据(evidence)，

在贝叶斯方法中，需要对参数的所有值进行积分。如上，贝叶斯方法通常考虑的是整个后验概率（这意味着我们不会因为与θ无关而忽略掉贝叶斯公式中的归一化项）。简单的说，对于贝叶斯方法，需要自始至终的使用加和和乘积。

如话题模型（上帝掷色子）

在对新的数据进行预测的时候：

$$p(x_{new}) = \int_\theta p(x|\theta)p(\theta|X)d\theta$$

为了得到后验分布$p(\theta|X)$,需要进行全局积分，需要用到采样。

5. 总结

Frequentist vs. Bayesian ：

- 频率学派:
    - 数据是可重复的随机采样(frequency)，eg. bootstrap
    - 参数固定

- 贝叶斯：
    - 数据是可信采样
    - 参数未知，服从特定分布
    - 数据不变

MLE vs. MAP vs. Bayesian

- 数据X，参数θ, 新数据 $\theta^\star$
- Maximum likelihood estimation:
    - 目标函数： $\mathrm{argmax}P(X|\theta)$
- Maximum a Posteri estimation:
    - 目标函数： $\mathrm{argmax}P(\theta | X)$
- Bayesian estimation：
    - 目标函数： $\int p(X|\theta)p(\theta) d \theta$

最大似然估计是最简单的形式，其假定参数虽然未知，但是为确定数值，就是找到使得样本的似然分布最大的参数。最大后验估计，和最大似然估计很相似，也是假定参数未知，但是为确定数值，只是目标函数为后验概率形式，多了一个先验概率项, **先验知识的加入,优化损失函数**。

而贝叶斯估计和二者最大的不同在于，假定把待估计的参数看成是符合某种先验概率分布的随机变量，而不是确定数值。在样本分布上，计算参数所有可能的情况，并通过计算参数的期望，得到后验概率密度:**ML和MAP只会给出一个最优的解， 然而贝叶斯模型会给出对参数的一个分布**

对样本进行观测的过程，就是把先验概率密度转化为后验概率密度，这样就利用样本的信息修正了对参数的初始估计值。在贝叶斯估计中，一个典型的效果就是，每得到新的观测样本，都使得后验概率密度函数变得更加尖锐，使其在待估参数的真实值附近形成最大的尖峰。——这个后续的同学讲解共轭和话题模型的时候会更多的涉及到。

supplement

需要补充的内容，用来加深理解。 在线性回归的问题中，给定了N个观察值${x_n}$，对应其目标值的${t_n}$的数据集，我们的目标是可以在给定新的x值的情况下，预测出t的值。直观的解决办法，构建函数$y(x)$,直接求出变量x的函数值即为预测值。

1. 线性基函数
   最简单的线性回归模型：

   $$y(x,w) = w_0 + w_1 \cdot x_1 + \ldots + w_D \cdot x_D$$

   这个模型中最关键的是参数集合 $\textbf{w}$ 和 变量集合 $\textbf{x}$, 为了克服线性模型的局限性，可以将输入变量转变为非线性的函数进行线性组合，即：

   $$y(x,w) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(x)$$

   其中，$\phi_j(x)$被称为基函数。通常会定义$\phi_0(x)=1$,这时候：

   $$y(x,w) = \sum_{j=1}^{M-1} w_j \phi_j(x) = \mathbf{w}^T \phi(x)$$

   M为参数的总数，基函数的选择可以多样,sigmoid, gaussian, polynormial

   对于给定的参数$\textbf{w}$，我们利用误差函数来衡量预测值与真实数据集的差别：

   $$E(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (y(x_n, w) - {t_n}^2$$

   通过最小化$E(\textbf{w})$来解决曲线拟合问题。

   进一步，为了控制过拟合的现象，会加入惩罚项（正则化）。

$$E(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y(x_n, w) - t_n{}^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

