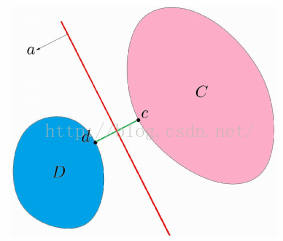
SVM

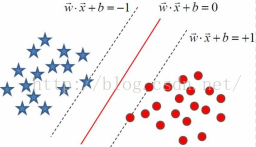
前人的功劳，我是后来者：<https://blog.csdn.net/american199062/article/details/51322852>

支持向量机是属于原创性、非组合的具有明显直观几何意义的分类算法，具有较高的准确率。使用SVM算法的思路：（1）简单情况，线性可分情况，把问题转化为一个凸优化问题，可以用拉格朗日乘子法简化，然后用既有的算法解决；（2）复杂情况，线性不可分，用核函数将样本投射到高维空间，使其变成线性可分的情形，利用核函数来减少高纬度计算量。

1. ****SVM相关基本概念****

****分割超平面****

设C和D为两不相交的凸集，则存在超平面P，P可以将C和D分离。 IMG_256两个集合的距离，定义为两个集合间元素的最短距离。 做集合C和集合D最短线段的垂直平分线。          （图像摘自七月算法）

但是， 如何定义两个集合的"最优"分割超平面？找到集合“边界”上的若干点，以这些点为“基础”计算超平面的方向，以两个集合边界上的这些点的平均作为超平面的“截距”。这些点被称作支持向量，点是可用向量方式表示。       （图像取自七月算法）

****输入数据****

假设给定一个特征空间上的训练数据集IMG_259

其中，IMG_260,IMG_261为第i个实例（若n>1，即x是多维度，具有多个属性特征，此时IMG_262为向量）； IMG_263为IMG_264的类标记，当IMG_265为+1时，称IMG_266为正例，当IMG_267为-1时，称IMG_268为负例。

****线性可分支持向量机****

给定线性可分训练数据集，通过间隔最大化得到的分离超平面为IMG_269，相应的分类决策函数IMG_270该决策函数称为****线性可分支持向量机****。其中，IMG_271****是某个确定的特征空间转换函数，它的作用是将x映射到（更高的）维度，最简单直接的：IMG_272****。事实上，求解分离超平面问题可以等价为求解相应的凸二次规划问题。

****整理符号****

        分割平面：IMG_273

        训练集：IMG_274

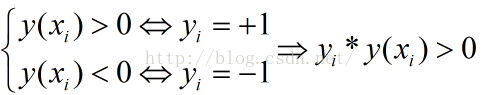
        目标值：IMG_275

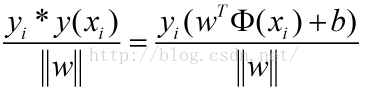
        新数据的分类：IMG_276

****二、SVM推导过程****

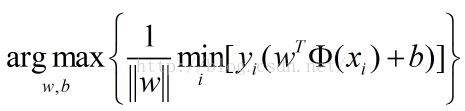
****推导目标函数****

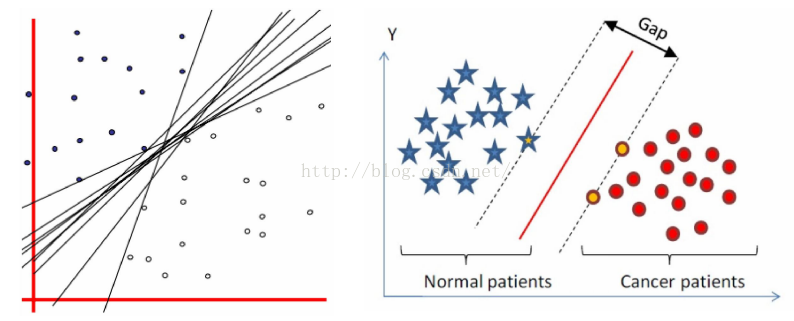
        根据题设IMG_277

        有：

****w,b等比例缩放，则t\*y的值同样缩放，从而****：

****最大间隔分离超平面****

        目标函数：，****表示最近点到直线距离尽可能大****

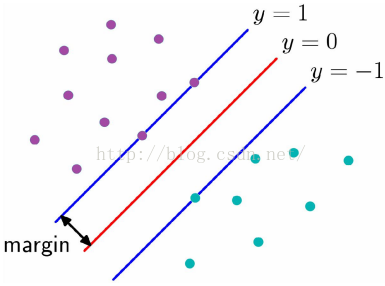


（图像取自七月算法）

****函数间隔和几何间隔****

       分割平面：IMG_282  (函数间隔)

