机器学习

于元隆、朱丹红

福州大学计算机与大数据学院
Email: yu.yuanlong@fzu.edu.cn

贝叶斯分类器的训练

贝叶斯分类器学习方式

- 给定带标签的训练样本，可以采用监督式学习技术训练先验概率和观测似然：
  - 参数化方法：给定概率分布的解析表达（例如，高斯），学习（估计）这些解析表达式中的参数（例如，高斯分布的均值和协方差）。该类方法也称为参数估计。
  - 非参数化方法：概率分布形式未知，基于概率密度估计技术，估计非参数化的概率分密度表达。
- 注意：贝叶斯分类器所需学习的先验概率和观测似然都是针对单独的每个类而言。

参数化方法

- 假设观测似然是高斯分布，需要估计的参数为：

$$\hat{\mu},\ \hat{\sigma}^2 \text{ for 1-D and } \underline{\hat{\mu}},\ \hat{\Sigma} \text{ for multi-D}$$

- 假设是二分类，训练样本有N1+N2=N个，先验概率需要估计的参数为P：

$$P(C_1) = P \qquad P(C_2) = 1 - P$$

- 常用的两种参数估计方法（监督式）：
  - 最大似然估计(Maximum Likelihood Estimation)
  - 贝叶斯估计（Bayesian Estimation）

最大似然估计

设：ω_i类的观测似然函数具有某种确定的函数形式；

θ是该函数的一个未知参数或参数集。

最大似然估计把θ当作确定的未知量进行估计。

1、似然函数

从ω_i类中独立地抽取N个样本，$X^N=\{\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N\}$

称在参数θ下观测到的样本集X^N的联合概率密度函数，$p(X^N|\boldsymbol{\theta})$，为相对于样本集$X^N$的$\theta$的似然函数。

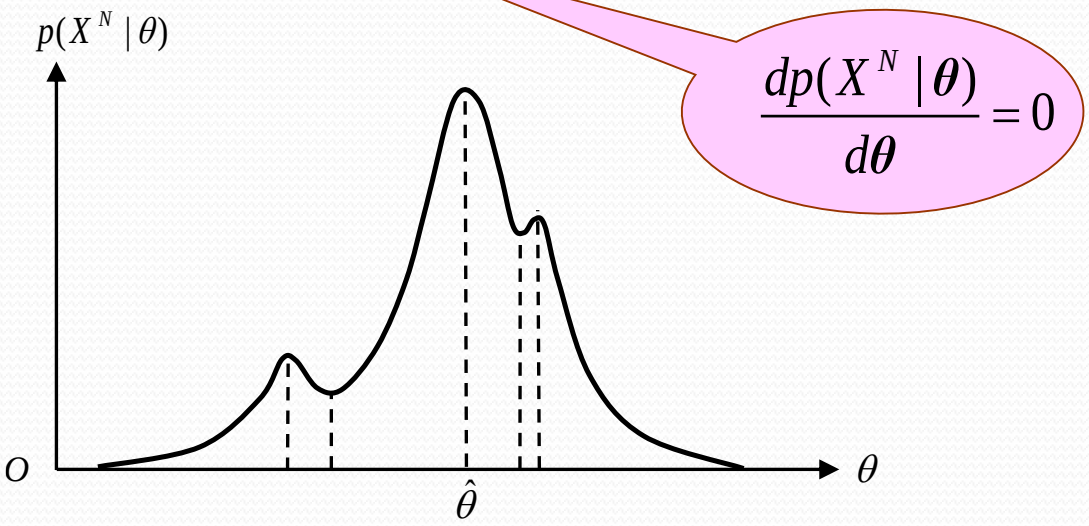
$$p(X^N|\boldsymbol{\theta})=p(\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N|\boldsymbol{\theta})=\prod_{k=1}^{N}p(\boldsymbol{X}_k|\boldsymbol{\theta})$$

