# 2.数学基础

## 贝叶斯主义与频率主义

概率论始终围绕着模型(参数)和案例展开，概率通常指某个事件发生的可能性。

频率主义认为：**概率及其模型(参数)是真实确定存在的，而事件本身是随机的；**因此，可以通过**最大似然估计**参数的值。比如，在随机试验过程中，我们把某事件发生的比例或频率作为该事件发生的概率。

贝叶斯主义则持完全不同的观点，他们认为：**真实出现的事件是一种确定性的存在，而模型及其参数反而是不确定的随机变量**。在分析过程中，总是先假设一个**先验的概率分布**，随着样本的增加，不断的修正先验的概率分布。

频率主义描述为事件在概率参数下出现的可能，通过最大似然估计参数，可通过**经验风险最小化**(ERM)描述

$$\theta = \arg\min_{\theta} L(X,\theta)$$

贝叶斯主义描述为概率参数在已有事件下的可信度，通过先验分布结合实际出现的事件修正概率参数，可通过**结构风险最小化**(SRM)描述。

$$\theta = \arg\min_{\theta}[L(X,\theta) + \lambda J(\theta)]$$

实际上就是将经验风险最小化**正则化**，防止了过拟合。常用的方法是**最大后验概率估计**(MAP)。

$$\theta = \arg\max_{\theta}[\log p(x|\theta) + \log p(\theta)]$$

第一项对应对数似然估计，第二项对应先验分布。

**经验风险最小化是无偏估计，结构风险最小化是有偏估计。**

在样本非常多时，两种方法的效果差不多，但当样本非常有限时，实践表明：贝叶斯方法总能取得较好的效果，通过添加先验分布实际上降低了系统的复杂度。

### 范数

# Vector Norms (范数)

- A norm is a function $\|\cdot\|$ that satisfies:
  - $\|\mathbf{x}\| \geq 0$ with equality if and only if $\mathbf{x} = 0$
  - $\|\mathbf{x}+\mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$
  - $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$

Can generally think of as the vector "**length**"

**Examples**: consider a vector x in $\mathbb{R}^n$ with elements $x = (x_1, x_2, \ldots, x_n)$. Then:

- The $l_2$ norm, or Euclidean distance:

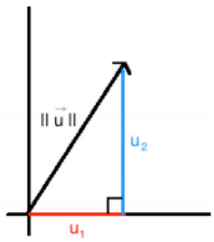
$$\|x\|_2 = \|x\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

- The $l_1$ norm:

$$\|x\|_1 = \sqrt{|x_1| + |x_2| + \cdots + |x_n|}$$

- The $l_p$-norm:

$$\|x\|_p = \sqrt[p]{|x|_1^p + |x|_2^p + \cdots + |x|_n^p}$$




https://zhuanlan.zhihu.com/p/26884695

Xiaodong Gu Machine Learning: Chapter 2

## 贝叶斯公式

通常，事件A在事件B(发生)的条件下的概率，与事件B在事件A的条件下的概率是不一样的；然而，这两者是有确定的关系,贝叶斯法则就是这种关系的陈述。

作为一个规范的原理，贝叶斯法则对于所有概率的解释是有效的；然而，频率主义者和贝叶斯主义者对于在应用中概率如何被赋值有着不同的看法：频率主义者根据随机事件发生的频率，或者总体样本里面的个数来赋值概率；贝叶斯主义者要根据未知的命题来赋值概率。一个结果就是，贝叶斯主义者有更多的机会使用贝叶斯法则。

贝叶斯法则是关于随机事件A和B的条件概率和边缘概率的。

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_j P(B|A_j)P(A_j)}$$

其中P(A|B)是在B发生的情况下A发生的可能性。

$$A_1, \cdots, A_n$$

为完备事件组，即

$$\cup_{i=1}^n A_i = \Omega, A_i A_j = \phi, P(A_i) > 0.$$

在贝叶斯法则中，每个名词都有约定俗成的名称：

Pr(A)是A的先验概率或边缘概率。之所以称为"先验"是因为它不考虑任何B方面的因素。

Pr(A|B)是已知B发生后A的条件概率，也由于得自B的取值而被称作A的后验概率。

Pr(B|A)是已知A发生后B的条件概率，也由于得自A的取值而被称作B的后验概率。

Pr(B)是B的先验概率或边缘概率，也作标准化常量（normalized constant）。

**按这些术语，Bayes法则可表述为：**

后验概率 = (似然度 * 先验概率)/标准化常量　也就是说，后验概率与先验概率和似然度的乘积成正比。

另外，比例Pr(B|A)/Pr(B)也有时被称作标准似然度（standardised likelihood），Bayes法则可表述为：

后验概率 = 标准似然度 * 先验概率。

## 梯度

设二元函数

$$z = f\left(x, y\right)$$

在平面区域D上具有一阶连续偏导数，则对于每一个点P（x，y）都可定出一个向量

$$\left\{\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right\} = f_x\left(x, y\right)\bar{i} + f_y\left(x, y\right)\bar{j}$$

,该函数就称为函数

$$z = f\left(x, y\right)$$

在点P（x，y）的梯度，记作gradf（x，y）或

$$\nabla f\left(x, y\right)$$

,即有：

gradf（x，y）=

$$\nabla f\left(x, y\right)$$

=

$$\left\{\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right\} = f_x\left(x, y\right)\bar{i} + f_y\left(x, y\right)\bar{j}$$

其中

$$\nabla = \frac{\partial}{\partial x}\bar{i} + \frac{\partial}{\partial y}\bar{j}$$

称为（二维的）向量[微分算子]或Nabla算子，

$$\nabla f = \frac{\partial f}{\partial x}\bar{i} + \frac{\partial f}{\partial y}\bar{j}$$

。

设

$$e = \{cos\alpha, \cos\beta\}$$

是方向l上的单位向量，则

$$\frac{\partial f}{\partial l} = \frac{\partial f}{\partial x}\cos\alpha + \frac{\partial f}{\partial y}\cos\beta = \left\{\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}\right\}\{\cos\alpha, \cos\beta\}$$

$$= gradf\left(x, y\right)e = \left|gradf\left(x, y\right)\right||e|\cos[gradf\left(x, y\right), e]$$

由于当方向l与梯度方向一致时，有

$$\cos[gradf\left(x,y\right),e]=1$$

所以当l与梯度方向一致时,方向导数

$$\frac{\partial f}{\partial l}$$

有最大值，且最大值为梯度的模，即

$$|gradf\left(x,y\right)|=\sqrt{\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2}$$

因此说，函数在一点沿梯度方向的变化率最大，最大值为该梯度的模。