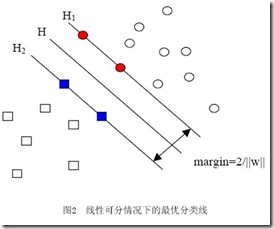
1. SVM: 寻找最优超平面将两类正确分开(训练错误率为0),且使分类间隔最大。

参考网址：<http://www.dataguru.cn/thread-371987-1-1.html>

1）线性可分情况

[](http://images.cnitblog.com/blog/458371/201212/31223651-cb3b6f9f79e34304a30c1042681d46f0.jpg)

  这个超平面可以用分类函数http://img.blog.csdn.net/20131107201211968表示，当f(x) 等于0的时候，x便是位于超平面上的点，而f(x)大于0的点对应 y=1 的数据点，f(x)小于0的点对应y=-1的点。

因此，最优超平面到点的距离为：http://img.blog.csdn.net/20131107201759093

带颜色的支持向量刚好在边界上，所以它们满足http://img.blog.csdn.net/20131111155244218。

因此，H1和H2之间的距离为:2/||W||.

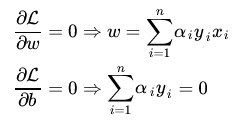
最大化此距离等价于最小化公式1：

http://img.my.csdn.net/uploads/201210/25/1351141994_1802.jpg

利用拉格朗日乘子法进行求解：

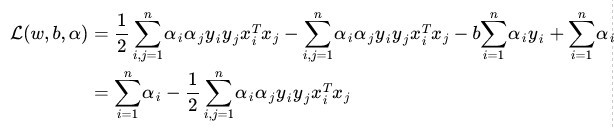


首先固定*http://img.blog.csdn.net/20131111195836468，*要让 L 关于 w 和 b 最小化，我们分别对w，b求偏导数，即令 ∂L/∂w 和 ∂L/∂b 等于零（对w求导结果的解释请看本文评论下第45楼回复）：

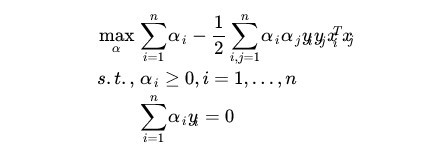


    将以上结果代入之前的L，得到：



从而有：  


由此，得到公式1的对偶问题：



求出http://img.blog.csdn.net/20131111195824031之后，根据：

http://img.my.csdn.net/uploads/201301/11/1357838666_9138.jpg

http://img.my.csdn.net/uploads/201301/11/1357838696_3314.png

就可以求出w和b，由此推出最优超平面。

2）在线性不可分的情况下，选择核函数通过将数据映射到高维空间，来解决在原始空间中线性不可分的问题。

http://my.csdn.net/uploads/201206/04/1338740718_7761.JPG这里*ϕ*：X->F是从输入空间到某个特征空间的映射，这意味着建立非线性学习器分为两步：

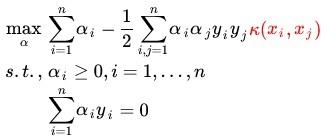
1. 首先使用一个非线性映射将数据变换到一个特征空间F，
2. 然后在特征空间使用线性学习器分类。

http://my.csdn.net/uploads/201206/04/1338741121_9867.JPG

  如果有一种方式可以**在特征空间中直接计算内积〈φ(x*i* · φ(x)〉**，就像在原始输入点的函数中一样，就有可能将两个步骤融合到一起建立一个非线性的学习器，**这样直接计算法的方法称为核函数方法：**核是一个函数K，对所有x，z，满足http://my.csdn.net/uploads/201206/04/1338741445_1451.JPG，这里φ是从X到内积特征空间F的映射。

