Web data mining部分：

第三章讨论部分：

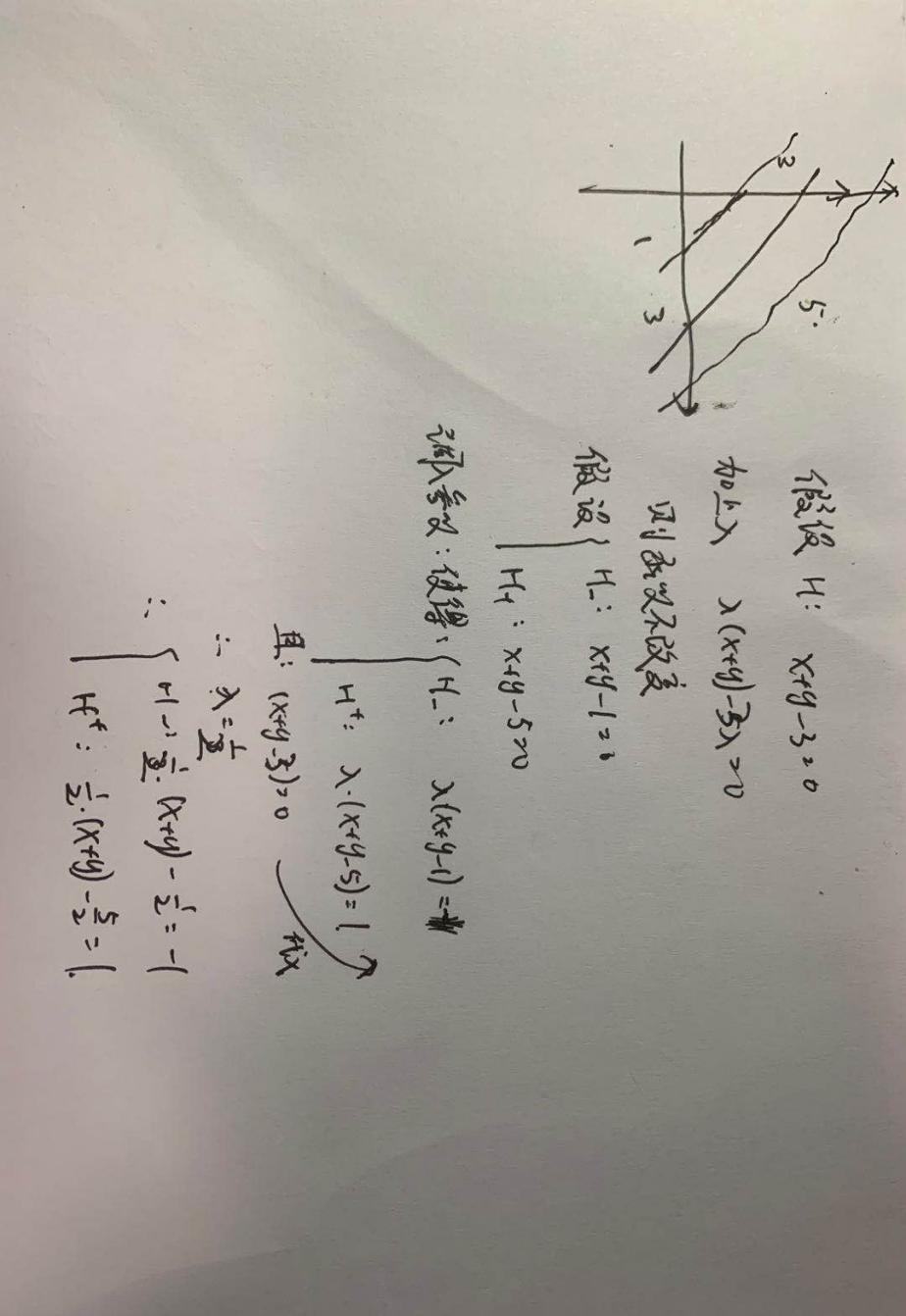
读书报告内容：

1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：

p83的第三行，通过加入参数，我们可以调节超平面，并且不改变函数，如何理解，参数为何一定是正实数？

讨论后的理解：

因为可以通过加入lamda调整超平面的位置左右调整，而不用改别函数，因为此处的函数lamda是可以消掉的，lamda一定是正数可以保证约束条件在是不等式的情况下，不等式不会变号。我实际进行手动模拟如下：