2、最大似然估计

根据已经抽取的N个样本估计这组样本**“最可能”**来自哪个密度函数。（“最似”哪个密度函数）

也即：要找到一个θ，它能使似然函数 $p(X^N|\theta)$ 极大化。

θ的最大似然估计 $\hat{\theta}$ 就是使似然函数达到最大的估计量。



θ为一维时的最大似然估计示意图

为便于分析，定义似然函数的对数为：$H(\boldsymbol{\theta}) = \ln p(X^N \mid \boldsymbol{\theta})$

θ的最大似然估计是下面微分方程的解：$\dfrac{dH(\boldsymbol{\theta})}{d\boldsymbol{\theta}} = 0$

设ω_i类的概率密度函数有p个未知参数，记为p维向量：$\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots, \theta_p]^{\mathrm{T}}$

此时 $H(\boldsymbol{\theta}) = \ln p(X^N \mid \boldsymbol{\theta}) = \sum_{k=1}^{N} \ln p(X_k \mid \boldsymbol{\theta})$

由：

$$\frac{\partial}{\partial \boldsymbol{\theta}} \left[\sum_{k=1}^{N} \ln p(X_k \mid \boldsymbol{\theta}) \right] = 0 \rightarrow \begin{cases} \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_1} \ln p(X_k \mid \boldsymbol{\theta}) = 0 \\ \cdots\cdots \\ \sum_{k=1}^{N} \dfrac{\partial}{\partial \theta_p} \ln p(X_k \mid \boldsymbol{\theta}) = 0 \end{cases}$$

解以上微分方程即可得到θ的最大似然估计值。

最大似然估计：先验概率估计

- 最大似然估计的基本思想：基于所有训练样本构建似然函数作为目标函数，目标函数以待训练参数为参数。通过最大化似然函数，求解最优的训练参数。

- 先验概率的估计
  - 先验概率的似然函数：给定N个训练样本和C_1 类的先验概率P，选取到N_1个属于C_1类样本的概率即为先验概率的似然函数，即Binomial分布。
  - N_1是自变量，N是已知的样本个数，P是待学习的参数。

$$P(N_1|N,P) = \binom{N}{N_1} P^{N_1}(1-P)^{N_2}$$

$$= \frac{N!}{N_1!(N-N_1)!} P^{N_1}(1-P)^{N-N_1}$$

Binomial分布

- Bernoulli分布：一个事件有两个状态，成功(1)和失败(0)，即随机变量x取值是1或者0。P表示该事件执行一次取值是1的概率。则该事件执行一次、成功的次数x的分布概率即为Bernouli分布：

$$\begin{cases} p(x=1) = P \\ p(x=0) = 1-P \end{cases} \Rightarrow p(x) = P^x(1-P)^{1-x}, \text{ where } x = 0, 1$$

$$E[x] = P, Var(x) = P(1-P)$$ P是该分布的参数。

- Binomial分布：一个事件有两个状态，成功(1)和失败(0)。如果该事件独立的执行N次，成功次数x的分布概率即为Binomial分布。

$$p(x) = \begin{pmatrix} N \\ x \end{pmatrix} P^x(1-P)^{N-x}, \text{ where } x = 0, \ldots, N$$

$$E[x] = NP, Var(x) = NP(1-P)$$ N和P是该分布的参数。

最大似然估计：先验概率估计

- 先验概率估计：给定N个样本及其标签，最大化先验概率的似然函数，从而求解得到P的估计值$\hat{P}_{ML}$。
- 为了最大化似然函数，首先求似然函数关于参数P的偏导，并设偏导为0。
- 最终得到，先验概率的最大似然估计是该类训练样本出现的频率。

$$\frac{\delta}{\delta P}\{P(N_1|N,P)\} =$$

$$\binom{N}{N_1}\{N_1P^{N_1-1}(1-P)^{N-N_1} - (N-N_1)P^{N_1}(1-P)^{N-N_1-1}\} = 0$$

$$N_1(1-P) - (N-N_1)P = 0$$

$$N_1 - NP = 0$$

$$\Rightarrow \hat{P}_{ML} = \frac{N_1}{N}$$

最大似然估计：高斯分布参数估计

- 高斯分布参数估计
  - 高斯分布的似然函数：给定一个类高斯分布的两个参数μ和Σ，可以把属于该类的所有训练样本的高斯联合分布作为似然函数。
  - 假设该类的各个训练样本之间相互独立；
  - 对于该似然函数，属于该类的训练样本$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$是自变量，$N$是该类的训练样本个数，$\mu$和$\Sigma$是待学习的参数。

$$p(\underline{x}_1 \cdots \underline{x}_N | \mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})}$$