**计算两个向量在隐式映射过后的空间中的内积的函数叫做核函数。转化为如下格式：**

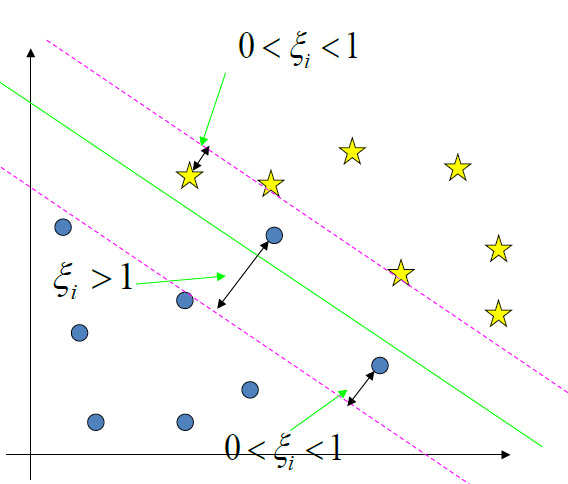
http://img.my.csdn.net/uploads/201304/03/1364956916_5948.jpg

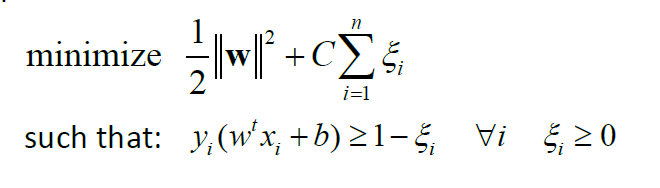


多项式核函数：[clip_image074](http://images.cnitblog.com/blog/458371/201212/31223713-c176af62aa454ae091efc240991bb615.jpg)

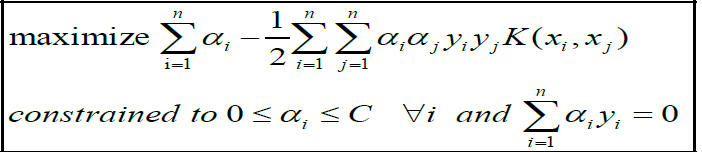
高斯核函数：[clip_image076](http://images.cnitblog.com/blog/458371/201212/31223713-36cd1aba86854217891760176063887b.jpg)