1. 提出的问题2：

拉格朗日乘数法如何用数学定理证明，有何实际意义

讨论后的理解：

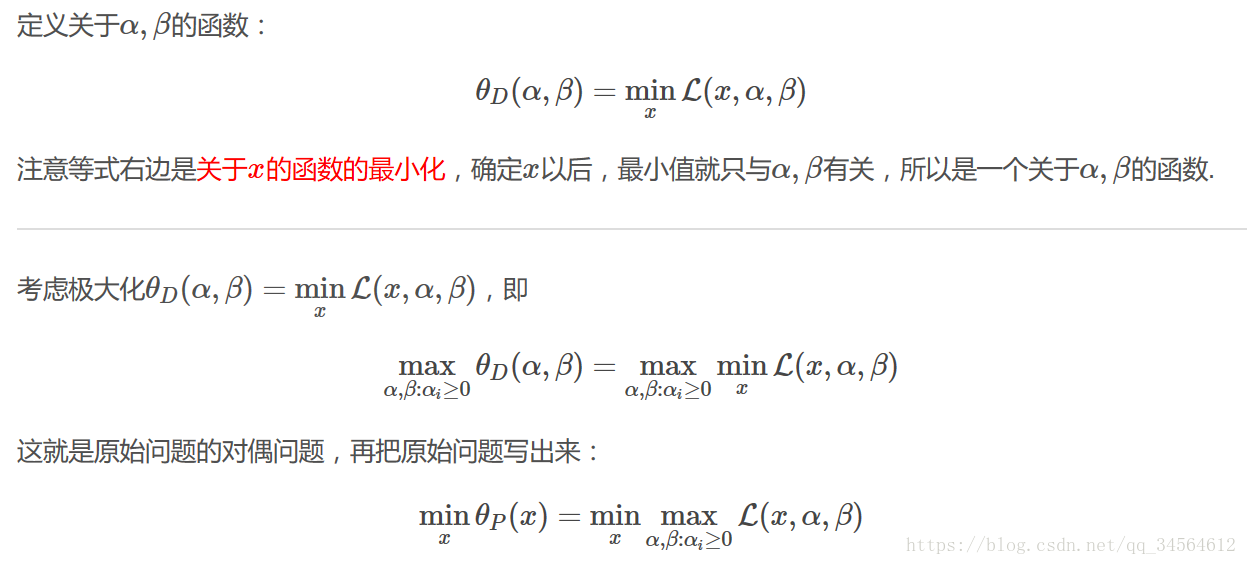
拉格朗日乘数法其实在高数计算中就有应用。我看了大量的资料整理了详细推导过程，因篇幅太长，在第四部分中具体呈现。

1. 提出的问题3：

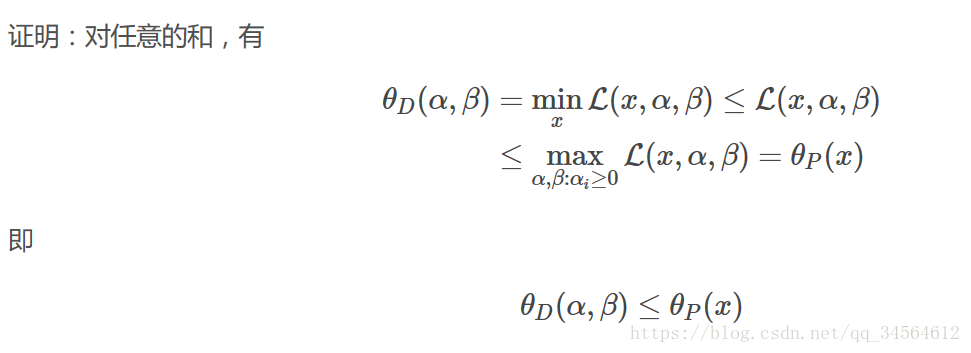
对偶问题那里理解有点问题？

讨论后的理解：

交换极大极小的顺序（其实就是交换代入求导的顺序），转化为拉格朗日极大极小问题：



不妨设对偶问题解为d\*。



求解对偶问题，先极小，对x求导，得出x关于α，β的表达式，再代入f(x)，再求极大，对α，β求导。又因为有严格证明保证当f(x)和c(x)为凸函数，h(x)为仿射函数时，p\*= d\*，所以可以转化对称问题求解。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题4：SVM的核方法把输入样本映射到高维的特征空间，在高维空间中还能保证算法的有效性吗？怎么评估这个误差是不是合理？