最大似然估计：高斯分布参数估计

- 高斯分布参数估计
  - 分别求似然函数关于两个参数μ和Σ 的偏导，设置偏导等于0。

$$\frac{\delta}{\delta \underline{\mu}} p(\underline{x}_1 \dots \underline{x}_N | \underline{\mu}, \Sigma) = 0$$

$$\frac{\delta}{\delta \Sigma} p(\underline{x}_1 \dots \underline{x}_N | \underline{\mu}, \Sigma) = 0$$

$$p(\underline{x}_1 \dots \underline{x}_N | \mu, \Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2} |\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}_i - \underline{\mu})^T \Sigma^{-1} (\underline{x}_i - \underline{\mu})}$$

- 高斯分布参数估计 $p(\underline{x}_1 \dots \underline{x}_N|\mu,\Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}_i-\underline{\mu})^T\Sigma^{-1}(\underline{x}_i-\underline{\mu})}$
  - 似然函数关于参数μ的偏导，设置偏导等于0.

$$\frac{\delta}{\delta\underline{\mu}}\left\{-N\,log\left((2\pi)^{n/2}|\Sigma|^{1/2}\right)-\frac{1}{2}\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})^T\Sigma^{-1}(\underline{x}_i-\underline{\mu})\right\}$$

$$=-\frac{1}{2}\frac{\delta}{\delta\underline{\mu}}\left\{\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})^T\Sigma^{-1}(\underline{x}_i-\underline{\mu})\right\}$$

$$=-\frac{1}{2}\sum_{i=1}^{N}(-2\Sigma^{-1}(\underline{x}_i-\underline{\mu})) \qquad = 0 \text{ to maximize}$$

$$\frac{\partial \mathbf{x}^T\mathbf{M}\mathbf{x}}{\partial \mathbf{x}} = [\mathbf{M}+\mathbf{M}^T]\mathbf{x}$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$$

$$\frac{\partial(\mathbf{x}_i-\boldsymbol{\mu})}{\partial \boldsymbol{\mu}} = -\mathbf{I}$$

$$\Rightarrow \underline{\hat{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i = \underline{m}$$

可见，高斯分布均值的最大似然估计等于样本的均值。

最大似然估计：高斯分布参数估计

- 似然函数关于参数Σ 的偏导。 $p(\underline{x}_1 \cdots \underline{x}_N|\mu,\Sigma) = \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\underline{x}_i-\underline{\mu})^T\Sigma^{-1}(\underline{x}_i-\underline{\mu})}$

$$\ln p = -\frac{dN}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \sum_{i=1}^{N}[\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}_i-\boldsymbol{\mu})]$$

$$\text{Consider } \boldsymbol{\theta} = \boldsymbol{\Sigma}^{-1}$$

$$\ln p = -\frac{dN}{2}\ln(2\pi) - \frac{N}{2}\ln\frac{1}{|\boldsymbol{\theta}|} - \sum_{i=1}^{N}[\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\theta}(\mathbf{x}_i-\boldsymbol{\mu})]$$

$$\ln p = -\frac{dN}{2}\ln(2\pi) + \frac{N}{2}\ln|\boldsymbol{\theta}| - \sum_{i=1}^{N}[\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu})^T\boldsymbol{\theta}(\mathbf{x}_i-\boldsymbol{\mu})]$$

$$\frac{\partial \ln p}{\partial \boldsymbol{\theta}} = \frac{N}{2}(\boldsymbol{\theta}^{-1})^T - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

$$\frac{N}{2}(\boldsymbol{\theta}^{-1})^T - \frac{1}{2}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T = 0$$

$$(\boldsymbol{\theta}^{-1})^T = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

$$\hat{\boldsymbol{\Sigma}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i-\boldsymbol{\mu})(\mathbf{x}_i-\boldsymbol{\mu})^T$$

可见，高斯分布协方差的最大似然估计等于样本的协方差。

$$|\mathbf{A}^{-1}| = |\mathbf{A}|^{-1}$$

$$\frac{\partial \ln|\mathbf{A}|}{\partial \mathbf{A}} = (\mathbf{A}^{-1})^T$$

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{A}} = \mathbf{x}\mathbf{x}^T$$

$$\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^T$$

$$\boldsymbol{\Sigma} = \boldsymbol{\theta}^{-1}$$

估计偏差与无偏估计

- 基于高斯观测似然的最大似然估计，通过最大化训练样本的观测似然乘积，得到对应的均值和协方差的估计值，结果是样本的均值和协方差。
- 但是，这些参数估计与该高斯分布的真实情况有偏差(bias)吗?

