# 机器学习算法相关记录

|  |  |
| --- | --- |
| 作者： | 刘炜 |
| 版本： | 0.1 |
| 时间： | 2016-07-26 |

# 名义变量与哑变量（dummy）

<http://baike.baidu.com/link?url=ftGEAhvFdPI-BInCblAadwwkZtUddjtIqFfQ2wMpeGZp4TKmsMlhgGlkkOBpQPmh2t74PbjNq4CqcFLwxVALeq>

<http://baike.baidu.com/view/2039878.htm?fromtitle=%E5%93%91%E5%8F%98%E9%87%8F&fromid=9621990&type=syn>

<http://www.dxy.cn/bbs/topic/26148597?ppg=1>

分类变量，大部分分类变量是描述变量

多分类变量又可分为有序（等级）或无序（也叫名义）分类变量

如果是有序（ordinal）分类变量，一般可按对因变量影响由小到大的顺序编码为1、2、3、……，或者按数据的自然大小，将它当作连续变量处理。如果是无序的（nomial）分类变量，则需要采用哑变量（dummy variables）进行编码。假如职业分类为工、农、商、学、兵5类，则可定义比分类数少1个，即5-1=4个哑变量。

# 统计学中RR、OR、AR、HR的区别

<http://www.medsci.cn/article/show_article.do?id=f22e1920083>

RR也叫危险比（risk ratio）或率比（rate ratio），是反映暴露与发病（死亡）关联强度的最有用的指标。RR适用于队列研究或随机对照试验。

AR归因危险度，又叫特异危险度、率差（rate difference, RD）和超额危险度（excess risk），是暴露组发病率与对照组发病率相差的绝对值，它表示危险特异地归因于暴露因素的程度。

OR比值比（Odds ratio）：也称优势比、比数比、交叉乘积比，是病例对照研究中表示暴露与疾病之间关联强度的指标，比值（odds）是指某事物发生的概率与不发生的概率之比。比值比指病例组中暴露人数与非暴露人数的比值除以对照组中暴露人数与非暴露人数的比值。

HR风险比（hazard ratio）［HR=暴露组的风险函数h1(t)/非暴露组的风险函数h2(t),t指在相同的时间点上］。而风险函数指危险率函数、条件死亡率、瞬时死亡率。Cox比例风险模型可以得到HR。资料的类型通常是临床治疗性研究，也可以是流行病学的队列观察性研究。

风险比和危害比：

1、多数认为HR与RR意思一样，但HR有时间因素在内，换句话说，包含了时间效应的RR就是HR；

2、可以这样理解，生存资料中，RR考虑了终点事件的差异，而HR不仅考虑了终点事件的有无，还考虑了到达终点所用的时间及截尾数据；

3、根据Parmar的文章，HR＝（Or/Oe）/(Cr/Ce)，O和C分别代表对照组和试验组，r代表实际发生事件的人数，e代表按log rank test中计算得出的理论上应该发生事件的人数。其实HR的计算还有一种方法就是

\*\*\* this is assumed based on the constancy of the ratio over time, as above.

对于提供生存率的文献，可以根据公式HR=[p0/(1-p0)]/[p1/(1-p1)],其中p0为对照组的5年生存率，p1为暴露组的5年生存率。

\*\*\* Hazard is an instaneous rate of a given outcome, which indeed factors time in. Hazard ratio is a raio of two hazards (Exp/Ctl) at that given time. By using hazard ratio as a summary estimate of the treatment effects in suvival data, we assume the hazard ratio is constant over time, which is a strong assumption and a basis for cox regression. This doesn't mean hazard would not change (actually, hazard always change over time), though - we just require the ratio of the hazards is constant.