自己的理解：

我们希望样本在特征空间内线性可分，因此特征空间的好坏对支持向量机的性能至关重要。但是，在不知道特征映射的形式时，我们并不知道什么样的核函数是合适的，而核函数也仅是隐式地定义了这个特征空间。于是，核函数选择成了支持向量机的最大变数。若核函数选择不合适，则意味着将样本映射到了一个不合适的特征空间，很可能导致性能不佳。

而且在现实任务中，即使找到了某个核函数使训练集在特征空间中线性可分，也很难断定这个貌似线性可分的结果是不是由于过拟合造成的。

1. 问题5：P77拉普拉斯修正是否有可能改变原先的预测结果？

自己的理解：首先拉普拉斯修正可以有效的避免因训练样本不足而导致概率估值为零的问题，并且在训练集变大的时候，修正所引入的先验的影响也会逐渐变得可忽略，使得估值趋向于实际概率值

并且因为拉普拉斯引入的分母n是类别c的个数，使得修正较为平滑，所以拉普拉斯修正是比较优秀的。

1. 问题6：p83式41下面说使用超平面上的任一点与边缘超平面的距离来作为d+，而不是x+到超平面的距离，为什么会这样计算？两种方式有什么本质上的区别？如果没有的话书上为什么会强调这一点？

自己的理解：

因为支持向量是边缘超平面上的点，最后在求边缘超平面的时候不仅仅是使用一个支持向量，而是使用全部支持向量的平均值，这样求出来的平均值不一定在原本的边缘超平面上，这样边缘超平面可能会变化，这样求出的平均的点可能不存在，而且数值上两种方法求出来的结果是一样的。

1. （必填）读书计划
2. 本周完成的内容章节：

（1）3.6到3.8完成并且梳理知识点，寻找问题，自己思考并且于小组会之前完成了自己的思考。

（2）为配合第三章的阅读，同时参考并且阅读了西瓜书的第六章，第七章，参考了视频《白板推导》和视频《一起啃书》理解了svm和贝叶斯方法背后的数学原理，并且放到一起综合学习

2、下周计划：3.9-3.10的阅读.

四、（选做）读书摘要及理解或伪代码的具体实现（读书摘要、伪代码的具体实现代码等可以写到这个部分）

1.代码实现

基于朴素贝叶斯的文本 分类：

def loadDataSet():

postingList=[['my', 'dog', 'has', 'flea', 'problems', 'help', 'please'],

['maybe', 'not', 'take', 'him', 'to', 'dog', 'park', 'stupid'],

['my', 'dalmation', 'is', 'so', 'cute', 'I', 'love', 'him'],

['stop', 'posting', 'stupid', 'worthless', 'garbage'],

['mr', 'licks', 'ate', 'my', 'steak', 'how', 'to', 'stop', 'him'],

['quit', 'buying', 'worthless', 'dog', 'food', 'stupid']]

classVec = [0,1,0,1,0,1]#1 侮辱性文字 ， 0 代表正常言论

return postingList,classVec

def createVocabList(dataSet):#创建词汇表

vocabSet = set([])

for document in dataSet:

vocabSet = vocabSet | set(document) #创建并集

return list(vocabSet)

def bagOfWord2VecMN(vocabList,inputSet):#根据词汇表，讲句子转化为向量

returnVec = [0]\*len(vocabList)

for word in inputSet:

if word in vocabList:

returnVec[vocabList.index(word)] += 1

return returnVec

def trainNB0(trainMatrix,trainCategory):

numTrainDocs = len(trainMatrix)

numWords = len(trainMatrix[0])

pAbusive = sum(trainCategory)/float(numTrainDocs)

p0Num = ones(numWords);p1Num = ones(numWords)#计算频数初始化为1

p0Denom = 2.0;p1Denom = 2.0 #即拉普拉斯平滑

for i in range(numTrainDocs):

if trainCategory[i]==1:

p1Num += trainMatrix[i]