$$\underline{\hat{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i = \underline{m} \qquad \hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\hat{\mu}})(\underline{x}_i - \underline{\hat{\mu}})^T$$

- 定义：如果一个参数估计的期望值是其真值，则该估计称作无偏差估计(unbiased estimates)。
- 无偏差估计意味着只要训练样本个数足够多，该估计值就是参数的真实值。
- 均值的最大似然估计：是无偏差估计！

$$E[\hat{\underline{\mu}}_{ML}] = E\left[\frac{1}{N}\sum_{i=1}^{N}\underline{x}_i\right] = \frac{1}{N}\sum_{i=1}^{N}E[\underline{x}_i] = \frac{1}{N}(N\underline{\mu}) = \underline{\mu}$$

因为期望值是关于样本 Xi 而言，所以 μ 相当于常数项。

补充：期望的定义与运算

- 针对一个随机变量（或向量）X，其期望（expectation, expected value or mean）是指 X 在大量实验情况下得到的平均值(average value)。

$$E[X] = \begin{cases} \sum_i x_i P(x_i) & \text{if } X \text{ is discrete} \\ \int x p(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

加权平均，权重是X取每个值时对应的概率。

- 期望计算属性：

$$\begin{aligned} E[aX + b] &= aE[X] + b \\ E[X + Y] &= E[X] + E[Y] \end{aligned}$$

- 自变量是随机量的函数g(X)期望：

$$E[g(X)] = \begin{cases} \sum_i g(x_i) P(x_i) & \text{if } X \text{ is discrete} \\ \int g(x) p(x) dx & \text{if } X \text{ is continuous} \end{cases}$$

- 假设所有样本x_i是iid(in*dependent and idential distributions*):

$$E[\mathbf{x}_i] = E[\mathbf{x}]$$

估计偏差与无偏估计

- 协方差的最大似然估计是无偏差吗?

$$E[\hat{\boldsymbol{\Sigma}}] = E\left[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left[(\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^T\right]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E\left\{[(\mathbf{x}_i - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})][(\mathbf{x}_i - \boldsymbol{\mu}) - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})]^T\right\}$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T - (\mathbf{x}_i - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T]$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] - E[\frac{1}{N}\sum_{i=1}^{N}(\mathbf{x}_i - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T]$$

$$- E[\frac{1}{N}\sum_{i=1}^{N}(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T$$

$$\text{Since } \frac{1}{N}\sum_{i=1}^{N}\mathbf{x}_i = \hat{\boldsymbol{\mu}}$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T + (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T$$

$$= \frac{1}{N}\sum_{i=1}^{N} E[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^T] - (\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})(\hat{\boldsymbol{\mu}} - \boldsymbol{\mu})^T$$

- 协方差的最大似然估计：有偏差估计!

- 协方差的最大似然估计：偏差是多少？

$$
\begin{aligned}
E[(\hat{\underline{\mu}}-\underline{\mu})(\hat{\underline{\mu}}-\underline{\mu})^T] &= E\left[\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})\frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})^T\right] \\
&= E\left[\frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\underline{x}_i-\underline{\mu})(\underline{x}_j-\underline{\mu})^T\right] \\
&= \frac{1}{N^2}E\left[\sum_{i=1}^{N}(\underline{x}_i-\underline{\mu})(\underline{x}_i-\underline{\mu})^T\right] \\
&= \frac{1}{N}\Sigma
\end{aligned}
$$

假设样本之间是独立分布的。

Since $x_i^m x_j^n = 0$ when $i \neq j$

where $m, n \in [1, d]$.