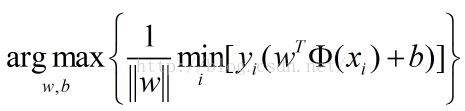
       总可以通过****等比例缩放w****的方法，使得两类点的函数值都满足IMG_283

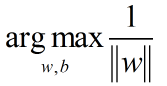
（图像取自七月算法）

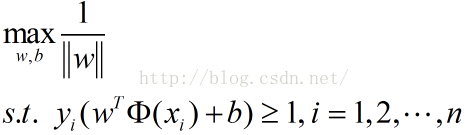
****建立目标函数****

       1.总可以通过等比例缩放w的方法，使得两类点的函数值都满足IMG_285

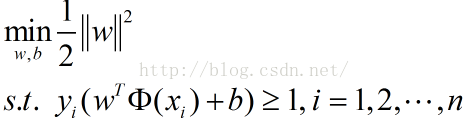
       2.约束条件：IMG_286

       3.原目标函数：

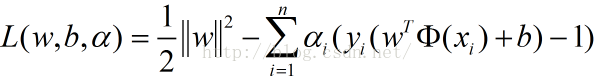
       4.新目标函数：



       5.目标函数变换一下：



        6.拉格朗日乘子法

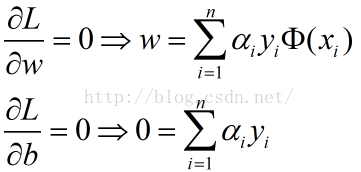


         7.原问题是极小极大问题

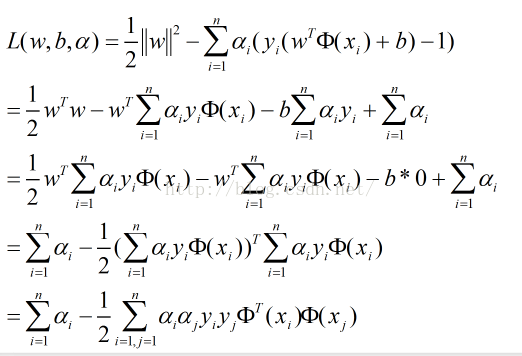
IMG_292

                       原问题的对偶问题是极大极小问题IMG_293

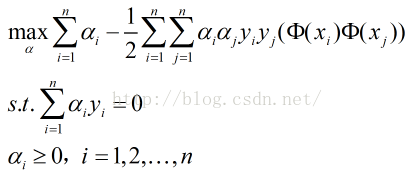
         8.将6中的拉格朗日函数分别对w, b 求偏导并令其为0：



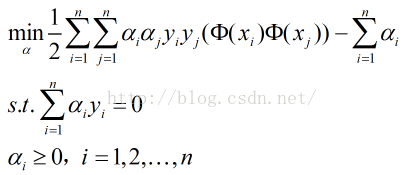
           9.计算拉格朗日的对偶函数



         10.继续求IMG_296的极大

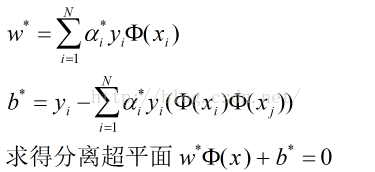


         11.整理目标函数：添加负号



        12.线性可分支持向量机学习算法

               计算结果如下



        13.分类决策函数

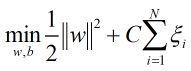
IMG_300

****三、线性不可分SVM****

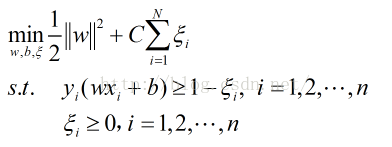
  1.若数据线性不可分，则增加松弛因子IMG_301，使函数间隔加上松弛变量大于等于1，