p1Denom += sum(trainMatrix[i])

else:

p0Num += trainMatrix[i]

p0Denom += sum(trainMatrix[i])

p1Vect = log(p1Num/p1Denom)

p0Vect = log(p0Num/p0Denom)

return p0Vect,p1Vect,pAbusive#返回各类对应特征的条件概率向量

#和各类的先验概率

def classifyNB(vec2Classify,p0Vec,p1Vec,pClass1):

p1 = sum(vec2Classify \* p1Vec) + log(pClass1)#注意

p0 = sum(vec2Classify \* p0Vec) + log(1-pClass1)#注意

if p1 > p0:

return 1

else:

return 0

def testingNB():

listOPosts,listClasses = loadDataSet()#加载数据

myVocabList = createVocabList(listOPosts)#建立词汇表

trainMat = []

for postinDoc in listOPosts:

trainMat.append(setOfWords2Vec(myVocabList,postinDoc))

p0V,p1V,pAb = trainNB0(trainMat,listClasses)#训练

#测试

testEntry = ['love','my','dalmation']

thisDoc = setOfWords2Vec(myVocabList,testEntry)

print testEntry,'classified as: ',classifyNB(thisDoc,p0V,p1V,pAb)

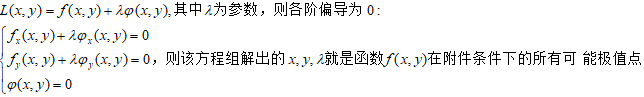
1、读书摘要及理解（选做）。

拉格朗日乘数法原理证明：

参考知乎文章：马同学高等数学《如何理解拉格朗日乘子法》，下述为自己理解后复盘。

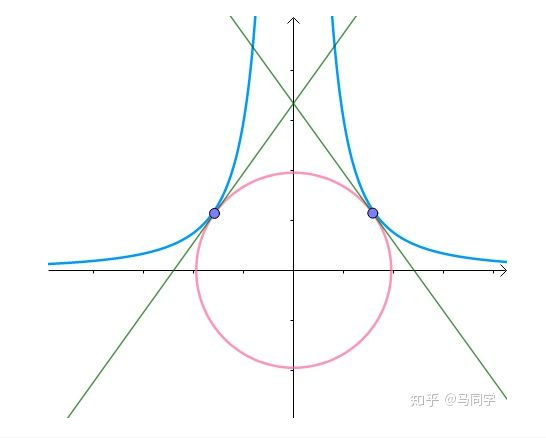
高数中学习的拉格朗日乘数法可以通过引入新的未知标量（拉格朗日乘数 [公式] ），直接求多元函数条件极值，不必先把问题转化为无条件极值的问题。

求函数f(x,y)在附加条件 [公式] 下可能的极值点，可以先做拉格朗日函数：



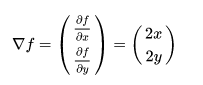
那么首先我们从二维进行分析

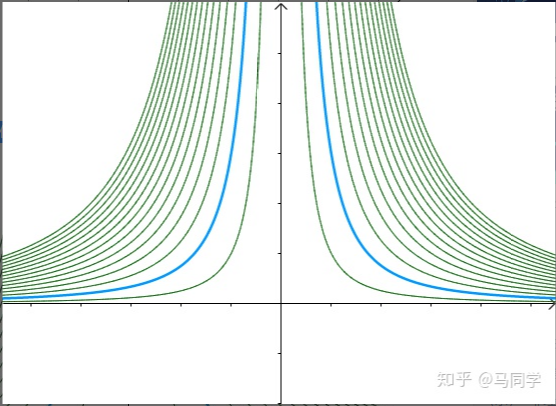
如图所示：求在约束条件x2\*y=3的情况下，求x2+y2的最值问题



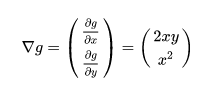
显然有：在极值点，圆与曲线相切。

下面我们引入等高线的概念，根据梯度的性质，梯度向量：

是等高线的法线，而另一个函数，g（x，y）=x2\*y 的等高线为下图所示，之前的曲线x2\*y=3就是其中值为3的等高线：



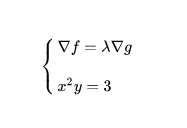
因此，必有这俩梯度向量相等，则为：



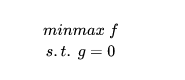
又因为有确切的证明可推导，梯度与等高线的切线垂直（可数学推导，运用求偏导的方法，此处忽略，高等数学知识）。