$$
\sum_i \sum_j \mathbf{x}_i \mathbf{x}_j^T =
$$

$$
\begin{bmatrix}
\sum_i \sum_j x_i^1 x_j^1 & \sum_i \sum_j x_i^1 x_j^2 & \cdots & \sum_i \sum_j x_i^1 x_j^d \\
\sum_i \sum_j x_i^2 x_j^1 & \sum_i \sum_j x_i^2 x_j^2 & \cdots & \sum_i \sum_j x_i^2 x_j^d \\
\vdots & \vdots & \ddots & \vdots \\
\sum_i \sum_j x_i^d x_j^1 & \sum_i \sum_j x_i^d x_j^2 & \cdots & \sum_i \sum_j x_i^d x_j^d
\end{bmatrix}
=
\begin{bmatrix}
\sum_i x_i^1 x_i^1 & \sum_i x_i^1 x_i^2 & \cdots & \sum_i x_i^1 x_i^d \\
\sum_i x_i^2 x_i^1 & \sum_i x_i^2 x_i^2 & \cdots & \sum_i x_i^2 x_i^d \\
\vdots & \vdots & \ddots & \vdots \\
\sum_i x_i^d x_i^1 & \sum_i x_i^d x_i^2 & \cdots & \sum_i x_i^d x_i^d
\end{bmatrix}
$$

- 协方差的最大似然估计：对于单个维度的偏差值

$$var(x_i) = \sigma^2$$

$$var(\frac{1}{N}\sum_{i=1}^{N} x_i) = \frac{1}{N}\sigma^2$$

The variance of the sample mean is the variance divided by the number of samples.

- 协方差的最大似然估计略小于实际的协方差。当N足够大时，可以看到做是一个较好的估计。

$$E[\hat{\Sigma}_{ML}] = \Sigma - \frac{1}{N}\Sigma$$

$$= \frac{N-1}{N}\Sigma$$

- 协方差的无偏差估计可以修正为：训练样本的协方差*N/(N-1)

$$\hat{\Sigma}_u = \frac{N}{N-1}\hat{\Sigma}_{ML} = \frac{1}{N-1}\sum_{i=1}^{N}(\underline{x}_i - \underline{m})(\underline{x}_i - \underline{m})^T$$

最大似然估计：总结

Maximum Likelihood Parameter Estimation

Choose as estimates the values of the parameters that maximize the likelihood (probability) of the observed set of training samples (labelled).

$$\hat{P}_{ML} = \frac{N_1}{N}$$

$$\underline{\hat{\mu}}_{ML} = \frac{1}{N}\sum_{i=1}^{N}\underline{x}_i = \underline{m}$$

$$\hat{\Sigma}_{ML} = \frac{1}{N}\sum_{i=1}^{N}(\underline{x}_i - \underline{\hat{\mu}})(\underline{x}_i - \underline{\hat{\mu}})^T \qquad \text{biased}$$

$$\hat{\Sigma}_u = \frac{N}{N-1}\hat{\Sigma}_{ML} = \frac{1}{N-1}\sum_{i=1}^{N}(\underline{x}_i - \underline{m})(\underline{x}_i - \underline{m})^T \qquad \text{unbiased}$$

贝叶斯估计

- 参数估计的目的是估计先验概率和观测似然分布，从而可以计算后验概率。

$$p(C_i|\mathbf{x}) = \alpha p(\mathbf{x}|C_i)p(C_i)$$

- 参数估计是针对如下情况：先验概率或观测似然的分布形式已知，但其中的参数未知。以观测似然为例，参数估计本质是估计如下条件概率：

$$p(\mathbf{x}|\theta, C_i)$$

- 假设各个类可以单独估计概率密度，则上述基于参数的观测似然写为：

$$p(\mathbf{x}|\theta)$$

- 最大似然估计：将待估计的参数θ当做固定的未知值。
- 贝叶斯估计：将待估计的参数θ当做概率分布，给定其分布的先验概率和训练样本，估计参数θ分布的后验概率。

贝叶斯估计：问题描述

- 两个假设：
  - 针对每个类各自单独估计其概率分布。
  - 各个训练样本之间假设是相互独立的。
- 假设θ 自己服从一个概率分布：
  - 该概率分布的先验概率已知：$p(\theta)$
  - 这个先验概率反映了基于现有知识关于该参数θ最初猜测以及关于该参数的不确定信息。
- 基于C_i类的训练样本，针对θ应用贝叶斯理论，得到其后验概率：

$$p(\theta|\mathcal{D}_i) = \frac{p(\mathcal{D}_i|\theta)p(\theta)}{p(\mathcal{D}_i)}, \text{ where } p(\mathcal{D}_i) = \int p(\mathcal{D}_i|\theta)p(\theta)d\theta$$

