# 课程1-2

## 损失函数与统计量之间关系的几个证明

【课上讲授】

假定损失函数，其中为参数的真实值，为参数的估计值。一般来说无法知道，但是可以根据样本集合计算出的后验分布，于是使用极小化的后验期望的方法来估计，即





命题（1）当是平方函数时，是的后验均值，即

命题（2）当是绝对值函数（1-范数）时，是的中位数（median），即

命题（3）当是dirac函数（函数）时，是的众数（mode），即

小知识：

随机变量的中位数的定义：若随机变量的某个取值满足：



则是随机变量的中位数（median）。注意：累积分布函数（CDF, cumulative distribution function）只具有右连续性。

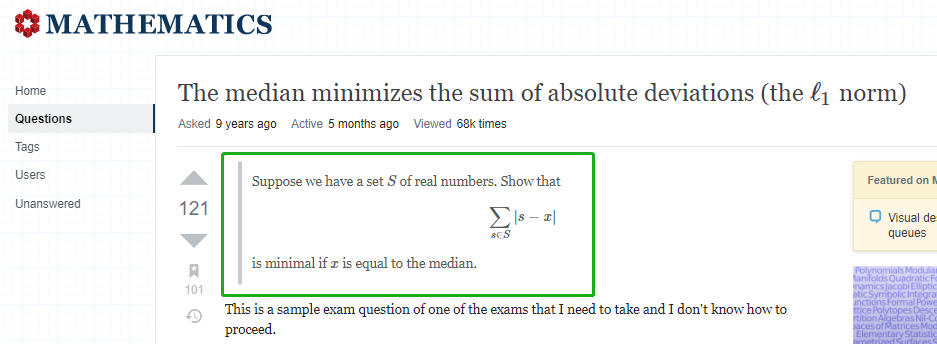
【课后资料】

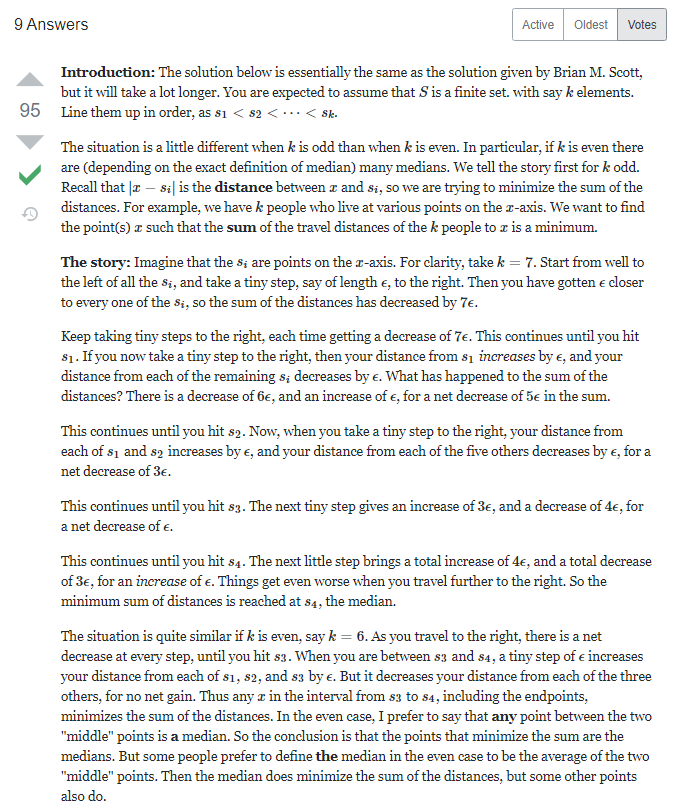
上述命题（1）注意到其实就是在求关于某个值的最小“方差”，该表达式的最小值就是的方差，取到最小值时的值就是的期望。

上述命题（2）的一个比较清晰易懂的证明：

<https://math.stackexchange.com/questions/113270/the-median-minimizes-the-sum-of-absolute-deviations-the-ell-1-norm>