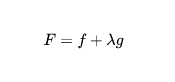
所以，其必有梯度向量平行，用数学符号表示为：1595823569(1)

故有联立方程：

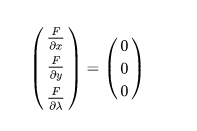


现在我们回头看拉格朗日乘子法的定义：



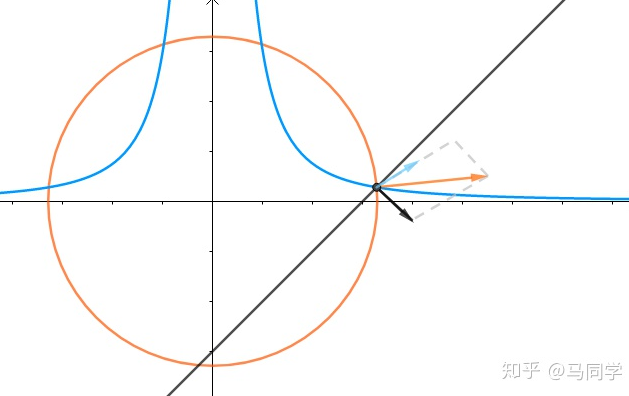
那么我们定义：

在对F的x，y，lamda求偏导，则有：

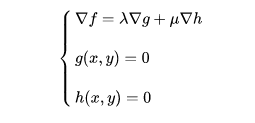


此处豁然开朗，突然理解了为什么之前学习高数，要构造拉格朗日F，再进行求偏导，之前学习只知其然不知其所以然，此时终于得其原理，感觉，数学真是妙不可言。

下面我们对拉格朗日乘子进行扩充增加多个约束条件，那么图为：



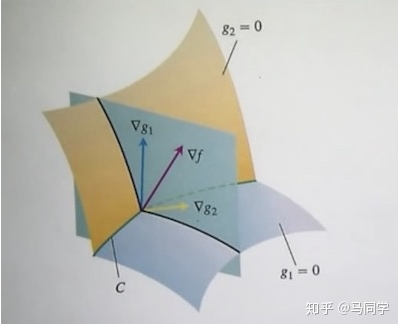
可得联立方程：



即可求解

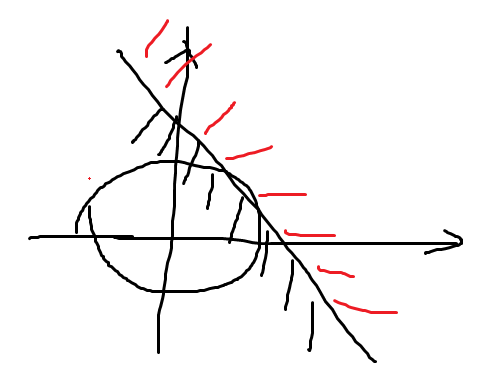
如果从高维向量高维超平面的角度讲直观上看

迪尔塔f必然在迪尔塔g1和迪尔塔g2形成的张开空间中，如图所示：



但是马同学原文这里就结束了，实际上，这只是在等式的情况下，而不等式的情况下，必然要考虑到lamda的梯度和不等式的梯度反向的情况下，其必有hj《=0，lamda》=0

这显然和我们所认知的支持向量在边界上所符合，在二维空间可表示为：

很显然，在向量同向的时候，即为图中的黑阴影部分，边缘边界上的点并不能作为支持向量，因为最小值点必定在原点（0，0），而反向时，即为红色阴影部分，才有作为支持向量。则有必要不充分条件lamda\*hj=0，这就是KTT条件里面的第三个约束条件。

综上，即为所证。

知识图谱部分：

**知识图谱自己完成了第五章的阅读。**

学习方式：

第一遍完成阅读，有不理解的地方圈画作标记，圈画自己认为重点做标记。

第二遍再次阅读，完成知识框架的构建。

第三遍回顾解决问题。

完成内容：

1实践了之前的sdtype代码

下周计划：

1继续完成至少一章的阅读

2 阅读论文DBpedia A的type inference部分。