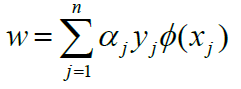
软间隔：允许有错分，

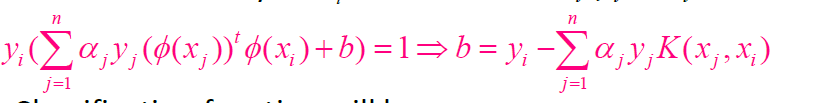


问题转化为：

其中c取无穷大时，表明需要严格遵守，c取有限值时，可以有部分不满足约束条件。



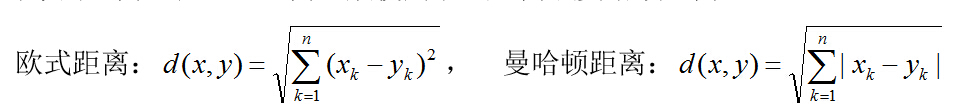
，



1. KNN

概念：监督学习方法，给定测试样本，基于某种距离度量找出训练集中与其最靠近的K个样本，然后基于这K个样本的信息确定其类别。

优势：1）通过计算对象间距离来作为各个对象之间的非相似性指标，避免了对象之间的匹配问题2）KNN通过依据k个对象中占优的类别进行决策，而不是单一的对象类别决策。



算法流程：

1）计算测试数据与各个训练数据之间的距离；

2）按照距离的递增关系进行排序；

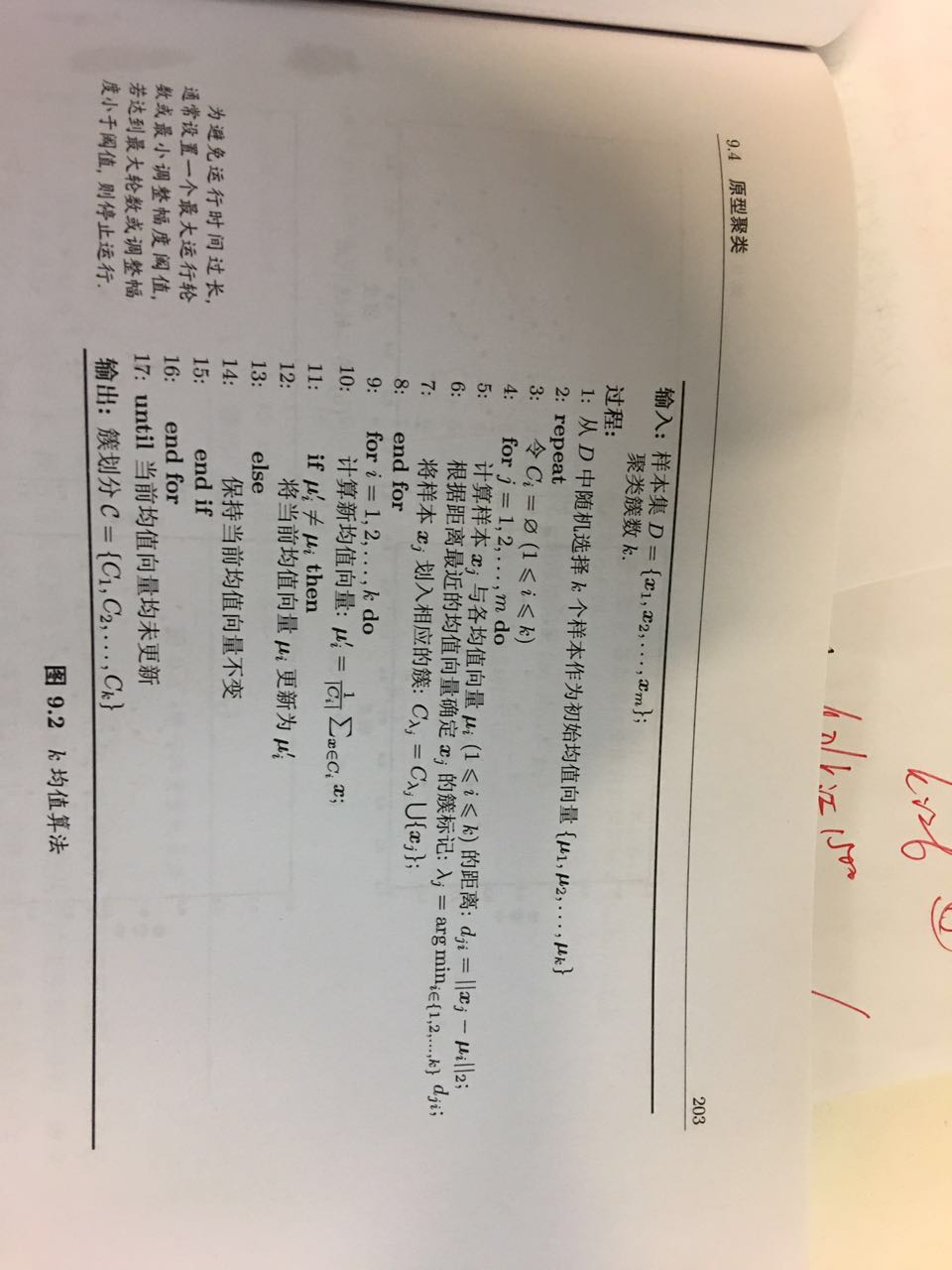
3）选取距离最小的K个点；

4）确定前K个点所在类别的出现频率；

5）返回前K个点中出现频率最高的类别作为测试数据的预测分类。

1. Kmeans(贪心策略,通过迭代优化求近似解)

算法流程:



1. 聚类算法（k-means和FCM相对优于其他，k-means聚类算法的初 始点选择不稳定，是随机选取的，这就引起聚类结果的不稳定。层次聚类虽然 不需要确定分类数，但是一旦一个分裂或者合并被执行，就不能修正，聚类质量受限制；FCM对初始聚类中心敏感，需要人为确定聚类数，容易陷入局部最优解；SOM与实际大脑处理有很强的理论联系。但是处理时间较长，需要进一步研究使其适应大型数据库。）

Kmeans：（k为参数，把n个对象分成k个簇，使簇内具有较高的相似度，而簇间的相似度较低。采用平方误差准则http://files.chinaaet.com/images/2010/10/18/4435423708146.gif）算法流程如3

层次聚类：先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到所有对象都在一个簇中，或者某个终结条件被满足。

(1) 将每个对象看作一类，计算两两之间的最小距离；  
(2) 将距离最小的两个类合并成一个新类；  
(3) 重新计算新类与所有类之间的距离；  
(4) 重复(2)、(3)，直到所有类最后合并成一类。

SOM聚类：假设在输入对象中存在一些拓扑结构或顺序，可以实现从输入空间(n维)到输出平面(2维)的降维映射，其映射具有拓扑特征保持性质,与实际的大脑处理有很强的理论联系。

(1) 网络初始化，对输出层每个节点权重赋初值；  
(2) 将输入样本中随机选取输入向量，找到与输入向量距离最小的权重向量；  
(3) 定义获胜单元，在获胜单元的邻近区域调整权重使其向输入向量靠拢；  
(4) 提供新样本、进行训练；  
(5) 收缩邻域半径、减小学习率、重复，直到小于允许值，输出聚类结果。

FCM聚类：用模糊数学的方法进行聚类分析，就是模糊聚类分析

(1) 标准化数据矩阵；  
(2) 建立模糊相似矩阵，初始化隶属矩阵；  
(3) 算法开始迭代，直到目标函数收敛到极小值；  
(4) 根据迭代结果，由最后的隶属矩阵确定数据所属的类，显示最后的聚类结果。

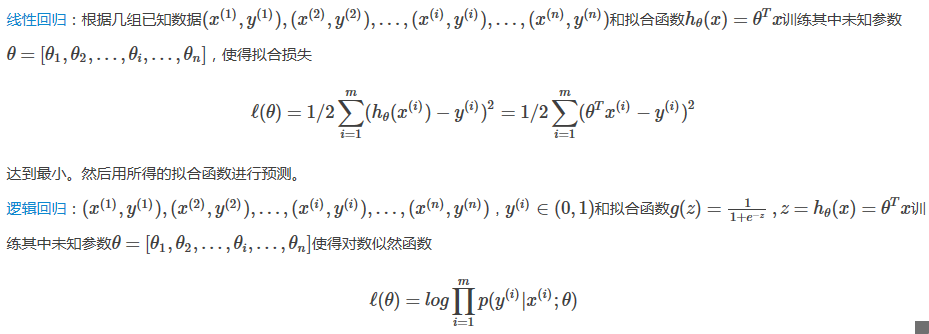
1. 逻辑回归:

Logistic回归目的是从特征学习出一个0/1分类模型，而这个模型是将特性的线性组合作为自变量，由于自变量的取值范围是负无穷到正无穷。因此，使用logistic函数（或称作sigmoid函数）将自变量映射到(0,1)上，映射后的值被认为是属于y=1的概率

http://img.my.csdn.net/uploads/201304/05/1365174192_8325.png

当我们要判别一个新来的特征属于哪个类时，只需求http://img.my.csdn.net/uploads/201304/05/1365175136_8232.png即可，若http://img.my.csdn.net/uploads/201304/05/1365175136_8232.png大于0.5就是y=1的类，反之属于y=0类。

概念: 面对一个回归或者分类问题，建立代价函数，然后通过优化方法迭代求解出最优的模型参数，然后测试验证我们这个求解的模型的好坏.





1. PCA：本质上是将方差最大的方向作为主要特征，并且在各个正交方向上将数据“离相关”，也就是让它们在不同正交方向上没有相关性。