4、根据Parmar的文章，HR＝（Or/Oe）/(Cr/Ce)，O和C分别代表对照组和试验组，r代表实际发生事件的人数，e代表按log rank test中计算得出的理论上应该发生事件的人数。

# 线性回归

<http://wenku.baidu.com/view/dc54f91cb7360b4c2e3f64a1.html?re=view> \*\*\*

<http://wenku.baidu.com/link?url=fuiGkRMnEMit-QSacQOdpq_XxS8aM-WQo9d_EEHpRfre22IhUUvhyaR2tLRfq7Zm9Rk6cuIEoyRn3i1Owcdb28D_V7t8BSag49aP9lDQdcW>

<http://wenku.baidu.com/link?url=fuiGkRMnEMit-QSacQOdpq_XxS8aM-WQo9d_EEHpRfre22IhUUvhyaR2tLRfq7Zm9Rk6cuIEoyRn3i1Owcdb28D_V7t8BSag49aP9lDQdcW>

<http://wenku.baidu.com/view/26d15d03bed5b9f3f90f1ccb.html?re=view>

## 估计方法

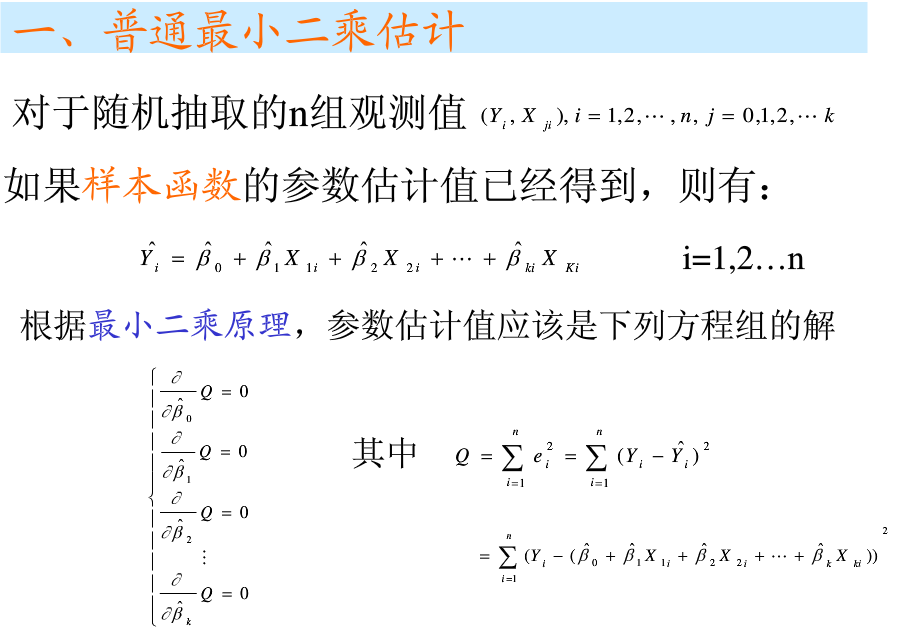
<http://blog.163.com/loongfee@126/blog/static/173697334201302211527594/>