截图如下：





上述命题（3）易证。

## 隐马尔科夫模型的（新）三个问题

【课上讲授】

已知马尔可夫链上的个顺序观测值，求模型的状态，按照时间与的大小关系，可以分为以下三个问题：

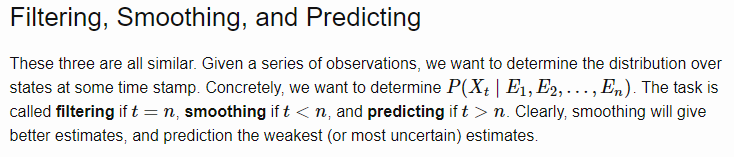
1. ，预测问题（prediction）
2. ，滤波问题（filtering）
3. ，平滑问题（smoothing）

【课后资料】

资料链接：

<https://danieltakeshi.github.io/2015-07-25-hidden-markov-models-and-particle-filtering/>

资料截图：



## 平方损失函数下的极大后验估计是一个最优估计子

【课上讲授】

HMM的状态在平方损失函数下的极大后验估计是一个最优估计子（Optimal Estimator），原因是：

1. 它是随机变量一个无偏估计；

证明：

1. 它是方差最小的估计。

证明：

【课下资料】

目前并未找到Optimal Estimator的严格定义，但是一些Optimal Estimation相关的资料显示，“无偏”和“最小方差”并不是构成Optimal Estimator的必要条件。

## 平方损失函数下的两个结论

本节将证明：在损失下，数学期望是随机变量的最优常数拟合，而条件数学期望则是Y的最优函数拟合。即：

1. 
2. 

注意这里的的表达式是：



### 4.1最优常数拟合

**证明**：

首先将（1）展开



利用方差的性质：，其中为随机变量的方差，得到



注意到上式最右侧不等号中的等号成立的充要条件是：，因此



**得证**。

### 4.2最优函数拟合

**证明**：

首先将（2）展开：







注意到：上式中的中的第一项是与无关的，因此在中可以消去此项，于是：



注意上式等号右侧中包含以下两部分：

 和 

接下来证明对于上面的第一部分，有。

注意到与随机变量无关，因此：









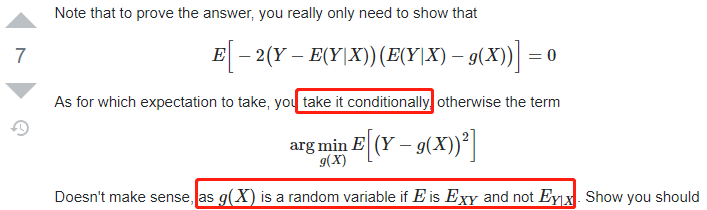


**得证**。

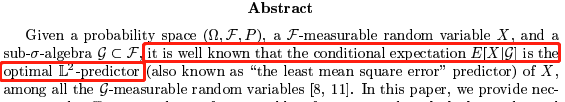
### 4.3说明

（1）本节介绍的结论的意义是，在以平方损失误差（条件）期望最小为目标时，随机变量关于随机变量（参数）的最优估计（拟合函数）表达式就是自身的条件期望。这相当于给出了一个回归问题机器学习的模型。后面还会证明，这个结论可以被推广到每一个属于Bregman Divergence（Bregman散度）的损失函数上。

（2）不少文献中将最有函数拟合式写为：，即没有强调等号右侧的期望实际上是条件期望，这一点在本节参考资料[1]中有说明。



（3）本节所介绍的两个结论式统计学中的重要结论，见参考资料[3]。



### 4.4参考资料

1. Problem with proof of Conditional expectation as best predictor, 超链接：<https://stats.stackexchange.com/questions/71863/problem-with-proof-of-conditional-expectation-as-best-predictor>
2. 期望和条件期望, 知乎, 超链接：<https://zhuanlan.zhihu.com/p/110871306>
3. On the Optimality of Conditional Expectation as a Bregman Predictor, Arindam Banerjee等.

