方骏-2020年4月19日-读书报告

1. 自己提出的问题的理解：
2. 提出的问题1：合页损失函数证明那里：

https://qqadapt.qpic.cn/txdocpic/0/02b4d5db6b56b1be63dd025a768566f5/0?w=554&h=63

这是怎么证明得到的？

讨论后的理解：这里的ζ是大于等于0的，那么就说明应该是，因此书上的应该是错误的。这里书上的意思应该是说，对于正确分类的点来说，是确实正确的，我们也应该用这种条件作为约束条件，这也符号原先最优化模型的约束条件。

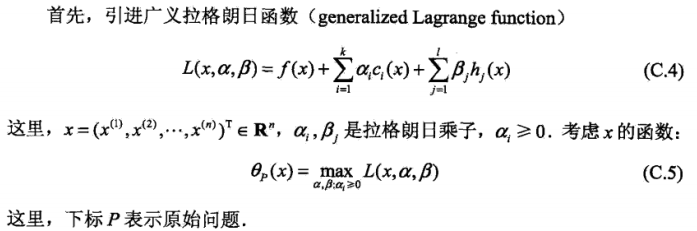
1. 提出的问题2：7.4.2的计算阈值b那里：

https://qqadapt.qpic.cn/txdocpic/0/de6d7014fe679889fd0d75a604d2ffe7/0?w=554&h=29

这是怎么得到两个b相等的？

讨论后的理解：这个问题是我理解错了。本身这两个b就是相等的，因为和本来就应该是一起计算出来的能够满足分类模型的式子，他们本来对应的b值就应该是相等的，只是如果他们两个均满足，那么说明计算出来的就是模型中的真正的b，只有一个值，这样的话两者自然是相等的，如果不满足上述条件，计算出来的值就是有范围的，但是仍然都是可以满足条件的，所以取平均值。

1. 别人提出的问题的理解：
2. 问题3：



这里的θp(x)，原问题是min，θp(x)是max，为什么能表示原问题？

自己的理解：这里的就是原问题中的，因为在满足约束条件的情况下，由于是等于0的，且也是小于0的，为了满足最大化情况，必须有均等于0，这样的话代入上式就可以得到了，因此该式完全可以替代原模型。

1. 问题4： 为什么使用核函数？核函数的本质是什么？

自己的理解：核函数的本质就是达成一个映射的作用，将一个欧拉空间或者是离散空间上的变量集或者是向量映射到一个高维或者是低维的希尔伯特空间当中，虽然在原空间中是无法线性可分的，但是在映射到希尔伯特空间后就可以用支持向量机分割数据了，这就是把非线性问题转化成线性问题去解决。

1. 问题5：支持向量的作用，是不是把正例点和负例点尽量分割的开一些？

自己的理解：支持向量就是离分割平面最近的几个数据点，这几个数据点代表着其他的数据点离分割平面更远，如果可以找到远一些的支持向量，那么这个分割就更加成功，可以让分割结果更加准确，找到支持向量的意思就是想把那些离分割平面最近，最不容易确定分类的那些点分类完，那么其他的点也就自然被分好类了。

1. 读书计划

1、本周完成的内容章节：《统计机器学习》第七章

2、下周计划：《统计机器学习》第八章

四、读书摘要及理解

1、支持向量机是一种二类分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使它成为实质上的非线性分类器，支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划问题。

2、如果数据是真正线性可分的，类似于感知机，可以有很多的把数据完全分割开来的平面，但是间隔最大化的超平面只有一个，这就是支持向量机求解的东西，让间隔最大化，才能做到分类的更加准确。