- 可见分母的归一化因子跟θ无关。
- 由于样本之间相互独立，可以得到：$p(\theta|\mathcal{D}_i) = p(\theta|\{\mathbf{x}_i\}) = \alpha \prod_{i=1}^{N} p(\mathbf{x}_i|\theta)p(\theta)$

贝叶斯估计：单维高斯参数分布

- 假设每类的观测似然分布是单维高斯分布，且方差σ^2已知，则θ就只是未知参数均值μ。

$$\theta = \mu$$

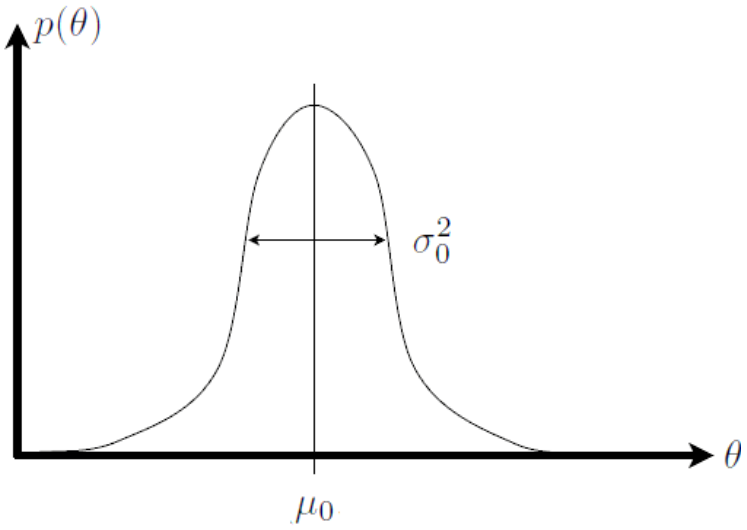
$$p(x|\theta) = N(\theta, \sigma^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2}$$

- 假设θ的先验概率分布也服从单维高斯分布，猜测该分布的均值μ_0和方差σ_0：

$$p(\theta) = N(\mu_0, \sigma_0^2)$$

$$= \frac{1}{\sqrt{2\pi}\sigma_0} e^{-\frac{1}{2}\left(\frac{\theta-\mu_0}{\sigma_0}\right)^2}$$



贝叶斯估计：单维高斯参数分布

- 基于θ的先验概率和该类的训练样本，计算后验概率密度：

$$p(\theta|\{x_i\}) = \alpha \prod_{i=1}^{N} p(x_i|\theta)p(\theta)$$

$$= \alpha \prod_{i=1}^{N} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{x_i-\theta}{\sigma}\right)^2\right]}_{p(x_i|\theta)} \underbrace{\frac{1}{\sqrt{2\pi}\sigma_0} exp\left[-\frac{1}{2}\left(\frac{\theta-\mu_0}{\sigma_0}\right)^2\right]}_{p(\theta)}$$

- 省略与θ无关的项：

$$\exp(a)\exp(b) = \exp(a+b)$$

$$= \alpha' exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\left(\frac{\theta-x_i}{\sigma}\right)^2 + \left(\frac{\theta-\mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' exp\left[-\frac{1}{2}\left[\left(\frac{N}{\sigma^2}+\frac{1}{\sigma_0^2}\right)\theta^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i + \frac{\mu_0}{\sigma_0^2}\right)\theta\right]\right]$$

贝叶斯估计：单维高斯参数分布

$$= \alpha' exp\left[-\frac{1}{2}\left(\sum_{i=1}^{N}\left(\frac{\theta - x_i}{\sigma}\right)^2 + \left(\frac{\theta - \mu_0}{\sigma_0}\right)^2\right)\right]$$

$$= \alpha'' exp\left[-\frac{1}{2}\left[\left(\frac{N}{\sigma^2} + \frac{1}{\sigma_0^2}\right)\theta^2 - 2\left(\frac{1}{\sigma^2}\sum_{i=1}^{N} x_i + \frac{\theta}{\sigma_0^2}\right)\theta\right]\right]$$