         则约束条件变成

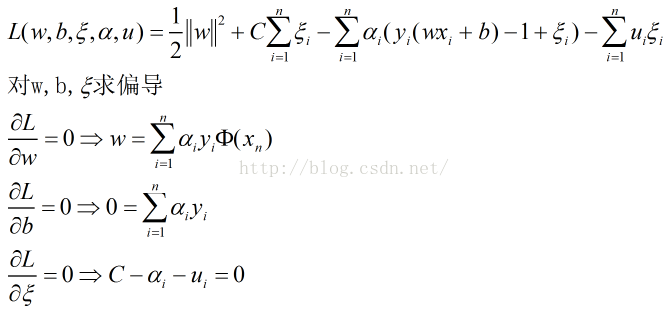
IMG_302

         目标函数：    （这里是为了保证松弛因子不至于过大）

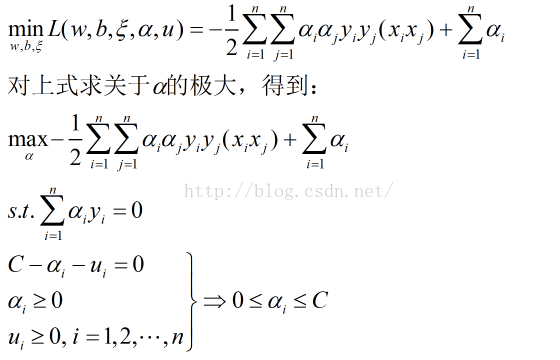
         2.此时的凸优化为



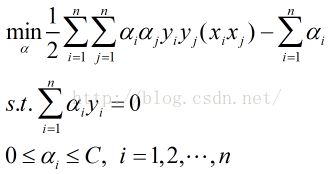
          3.拉格朗日函数



        4.将三式代入L中，得到

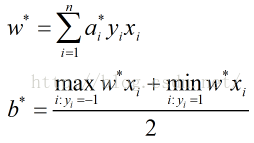


       5. 整理，得到对偶问题的最优化问题



      求得最优解IMG_308

      6.计算



       实践中往往取支持向量的所有值取平均，作为b\*

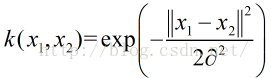
       7.求得分离超平面IMG_310

       8.分类决策函数为

IMG_311

****核函数****：可以使用核函数，将原始输入空间映射到新的特征空间，从而使得原本线性不可分的样本可在核空间可分。

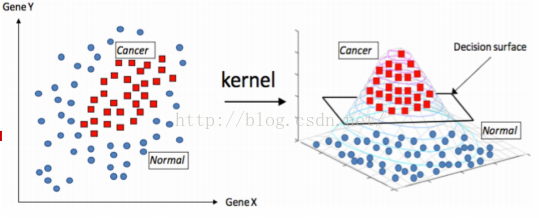
       有多项式核函数IMG_312

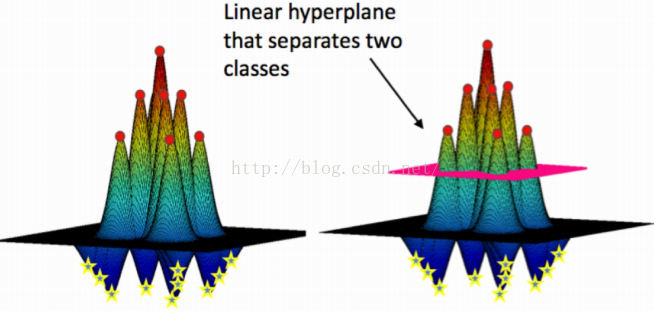
       高斯核函数RBF 

       字符串核函数

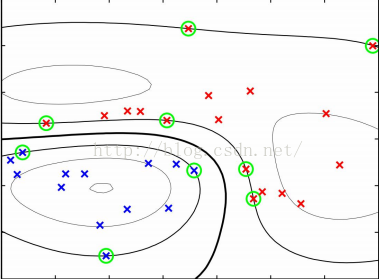
        在实际应用中，往往依赖先验领域知识或交叉验证等方案才能选择有效的核函数。没有更多先验信息，则使用高斯核函数。

核函数映射：

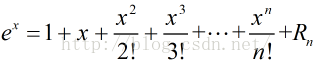
（图像取自七月算法）

（图像取自七月算法）

                  高斯核

（图像取自七月算法）

          粗线是分割超“平面”，其他线是y(x)的等高线，绿色圈点是支持向量点。

          高斯核是无穷维的，因为

          注：SVM和Logistic回归的比较：（1）经典的SVM，直接输出类别，不给出后验概率；（2）Logistic回归，会给出属于哪一个类别的后验概率；（3）比较重点是二者目标函数的异同。