## 推广平方损失下的估计结论

本节将证明，上一节中介绍的结论：实际上可以从平方损失函数推广到所有满足Bregman Divergence的损失函数上。

### Bregman散度

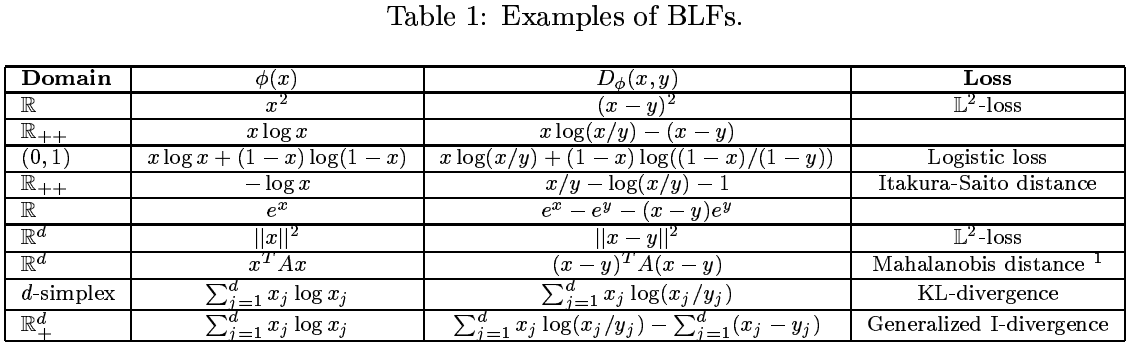
以下的定义1和定义2均来自参考资料[1]。

**定义1（Bregman Divergence）**. 设是一个定义在凸集上的严格凸函数，如此也就在上可微。对任意两点，关于的Bregman散度（Bregman Divergence）定义为：

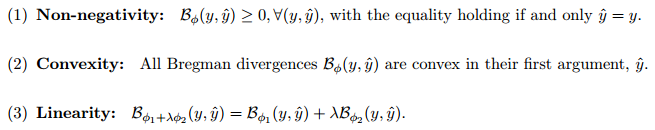
.

其中表示向量与向量的内积。

**定义2（Bregman Loss Function）**. 如果损失函数满足Bregman散度，则损失函数可以被称为Bregaman Loss Function（BLF）。典型的BLF包括（来自参考资料[2]）：



**性质.**



### 结论推广

以下的定理1和定理2均来自参考资料[1]，本节中不再给出证明。

**定理1**. 设是一个严格凸的可微函数且是对应的Bregman散度。那么对于的所有函数，条件期望就是Bregman损失函数期望的唯一**最小子**（unique minimizer）

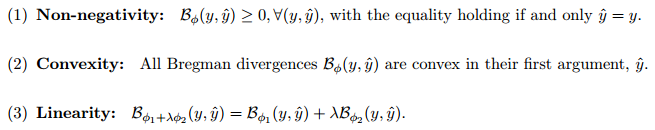
.

**定理2**. 设是非负函数。如果对多有的随机变量，是的唯一**最小子**（unique minimizer），对于所有均成立，那么一定存在严格凸的可微函数使得是的Bergman散度。

【说明】

定理1是前一节中介绍的平方损失函数下最优估计结论的推广；

定理2一般被看做是一个理论上更惊人的结论，在参考资料[3]，这里仅给出截图：



### 5.3参考资料

1. The Big Picture: Loss Functions at the Dataset Level, Karthik Kannan, 硕士论文.

[2] On the Optimality of Conditional Expectation as a Bregman Predictor, Arindam Banerjee等.

[3] 如何理解Bregman divergence？, 知乎回答, 超链接：<https://www.zhihu.com/question/22426561/answer/209945856>

# 课程3-4

## 完全贝叶斯

【课上讲授】

最大后验估计（maximum a posteriori estimation，MAP）与完全贝叶斯估计（Full Bayesian Estimation）是不同的，完全贝叶斯估计不是直接取使后验概率最大的参数作为估计结果，而是基于参数的后验分布对参数进行平滑（关于参数的积分），通过积分消去原始模型中的参数项，得到新的模型估计表达式。

【课后资料】

将课上老师讲到的内容进行展开解释。一般统计学习教材种都会介绍以下三种与概率相关的参数估计方法：

1. 极大似然估计（ML）；
2. 极大后验估计（MAP）；
3. （完全）贝叶斯估计（Bayesian Estimation）。

以上三类估计方法的基础模型都是：若待估计量是，那么给定包含参数的概率模型，和观测集合。

### 极大似然估计

极大似然估计中首先定义样本集合上的似然函数：

: 

（1）极大似然估计的参数计算（学习、训练）公式：



（2）极大似然估计训练后的模型：



### 1.2极大后验估计

在ML方法中，上式的被看作是带估计的参数（常量），而在MAP和Bayesian Estimation种会将参数看作是随机变量，为此还需要引入的先验分布：

: 