OLS：Oridinary least squares 普通最小二乘估计

ML：Maximum likelihood 最大似然估计

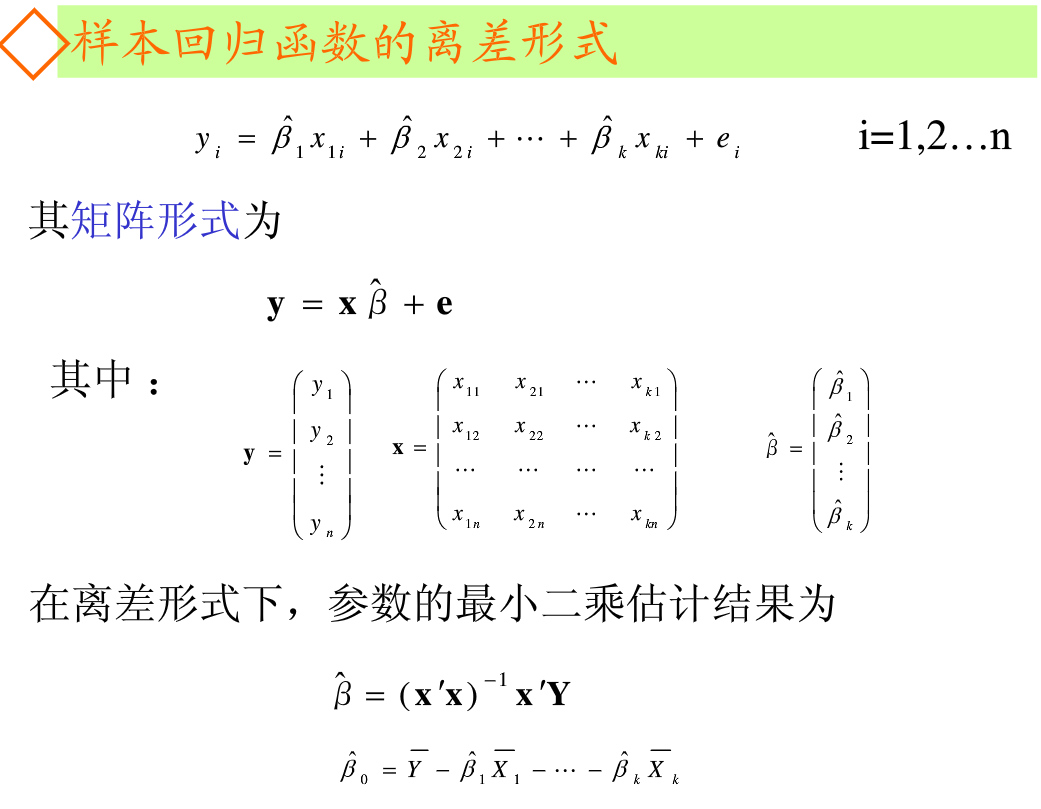
MM：Moment method 矩估计

## OLS



注意这里Q是实际值与预测值的误差平方和，这里看作关于的方程，即最终拟合曲线方程的参数。

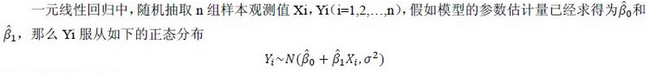
要使拟合曲线更贴近实际，误差平方和就要最小，于是对Q关于进行求偏导数，根据实际场景知道存在对的偏导为零使得误差达到最小，从而求出的估计值，进而得到拟合方程。



此方法要求X是列满秩的，而且求矩阵的逆比较慢。若X不是列满秩则求不出来拟合曲线的所有参数，当然这表示其中参数是相关联的，可以消去，或者说相应参数在式子中没有起到作用，可以通过重新确定参数也求解。

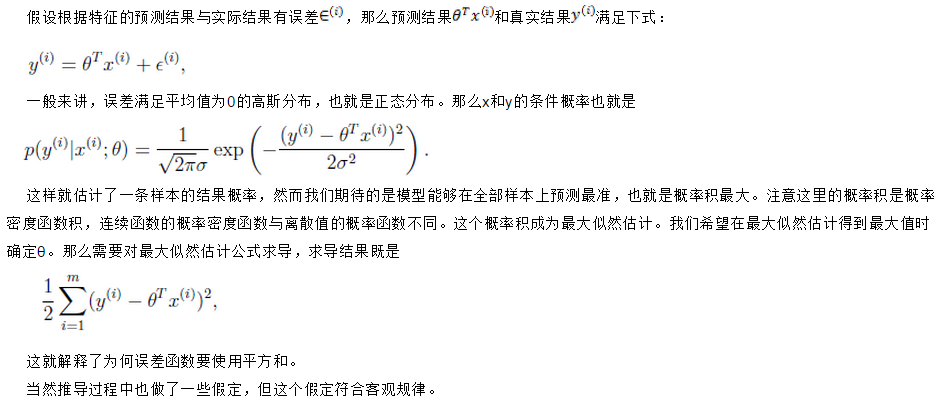
<http://blog.csdn.net/xidianzhimeng/article/details/20847289>

