1.PCA降维的公式及原理分析:

假设对称矩阵A的所有特征值都不一样，那么: -> 对角矩阵

对于矩阵Y，它的协方差

假设Y=QX（此时不考虑降维），则 = =

PCA的本质:让协方差最小，方差最大 -> 协方差矩阵的对角元最大，非对角元为0->对角矩阵满足

我们已经得到两个公式: 和 ，假设Q =

可以得到: = ，因为协方差矩阵是对称半正定矩阵(特征值>=0)

所以可以得到:当Q = 时，是对角矩阵，因此如果Q少取几行就实现了降维

2.极大似然估计公式及原理分析:

极大似然估计起源于贝叶斯

假设特征为D，标签为A，P为概率

贝叶斯公式为:P(A|D) =

当样本给定的时候P(D)为常量，同时假定先验概率P(A)近似相等，那么我们可以得到:

P(A|D) = P(D|A)

对于机器学习的意义：将已知特征求标签的概率转化为已知标签求特征的概率，从无监督问题变成了有监督问题

但是极大似然的最大问题是:不同参数的P(A)不一定是相等的，只能说在样本不够大的时候是近似相等的

贝叶斯算法考虑了先验概率P(A),但同时要求先验概率要准确

极大似然的两个假设：

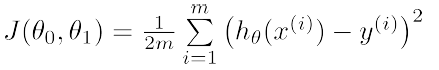
1. 已经发生的事件是独立重复事件，符合同一分布
2. 已经发生的事件是可能性（似然）最大的事件

3.线性回归

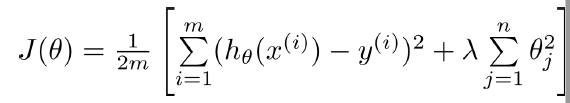
公式:

h(x) =

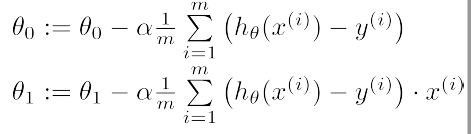
损失函数:



加上L2正则化后：



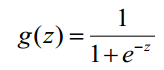
梯度下降:分别对求导，



4.逻辑回归

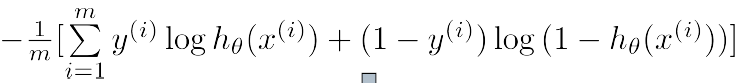
LR实际上是sigmoid+线性回归，用于判断边界，起到分类的作用

sigmoid公式：

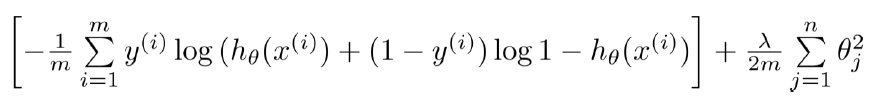


如果输入z = 线性回归，那么就是LR

损失函数:



损失函数+正则化项:



逻辑回归使用实例:

5.贝叶斯

公式：

P(A|D) =

推导:

X,Y同时发生的概率即联合概率P(X,Y) = P(Y|X)P(X) = P(X|Y)P(Y)->P(Y|X) = （贝叶斯）

朴素贝叶斯(NB):条件独立性假设，很“朴素”的近似，即:

P("邱佳豪"|S)=P（“邱”|S）P（“佳”|S）P（“豪”|S）

以垃圾邮件为例:

利用分词向量将“我司发票”拆成一个个分词

C1=P（“垃圾邮件”|“我”，“司”，“发票”） = P(“我”，“司”，“发票”|“垃圾邮件”)P(“垃圾邮件”) | P（“我”，“司”，“发票”）

C2=P ("正常邮件" | “我”，“司”，“发票”) = P（“我”，“司”，“发票”|"正常邮件"）P（"正常邮件"）| P(“我”，“司”，“发票”)

C1/C2可以把分母约掉，再加上独立性假设

C1/C2 = P（“我”|S）P（“司”|S）P（“发票”|S）P（S）除以P（“我”|H）P（“司”|H）P（“发票”|H）P（H），其中H，S分别代表垃圾邮件和正常邮件

这样就化简后，每一项就特别好求，比如

P（“发票”|“垃圾邮件”） = 垃圾邮件中出现发票的次数/垃圾邮件中所有词谱出现的次数总和

由于乘法的交换率， P（“我”|S）P（“司”|S）和 P（“司”|S） P（“我”|S）是一样的，这样一来，贝叶斯就失去了语言里的顺序问题

面试必问:1.三种模型 2.平滑 3.工程上的灵活应用

P(（"代开普通发票增值税发票”）|S)这个拆开后，发票出现了2次

如果统计与判断的时候都计算重复的次数，称为多项式模型，即P（发票）\*P（发票）

如果训练和判断都只算一次，称为伯努利模型，即P（发票）

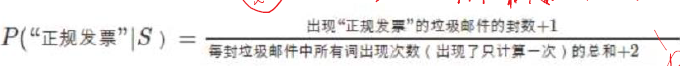
如果训练的时候考虑重复的，判断的时候不考虑，即高斯分布型（假定属性/特征是服从正态分布的）

工程中比理想情况更加复杂，这个时候要考虑平滑：如何在样本数量不是那么充足的情况下，给未出现的新词一个概率，就是平滑最重要的事情

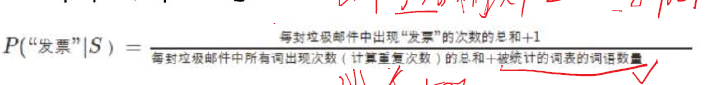
平滑的本质:统一给未出现的词一个概率，同时也要减少已知的概率（所有分子相加等于分母）

