## 第六章 支持向量机

6.1 试比较SVM与logistic回归的异同。

参考答案：

|  |  |  |
| --- | --- | --- |
|  | SVM | logistic回归 |
| 相同点 | 两者都是常见的分类算法。就优化目标而言，两者都构造了代价函数来增加对分类影响较大的特征的权重，而减小对分类影响较小的特征的权重。 | |
| 不同点 | 更侧重突出“间隔”的概念，只考虑支持向量，即采用少数点去训练分类器。 | 采用非线性映射的方法，对线性回归的基本思想进行推广来训练分类器。 |

6.2 试证明样本空间中任意点***x***到超平面的(***w***, *b*)的距离为。

参考答案：

设***x***0为超平面上任一点，***w***为平面的法向量，则点***x***到超平面的距离

6.3 对于样本空间中的一划分超平面***w***T***x*** + *b* = 0，有***w*** = (-1, 3, 2)，*b* = 1。则判断如下向量是否为支持向量，并求出间隔。

(1) ***x***1 = (4, -2, 2)

(2) ***x***2 = (2, 5, -6.5)

(3) ***x***3 = (4, -2, 4)

参考答案：

(1) = -5 < -1，不是支持向量。

(2) = +1，是正例一侧的支持向量。

(3) = -1，是负例一侧的支持向量。

间隔

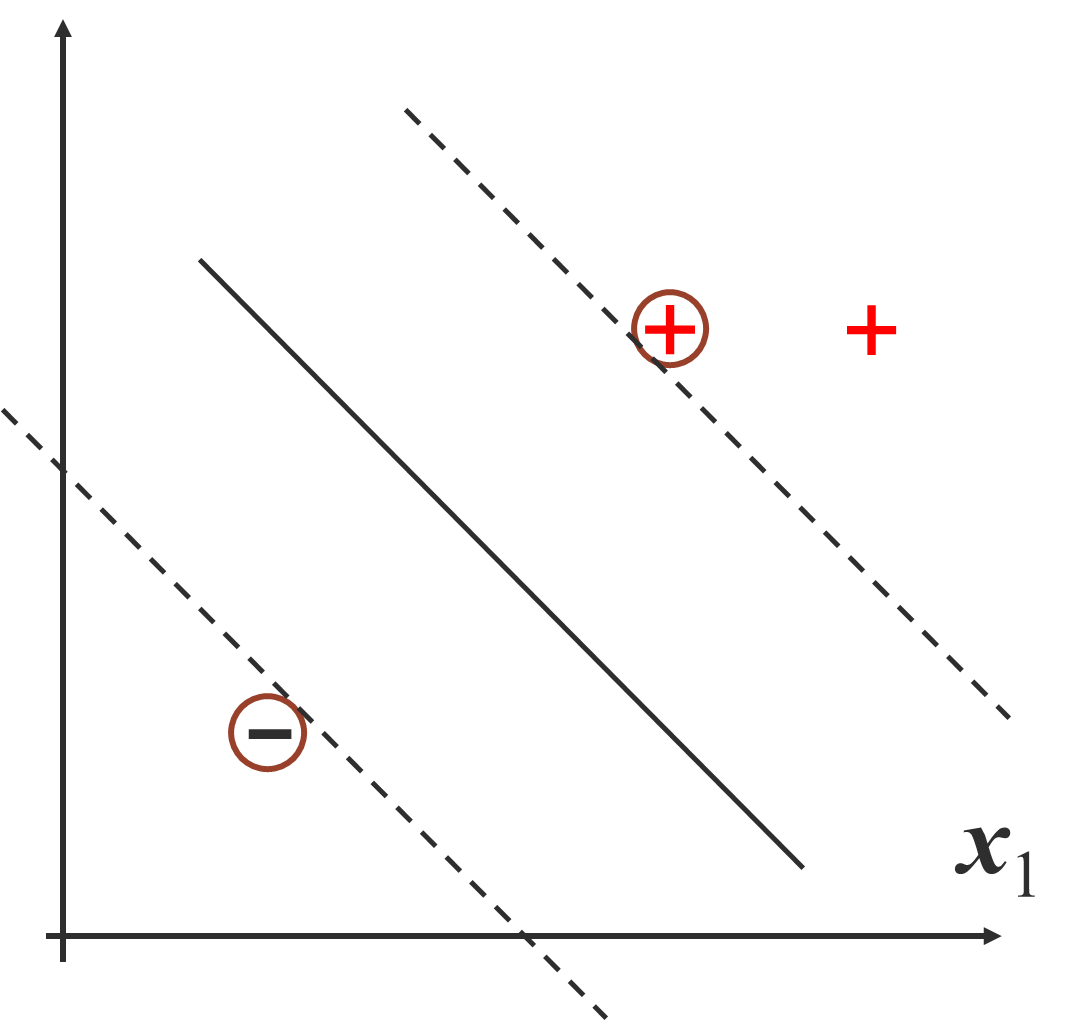
6.4 已知正例点***x***1 = (3,3)T, ***x***2 = (4,3)T 负例点***x***3 = (1,1)T试求最大间隔分离超平面和分类决策函数，并在图上画出分离超平面、间隔边界及支持向量。