最大似然估计和最小二乘估计的区别和联系



残差服从的正态分布，而

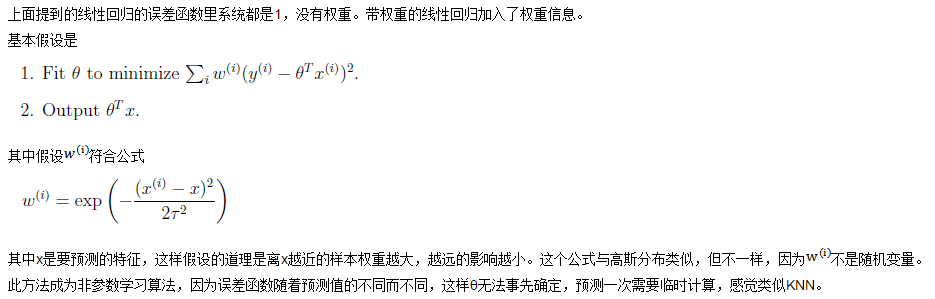
## 误差函数为平方和的概率解释



<http://blog.csdn.net/statdm/article/details/7585153>

假设误差满足平均值为0的正态分布，这也服从使误差最小化以达到贴近实际的目的。

## 带权重的线性回归



# Logistic regression(Logistic 回归)

<http://scau200630760309.blog.163.com/blog/static/191217862201210147521204/>



表示在所有特征均有的情况下发生的情况。

## 阶跃函数

<http://blog.sina.com.cn/s/blog_455c7a600101ngt1.html>

<https://en.wikipedia.org/wiki/Heaviside_step_function>

赫维赛德阶跃函数( Heaviside step function ) ，单位阶跃函数。图像为阶梯状，可对事物进行分类，值域为0-1之间的离散值，定义域为(-INF,INF)。

## Sigmoid函数

<http://blog.csdn.net/statdm/article/details/7585153>

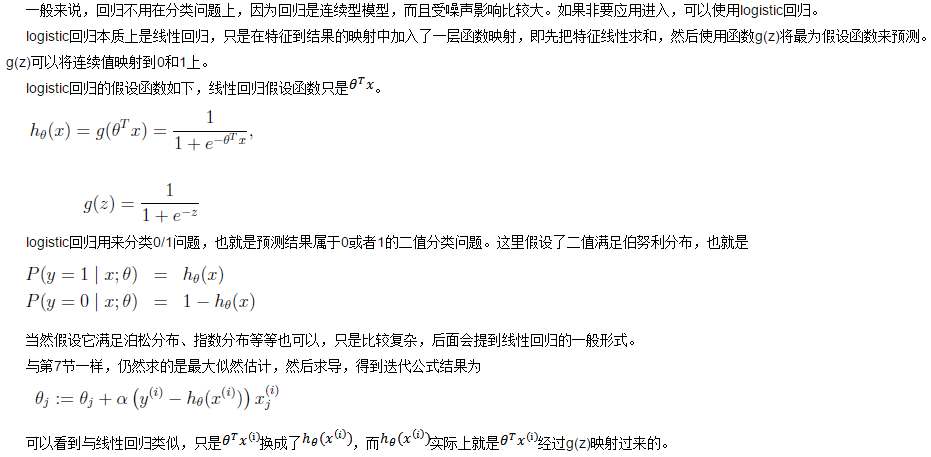
西格马函数，即S型函数，被用于Logistic regression，因此也叫Logistic函数。基本形式为：



值域为(0,1)，定义域为(-INF,INF)，由于其函数曲线特性与正态分布的概率曲线特性类似，因此可以模拟事件发生概率。又由于其阶跃性，所以容易对事物进行分类。