- 可见，该公式是二次exp函数，可以把θ的后验概率写作高斯分布的形式：

$$p(\theta|\{x_i\}) = \frac{1}{\sqrt{2\pi}\sigma_N} exp\left[-\frac{1}{2}\left(\frac{\theta - \mu_N}{\sigma_N}\right)^2\right]$$

- 对比上述两个公式，可以得到：

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_N}{\sigma_N^2} = \frac{N}{\sigma^2}m + \frac{\mu_0}{\sigma_0^2}$$

m是样本均值

贝叶斯估计：单维高斯参数分布

$$\frac{1}{\sigma_N^2} = \frac{N}{\sigma^2} + \frac{1}{\sigma_0^2} \qquad \frac{\mu_N}{\sigma_N^2} = \frac{N}{\sigma^2} m + \frac{\mu_0}{\sigma_0^2}$$

m是样本均值

- 由上述两个公式解出μ_N 和σ_N ：

$$\mu_N = \frac{N\sigma_0^2 m + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} \qquad \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}$$

- 可见，给定该类的N个样本，参数θ概率分布的均值等于样本均值和该参数先验概率均值的加权和。
- 给定该类的N个样本，参数θ概率分布的方差是由该类观测似然分布的方差、该参数的先验概率方差、样本个数共同决定。

贝叶斯估计：单维高斯参数分布

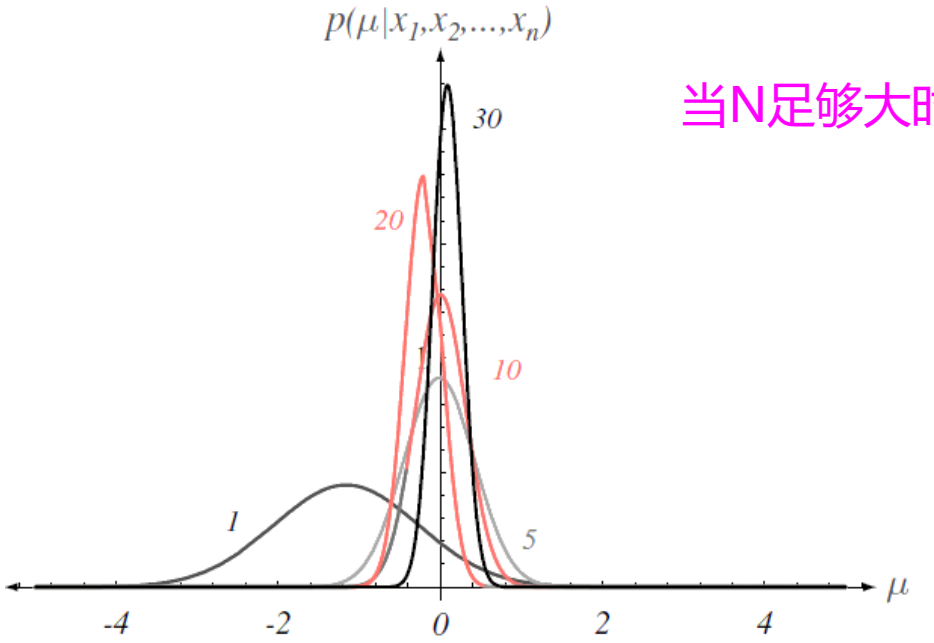
- 当该类的训练样本个数N非常大时，会发生什么？

$$\lim_{N\to\infty} \mu_N = m$$

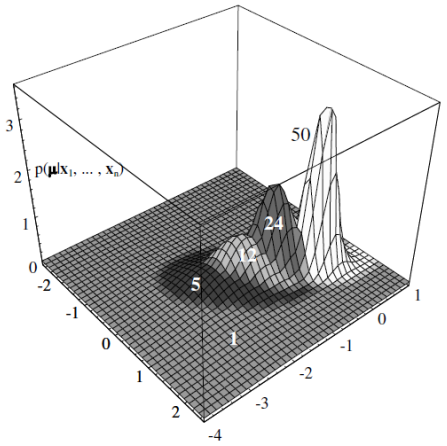
$$\lim_{N\to\infty} \sigma_N^2 = 0$$

$$\mu_N = \frac{N\sigma_0^2 m + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2}$$

$$\sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}$$



当N足够大时，样本均值就是均值参数θ的无偏估计。



贝叶斯估计：单维高斯参数分布

- 贝叶斯学习：贝叶斯估计具备不断学习的能力，它允许最初的、基于少量训练样本的、不太准的估计，随着训练样本的不断增加，可以串行的不断修正参数的估计值，从而达到该参数的期望真值。

$$\mu_N = \frac{N\sigma_0^2 m + \sigma^2 \mu_0}{N\sigma_0^2 + \sigma^2} \qquad \sigma_N^2 = \frac{\sigma_0^2 \sigma^2}{N\sigma_0^2 + \sigma^2}$$