参考答案：

对偶问题

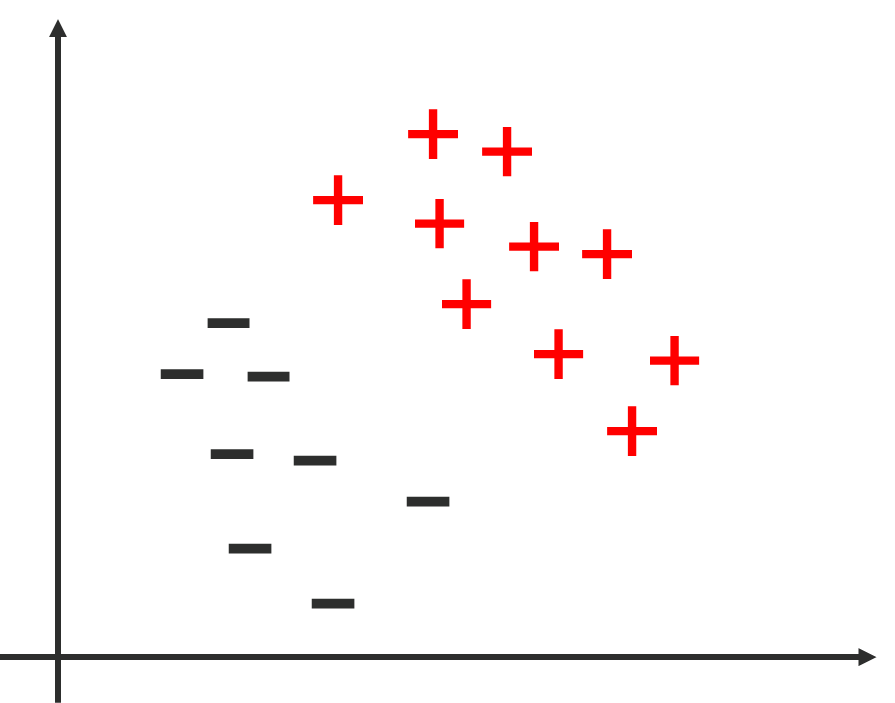
将代入

则问题转化为，求得。则可知***x***1，***x***3为支持向量。。则分隔超平面，分类决策函数。

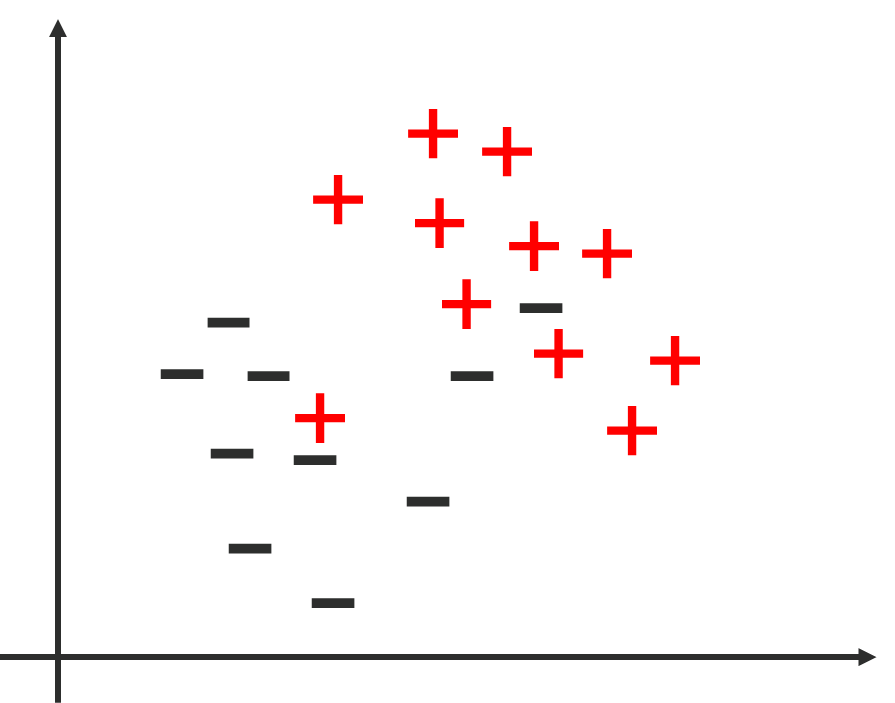


6.5 判断下列数据集是否线性可分。若线性可分，定性地画出分隔超平面及间隔；若不可分，列举用SVM处理时的改进方法，并简要说明原理。

（1）

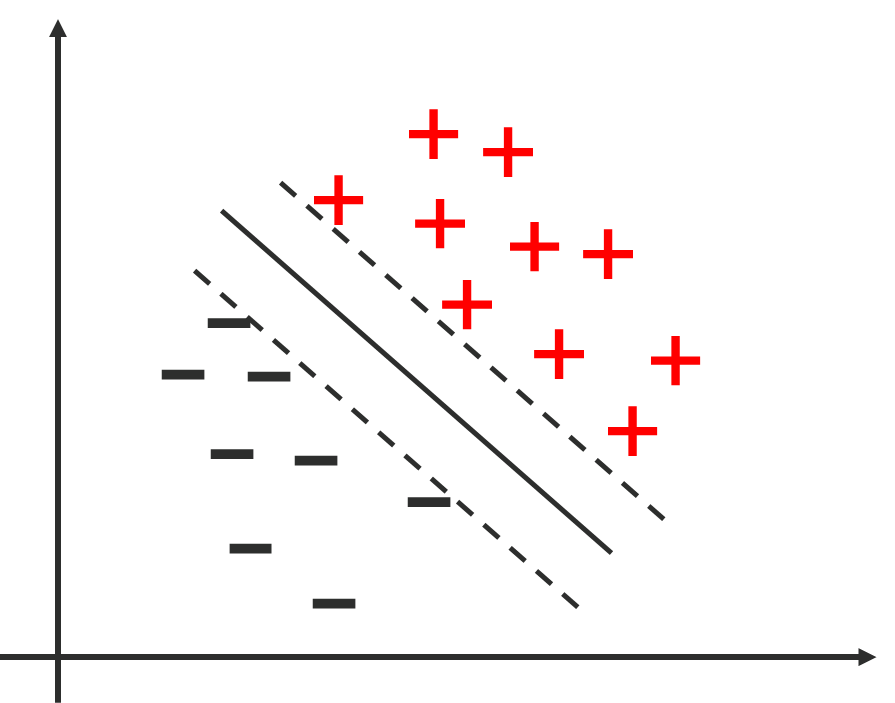


（2）



参考答案：

（1）线性可分



（2）线性不可分，利用SVM处理时可引入软间隔或选取合适的核函数。

6.6 假设输入空间是R2，核函数是***z***) = (***x***T***z***)2 ，试找出其相关的特征空间H和映射Φ(***x***)：**R**2 → H.

参考答案：

取特征空间H = **R**3 ，记***x*** = (*x*1, *x*2)T, ***z*** = (*z*1, *z*2)T,

(***x***T***z***)2 = (*x*1*z*1 + *x*2*z*2)2 = (*x*1*z*1)2 + 2*x*1*z*1*x*2*z*2 + (*x*2*z*2)2