线性可分支持向量机就是给定线性可分数据集，通过间隔最大化或等价地求解相应的凸二次规划问题学习得到的分离超平面为：



以及相应的分类决策函数：



称为线性可分支持向量机。

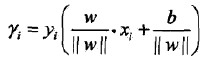
3、函数间隔就是超平面关于样本点的间隔，为：



定义超平面关于训练数据的函数间隔为超平面关于数据集中所有样本点的函数间隔之最小值，即：

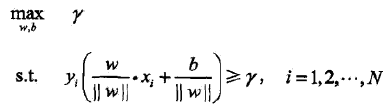


显然光有函数间隔是不够的，可以加一些规范化让距离值可以唯一的得到，就是几何间隔，即：

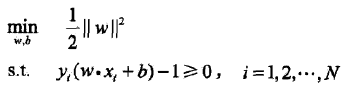


最小化问题变成最小化它就行了。

4、最大间隔分离超平面就是让函数间隔值最大，找到这些点中最小的函数间隔，让这些最小函数间隔点最大化，就是线性可分向量机的思路。即：



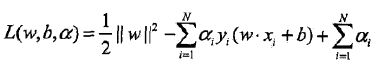
实际上可以取γ为1，对约束条件并没有本质的影响，那么最优化问题变成：



这里计算得到的分离超平面是唯一的，详细证明在书上。

支持向量就是让上述约束条件等号成立的点，即在距离分割平面最近的那一个平面上，让这个平面最远就是最优化的目标。这两个平面之间的距离就是间隔边界，自然就是越大越好。

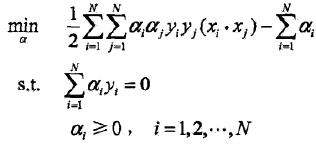
5、定义拉格朗日函数，可以得到新的对偶问题，拉格朗日函数如下：



原始问题的对偶问题就是极大极小问题：

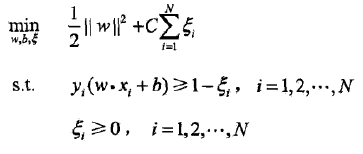


通过求偏导由内到外，可以得到：



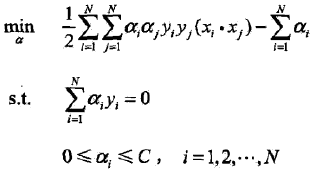
这就是原问题的对偶问题。

6、线性支持向量机就是说数据并不是严格线性可分的，但是抛弃某几个点之后，数据就可以基本满足线性可分了，那么就可以用线性可分向量机来做了，问题为：



这里的就是松弛变量，替不满足线性可分的点的距离补足一个差距，让它基本满足线性可分，这就是一个惩罚项，目标函数中C就是起到惩罚的权重作用。那么既然有误分类的店，说明b的值就不是唯一的了，在一个范围之内。

和之前的对偶问题类似，可以得到线性支持向量机的对偶问题为：



找到一个满足上述条件的，带入原模型式子里就可以求得一个b，那么有不同的及说明肯定不止一个b，所以实际计算时可以这些满足条件的样本店对应的b的平均值。

利用合页损失函数可以得到另一种最优化问题：



7、非线性支持向量机就要依靠核函数来进行转化了，因为数据是完全非线性的，那么单纯的进行补足就不起作用了，误差太大，因此就需要核函数来将非线性数据进行映射，映射到一个希尔伯特空间上，在这个空间上可以应用线性支持向量机来做。

核函数的定义就是，在输入空间中，如果存在一个从到的映射：



使得对所有的，函数满足条件



则称为核函数，为映射函数，式中是内积。

核技巧的想法就是，只定义核函数，不显示地定义映射函数，这样直接计算核函数就比较容易，可以在约束条件和目标函数中进行替换，完成映射的过程：

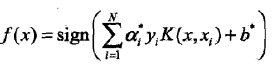


8、怎么找到核函数呢？可以先定义一个映射，让它构成向量空间，然后在这个空间上定义内积构成内积空间，最后再完备化成希尔伯特空间。那么一个函数为正定核函数的充要条件就是对任意的，对应的Gram矩阵：

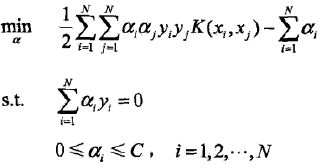


是半正定矩阵。

有很多的常用的核函数，因为自己构造一个核函数往往很难，因为证明半正定矩阵很困难，所以可以一些常用的函数，比如多项式核函数，高斯核函数，对于文本分类问题，还可以用字符串核函数。那么得到的非线性支持向量机就可以变成如下的模型：



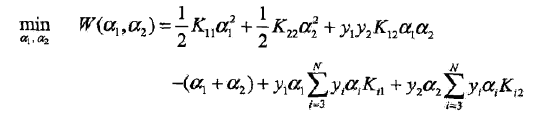
优化问题也变成：

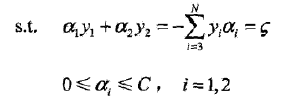


9、在最优化上述问题的过程中，有许多最优化算法可以求得上述的最优解，但是一旦数据量很大就不再好用了，所以这里书上介绍了SMO算法，可以简单的计算出最优解。

SMO算法是一种启发式算法，如果所有的变量都满足约束条件，那么解就已经得到了，如果不是，那么选择两个变量，固定其他的变量，针对这两个变量构建一个二次规划问题。这个二次规划问题关于这两个变量的解应该更接近原始二次规划问题的解。这样，只通过结算子问题的解就可以大大提升问题求解的速度。子问题有两个变量，一个是违反KTT条件最严重的一个，另一个由约束条件自己确定。

这样，SMO关于最优化的问题就变成了：





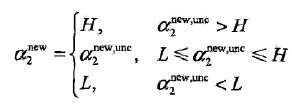
最终得到的未经剪辑的解是：



其中，



经过剪辑后的解是：



那么相应的另一个解就是：



至于变量的选择，可以先在外层循环，选出KKT违反最严重的样本点，将其作为第一个变量。第二个变量就是在内存循环中找，虚妄可以让的变化足够大，即让最大。之后重新计算阈值b，在更新之前要用到的E值，这样就可以完成SMO算法的过程。