一般来说，回归不用在分类问题上，因为回归是连续型模型，而且受噪声影响比较大。如果非要应用进入，可以使用logistic回归，因为其阶跃性，可以较好区别事物。

logistic回归本质上是线性回归，只是在特征到结果的映射中加入了一层Sigmoid函数映射，即先把特征线性求和，然后使用函数作为假设函数来预测。



得到迭代公式。

## 求解

<http://blog.csdn.net/ariessurfer/article/details/41310525>

<http://www.cnblogs.com/fengbing/archive/2013/05/18/3086324.html>

<http://m.blog.csdn.net/article/details?id=51179381>

牛顿法Newton-Raphson method

梯度下降法：沿着梯度方向(相应导数方向)，要设定步长。

<http://www.cnblogs.com/fengbing/archive/2013/05/15/3079033.html>

# 贝叶斯分类

## 思想

根据以往经验，将个体样本归属于它最有可能的分类当中。比如路上遇到个人会讲成都话，那么根据你的经验判断他很有可能就是成都人，于是你认为他是个成都人。

## 贝叶斯定理



## 病人分类例子

让我从一个例子开始讲起，你会看到贝叶斯分类器很好懂，一点都不难。某个医院早上收了六个门诊病人，如下：

　　症状　　职业　　　疾病

　　打喷嚏　护士　　　感冒

　　打喷嚏　农夫　　　过敏

　　头痛　　建筑工人　脑震荡

　　头痛　　建筑工人　感冒

　　打喷嚏　教师　　　感冒

头痛　　教师　　　脑震荡

现在又来了第七个病人，是一个打喷嚏的建筑工人。请问他患上感冒的概率有多大？

根据贝叶斯定理可得：

　　　P(感冒|打喷嚏x建筑工人)

　　　　= P(打喷嚏x建筑工人|感冒) x P(感冒)

　　　　/ P(打喷嚏x建筑工人)

假定"打喷嚏"和"建筑工人"这两个特征是独立的，因此，上面的等式就变成了：

　　　P(感冒|打喷嚏x建筑工人)

　　　　= P(打喷嚏|感冒) x P(建筑工人|感冒) x P(感冒)

　　　　/ P(打喷嚏) x P(建筑工人)

这是可以计算的。

　　 P(感冒|打喷嚏x建筑工人)

　　　　= 0.66 x 0.33 x 0.5 / 0.5 x 0.33

　　　　= 0.66

因此，这个打喷嚏的建筑工人，有66%的概率是得了感冒。同理，可以计算这个病人患上过敏或脑震荡的概率。比较这几个概率，就可以知道他最可能得什么病。

这就是贝叶斯分类器的基本方法：在统计资料的基础上，依据某些特征，计算各个类别的概率，从而实现分类。

## 贝叶斯分类

<http://www.ruanyifeng.com/blog/2013/12/naive_bayes_classifier.html>

<http://www.cnblogs.com/leoo2sk/archive/2010/09/17/naive-bayesian-classifier.html>

假设某个体有n项特征（Feature），分别为，即。现有m个类别（Category），分别为。贝叶斯分类器就是计算出对该个体概率最大的那个分类，也就是求下面这个算式的最大值：



则：该个体属于

由于 对于所有的类别都是相同的，可以省略，问题就变成了求的最大值。

## 朴素贝叶斯分类

假设所有特征都彼此独立，因此：



上式等号右边的每一项，都可以从统计资料中得到，由此就可以计算出每个类别对应的概率，从而找出最大概率的那个类。

## 对应例子

本例摘自张洋的《算法杂货铺----分类算法之朴素贝叶斯分类》。

<http://www.cnblogs.com/leoo2sk/archive/2010/09/17/1829190.html>

根据某社区网站的抽样统计，该站10000个账号中有89%为真实账号（设为），11%为虚假账号（设为）。





接下来，就要用统计资料判断一个账号的真实性。假定某一个账号有以下三个特征：

: 日志数量/注册天数

: 好友数量/注册天数

: 是否使用真实头像（真实头像为1，非真实头像为0）

若该帐号统计结果：　　　　　　　

请问该账号是真实账号还是虚假账号？

方法是使用朴素贝叶斯分类器，计算下面这个计算式的值。



虽然等式右边这些值可以从统计资料得到，但是这里有一个问题：和是**连续变量**，不适宜按照某个特定值计算概率。

一个技巧是将连续值变为离散值，计算区间的概率。比如将分解成[0, 0.05]、(0.05, 0.2)、[0.2, +∞]三个区间，然后计算每个区间的概率。在我们这个例子中，等于0.1，落在第二个区间，所以计算的时候，就**使用第二个区间的发生概率**。