极大后验估计的参数估计公式使用了贝叶斯公式，如下：

（1）极大后验估计的参数计算公式：



需要特别强调注意的是，上式推导的依据是：右侧分母部分是与参数无关的，所以求时可以不考虑分母。

为了计算简便，经常对求对数，这同样不影响的结果，如下：



（2）极大后验估计训练后的模型：



（3）总结

a.) MAP与Bayesian一样，都是将看作随机变量，这与ML种将看作常量不同；

b.) MAP在理论描述时虽然形式上需要依靠贝叶斯公式计算参数后验分布，但实际上由于参数的极大后验估计只需要知道，因此实际上无需严格按照贝叶斯公式进行计算（特别是无需计算贝叶斯公式的分母）。这一点是MAP与Bayesian最重要的区别。

### 1.3贝叶斯估计

本小节介绍的Bayesian Estimation也被称为完全贝叶斯方法（Full Bayesian），它和MAP一样将看作随机变量，但Bayesian Estimation不直接估计参数的值，而是先计算的后验分布，之后通过在后验分布上的“平滑”（smooth，积分、期望）计算分布。

（1）计算的后验分布：



（2）基于后验分布的“平滑”：



将上面的两步骤公式合并，得到



Bayesian Estimation与MAP最重要的不同是需要依据贝叶斯公式计算出的解析形式，而这种计算往往实际是困难的，因为贝叶斯公式中的分母是一个积分：



当随机变量是高维向量时，上面的计分时难于计算的，一般只能使用近似算法求解。

### 1.4 参考资料

[1] ML, MAP, and Bayesian — The Holy Trinity of Parameter Estimation and Data Prediction, Avinash Kak, Purdue University, 2017.

## Batch vs Recursive

先通过MDP（Markov Decision Process）滤波（filtering）问题的递归解法（感觉这里相当于增量、online），介绍递归算法的感性认识，之后介绍本节课的主旨：对比批量（batch, 矩阵）算法和递归（recursive, 增量）算法。连续举了三个例子，每个例子都先后使用批量再用增量算法。

### MDP滤波

待补充

### 2.2带高斯噪声的线性回归

【问题】

对于包含高斯噪声的线性回归问题，面向过程的增量描述如下：对于逐一到来的样本，假设满足：

，，，，

随机变量和分别服从：

，

使用Bayesian Estimation估计



【分析】

Bayesian Estimation估计的基本思路是先计算，之后利用





得到预测分布（Predictive distribution），进行预测。这里需要注意，上面计算预测分布的公式中，应该把看作是在已知（给定）的数据上对参数的后验，的定义为



批量（batch）算法和增量（recursive）算法在计算预测分布上是相同的，二者的区别集中在计算后验的方式上。关于批量算法的计算推导过程可以参考[1]（PRML, Pattern Recognition and Machine Learning）中的3.1.5，接下来重点描述根据板书整理出的增量算法计算思路。

考虑到线性回归问题中将视作常量而非随机变量，同时为了和课堂板书推到表达形式保持一致，下面将省略中的，简记为。

假设已知，那么根据贝叶斯公式有：







注意到上面推到结果中，第1个因子中，其实与无关，因此



再对上式等号右侧的概率使用贝叶斯公式



用的展开代替展开式中第1项因子，消去后得到





上面推导的第2步利用了与随机变量无关这一特性。注意到上式中的表达式实际上是已知的（利用题干中、且与相互独立得出），即



综上，证明了可以基于已知的（和）计算出。是用数学归纳法，不仅可以证明增量算法是可行的，而且实质上给出了增量算法计算的一个构造性证明。

增量算法最终计算得到的表达式与批量算法一致，具体计算过程不做展开。此外，增量方法得到的递推方程组的形式

与最小均值平方算法（LMS, Least Mean Square Algorithm）存在一定的关联。

### 2.3 卡尔曼滤波

卡尔曼滤波（Kalman Filtering），待补充。

### 2.4参考资料

1. Pattern Recognition and Machine Learning, Bishop, 2006.
2. Introduction to recursive Bayesian filtering, Michael Rubinstein, PPT.