, 其中 $||\textbf{w}||^2=\textbf{w}^T\textbf{w}$

2. 曲线拟合

我们首先考虑高斯分布的有偏估计的情况。无偏估计中，参数的期望值应该和参数值相同，—— 理解样本方差和分布方差的区别。

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

所以对高斯分布的概率估计是有偏估计，整个推导也比较简单：

$$E(\sigma^{\star 2}) = E[\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2] = E[\sum_{i=1}^{n}X_i^2 - 2\sum_{i=1}^{n}X_i\bar{X} - \sum_{i=1}^{n}\bar{X}^2]$$

$$= E[\sum_{i=1}^{n}X_i^2] - E[2\hat{X}\sum_{i=1}^{n}X_i] - E[n\bar{X}^2] = \frac{1}{n}(nE[X_i^2] - nE[\hat{X}^2])$$

所以有：

$$E(\sigma^{\star 2}) = E[X_i^2] - n[\bar{X}^2]$$

又有： $V(x) = E(x^2)-(E(x))^2$, $E(x_i^2) = V(x_i^2) + (E(x_i))^2 = \sigma^2 + \mu^2$, $E(\hat{x}^2) = V(\bar{x}^2) + (E(\bar{x}))^2 = \frac{\sigma^2}{n} + \mu^2$ ，带入即得。加入因子$\frac{N-1}{N}$可变为无偏。

在另一方面，我们可以使用概率分布来表达关于目标变量的值的不确定性。即假设目标变量满足均值为$y(x,w)$，方差为超参数$\beta = \frac{1}{\sigma^2}$ 的高斯分布——中心极限定律。这时候给出变量x,得到目标值t的概率：

$$p(t|x, \mathbf{w}, \beta) = \mathcal{N}(t|y(x, (w), \beta))$$

利用最大似然估计的方法，得到对数似然函数：

$$\ln L(t|x, \mathbf{w}, \beta) = -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y(x_n, (w) - t_n))^2 - \frac{n}{2}ln(2\pi) - \frac{n}{2}ln(\sigma^2)$$

求得多项式系数的最大似然解（后两项无关），等价的最小化平方和误差函数。接着确定精度（方差），然后带入预测公式即可。

同样的，我们可以进一步引入多项式系数$\textbf{w}$的先验分布，假设分布为均值为0，精度为α（方差倒数）。使用贝叶斯定律，$\textbf{w}$ 的后验概率正比于先验概率和似然函数的乘积：

$$p(\mathbf{w}|x, t, \alpha, \beta) \propto p(t|x, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)$$

同理，此时我们最大化后验概率，只需要最小化下式：

$$\frac{\beta}{2}\sum_{n=1}^{N}y(x_n, \mathbf{w}) - t_n{}^2 + \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}$$

显然，等价于正则后的平方和误差函数。

3. Beyesian 拟合

当我们讨论了曲线拟合中的最大似然估计和最大后验估计的方法后，其实依然在进行对参数的点估计，在前面讨论贝叶斯估计的时候，我们给出的预测概率： $p(x_{new})=\int_\theta p(x|\theta)p(\theta|X)d\theta$应该改写成：

$$p(t|x, \mathbf{x}, \mathbf{t}) = \int_w p(t|x, w)p(w|\mathbf{x}, \mathbf{t})d\mathbf{w}$$

, 其中，$\textbf{x,t}$为训练集合，而x为新的测试点，t为预测值(省略了超参数α,β)。 类似的，对积分进行解析求解，得到预测分布的高斯形式：

$$p(t|x,\mathbf{x},\mathbf{t}) = \mathcal{N}(t|m(x), s^2(x))$$

其中，均值和方差分别为：

$$m(x) = \beta\phi(x)^T S \sum_{n=1}^{N} \phi(x_n) t_n$$

$$s^2(x) = \beta^{-1} + \phi(x)^T S \phi(x)$$

可以看出，预测分布的均值和方差都依赖与x，而且，方差的第一项表示了预测值的不确定性，有目标变量的噪声造成。第二项为参数的不确定性。具体的推导和S的含义，还请继续阅读PRML-2.3和3.5节，这里不多介绍了。

---

reference：

1. PRML-1.2
2. https://blog.csdn.net/zengxiantao1994/article/details/72787849
3. https://www.cnblogs.com/sylvanas2012/p/5058065.html