根据统计资料，可得：



因此：

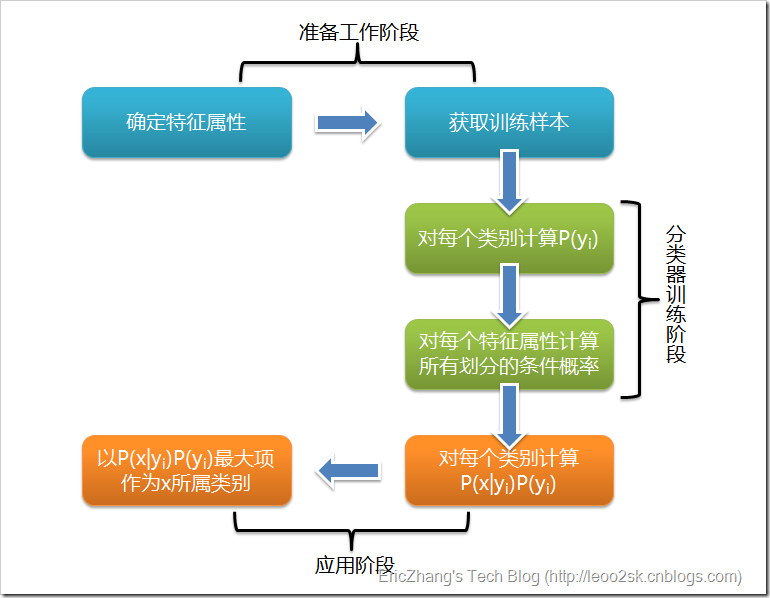






可以看到，虽然这个用户没有使用真实头像，但是他是真实账号的概率，比虚假账号高出30多倍，选择概率最大的分类作为其分类 ，因此判断这个账号为真，。

## 朴素贝叶斯分类流程



**第一阶段**——据具体情况确定特征属性，并对每个特征属性进行适当划分，然后由人工对一部分待分类项进行分类，形成训练样本集合。

输入：待分类数据，

输出：特征属性和训练样本。

人工完成，质量对整个过程将有重要影响，分类器的质量很大程度上由特征属性、特征属性划分及训练样本质量决定。

**第二阶段**——分类器训练阶段，这个阶段的任务就是生成分类器，主要工作是计算每个类别在训练样本中的出现频率及每个特征属性划分对每个类别的条件概率估计，并将结果记录。

输入：特征属性和训练样本

输出：分类器（经验积累，之后用于判断）。

**第三阶段**——应用阶段。使用分类器对待分类 项进行分类。

输入：分类器和待分类项

输出：待分类项与类别的映射关系。

# 模糊集fuzzy set

对象有某些模糊的概念而无法明确该对象对集合的隶属关系。

用来表达模糊性概念的集合。 又称模糊集[1] 、模糊子集。普通的集合是指具有某种属性的对象的全体。这种属性所表达的概念应该是清晰的，界限分明的。因此每个对象对于集合的隶属关系也是明确的，非此即彼。但在人们的思维中还有着许多模糊的概念，例如年轻、很大、暖和、傍晚等，这些概念所描述的对象属性不能简单地用“是”或“否”来回答，模糊集合就是指具有某个模糊概念所描述的属性的对象的全体。由于概念本身不是清晰的、界限分明的，因而对象对集合的隶属关系也不是明确的、非此即彼的。这一概念是美国加利福尼亚大学控制论专家L.A.扎德于 1965 年首先提出的。模糊集合这一概念的出现使得数学的思维和方法可以用于处理模糊性现象，从而构成了模糊集合论（中国通常称为模糊性数学)的基础。

# 数据清洗

OpenRefine：之前由google提供支持，免费开源数据清洗工具，帮助用户转换数据集的工具，优化数据的质量以便于在真实场景中使用。可以观察和操纵数据的工具。它类似于传统Excel的表格处理软件，但是工作方式更像是数据库，以列和字段的方式工作，而不是以单元格的方式工作。

将调数据清洗，可视化并不突出。

<http://selection.datavisualization.ch/>

# 可视化

<http://inspire.blufra.me/big-data-visualization-review-of-the-20-best-tools/>

## 数据可视化工具：零编程

Data Visualization Tools for Presentations: Zero Coding Required

### Tableau -- Desktop and Online(商业)

### Infogram(商业)

### ChartBlocks -- Online

### Datawrapper(商业)

### Plotly (还需要详细了解)

### RAW

### Visual.ly(商业)

## Javascript 库：需要编程

Data Visualization Tools for Developers: JavaScript libraries

### D3.js

### Sigma JS

### Polymaps

### Other很多