- 极值情况：
  - 如果参数的先验方差$\sigma_0 = 0$，则$\mu_N \to \mu_0$，意味着先验的确定性会强大，使得后续训练样本的不断进入也不太可能对参数估计有太多改变。
  - 如果参数的先验方差$\sigma_0 \gg \sigma$，则$\mu_N \to m$，意味着先验的确定性会非常小，使得刚开始的参数估计不准，因为训练样本个数太少。

贝叶斯估计：Class-conditional Density

- 最后，根据贝叶斯估计得到的参数后验概率分布，计算该类的class-conditional density，即观测似然与参数联合分布的边际（marginal）概率：

$$p(\mathbf{x}|\mathcal{D}_i) = \int_\theta p(\mathbf{x}|\theta, \mathcal{D}_i)d\theta = \int_\theta p(\mathbf{x}|\theta)p(\theta|\{\mathbf{x}_i\})d\theta$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \frac{1}{\sqrt{2\pi}\sigma_n} \exp\left[-\frac{1}{2}\left(\frac{\mu-\mu_N}{\sigma_N}\right)^2\right] d\mu$$

$$= \frac{1}{2\pi\sigma\sigma_N} \exp\left[-\frac{1}{2}\frac{(x-\mu_N)^2}{\sigma^2+\sigma_N^2}\right] f(\sigma, \sigma_N)$$

$$\text{where } f(\sigma, \sigma_N) = \int \exp\left[-\frac{1}{2}\frac{\sigma^2+\sigma_N^2}{\sigma^2\sigma_N^2}\left(\mu - \frac{\sigma_N^2 x + \sigma^2\mu_N}{\sigma^2\sigma_N^2}\right)^2\right]$$

- 可见，观测似然的边缘（marginal）概率可以看做是高斯分布：

$$p(\mathbf{x}|\mathcal{D}_i) = \mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$$

贝叶斯估计：观测似然分布的最终估计

- 贝叶斯估计是估计观测似然的边缘概率，即考虑观测似然分布和其参数分布的联合概率，把参数看做参数空间的一个概率分布：

$$p(\mathbf{x}|\mathcal{D}_i) = \mathcal{N}(\mu_N, \sigma^2 + \sigma_N^2)$$

- 最大似然估计只是估计给定参数情况下的观测似然的分布，不考虑参数的分布情况，只是把参数看做参数空间的一个固定的点：

$$p(\mathbf{x}|\theta, \mathcal{D}_i) = \mathcal{N}(\mu, \sigma^2)$$

- μ_N 随着样本个数的逐渐增大，趋近与真实均值；在已知的方差σ上加入σ_N，代表对于未知均值μ的不确定性。
- 样本个数逐渐增大时，贝叶斯估计越来越能代表真实的观测似然分布：

$$p(\mathbf{x}|\mathcal{D}_i) \to p(\mathbf{x}|C_i, \mathcal{D}_i)$$

贝叶斯估计：总结

- 贝叶斯估计（学习）：不仅仅只是估计参数的分布，最终估计是的观测似然和参数的联合分布的边缘概率。
- 给定量：观测似然分布的形式、参数的先验概率、训练样本。
- 贝叶斯估计的步骤：
  - 1.估计参数的后验概率：

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int p(\mathcal{D}|\boldsymbol{\theta})p(\boldsymbol{\theta})\ d\boldsymbol{\theta}} \qquad p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{k=1}^{n} p(\mathbf{x}_k|\boldsymbol{\theta}).$$

  - 2.估计观测似然的边缘概率：

$$p(\mathbf{x}|\mathcal{D}) = \int p(\mathbf{x}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathcal{D})\ d\boldsymbol{\theta}$$