如果真实情况中出现了一个测试中没有出现的词就会导致某一项P=0，导致整个结果为0

对于伯努利模型:（简称拉普拉斯平滑）

因为考虑重复所以是封数，分母加2（2效果比1好，未知的概率增加，已知的概率就要减少）

对于多项式模型:（加上被统计词谱的数量，实际上是降低已知词的概率）



多项式模型平滑举例:

训练集1.(a,b,c,a,a) 2.(a,b,b) 3.(b,c,c) 测试集:4.(d,a,c)

P(d)=(0+1)/(11+4)=1/15,其中11是训练集中所有词出现次数的总和，4代表有a，b，c，d四个词

P(a) =(4+1)/15=5/15，4是训练集出现了4次，旧的已有的次数也要加1

P(b) =(4+1)/15 = 5/15

P(c) = (3+1)/15 = 4/15

总和的概率为1

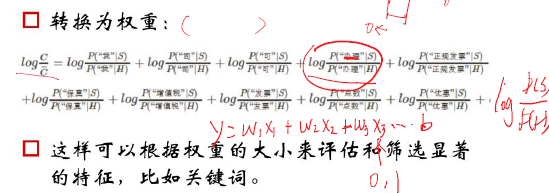
工程上的一些技巧:

1. 取对数

把乘法变成加法，运算速度提高

1. 取权重

这样把两张表变成了一张表，测试的时候只需要取出权重做加法就行



这样一转化之后，你会发现每一个log其实就相当于一个wx，而最后的logP(S)/P(H)就相当于b，于逻辑回归相似

面试题:NB与LR的差异

1. 最核心的区别是:LR是Loss最优化求出的权重，而朴素贝叶斯是基于统计，跳过了Loss最优化，直接得出权重
2. NB比LR多了一个条件独立假设
3. LR是判别模型，NB是生成模型

朴素贝叶斯注意点

1.很多特征是连续数值型的，但是它们不一定服从正态分布，一定要想办法把它们变换调整成满足正态分布！！

2.对测试数据中的0频次项，一定要记得平滑，简单一点可以用『拉普拉斯平滑』。

3.先处理处理特征，把相关特征去掉，因为高相关度的2个特征在模型中相当于发挥了2次作用。

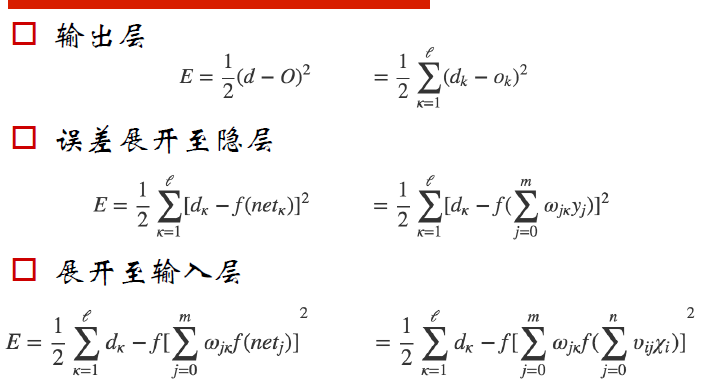
4.一般其他的模型(像logistic regression，SVM等)做完之后，我们都可以尝试一下bagging和boosting等融合增强方法。这对朴素贝叶斯里这些方法都没啥用。原因是这些融合方法本质上是减少过拟合，减少variance的。朴素贝叶斯是没有variance可以减小。

6.神经网络之BP算法

正向传播求损失，反向传播求误差

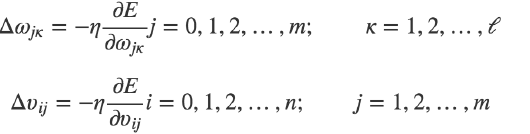
核心:链式求导法则

损失函数:d（真实的结果）与0（输出）都是概率，两者越接近，交叉熵E越小，（实际运算时会记录中间的运算结果，所以不会这么繁琐）



梯度下降:（因为损失函数不是凸函数，存在很多局部变量，如果采用所有样本计算，一个是效率低，一个是容易陷入局部最优）

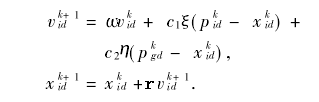
SGD,随机梯度下降，一般的梯度下降需要求出所有样本当前应该朝着哪个方向走，随机梯度下降只是找一部分样本来计算朝哪个方向走



7.神经网络之粒子群优化算法

所以的鸟都不知道食物在哪，但是知道自己当前离食物的举例，最有效的策略就是寻找离食物最近的鸟的附件区域。

下面给出粒子群算法的位置速度更新公式：



http://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_w.bmp是保持原来速度的系数，所以叫做惯性权重。http://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_c1.bmp是粒子跟踪自己历史最优值的权重系数，它表示粒子自身的认识，所以叫“认知”。通常设置为2。http://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_c2.bmp是粒子跟踪群体最优值的权重系数，它表示粒子对整个群体知识的认识，所以叫做“社会知识”，经常叫做“社会”。通常设置为2。http://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_e.bmphttp://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_n.bmp是[0,1]区间内均匀分布的随机数。http://p.blog.csdn.net/images/p_blog_csdn_net/niuyongjie/295595/o_r.bmp是对位置更新的时候，在速度前面加的一个系数，这个系数我们叫做约束因子。通常设置为1。