则可取映射

特征空间H = **R**3时，

6.7 考虑XOR问题，其中有四个样本点：

(0, 0), -; (0, 1), +; (1, 0), +; (1, 1) -;

利用核函数***z***) = (1+***x***T***z***)2对原始空间进行变换，线性映射函数Φ(***x***) (Φ(***x***)∈**R**6)，求映射后的样本点Φ(***x***1)，Φ(***x***2)，Φ(***x***3)，Φ(***x***4)。

参考答案：

由核函数的表达式可得，

则

Φ(***x***1) = (1, 0, 0, 0, 0, 0), -;

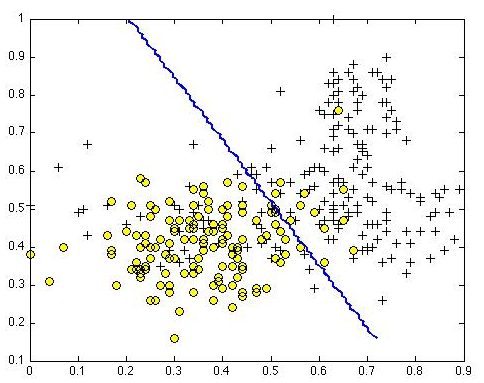
Φ(***x***2) = (1, 0, , 0, 0, 1), +;

Φ(***x***3) = (1, , 0, 0, 1, 0), +;

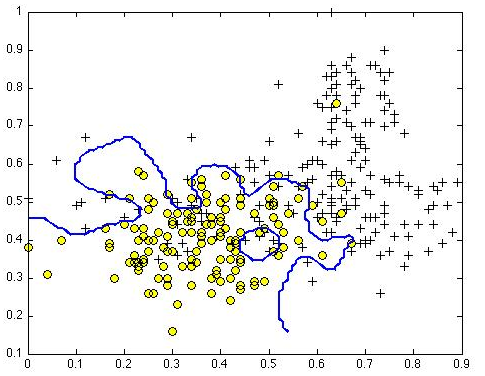
Φ(***x***4) = (1, , , , 1, 1). -.

6.8 假设利用高斯核训练一个SVM模型，数据集和决策边界如下，为实现更好的分类效果，请问应当如何调整模型参数*C*和σ2？

（1）



（2）下图所示的决策边界在交叉验证集性能表现很差。



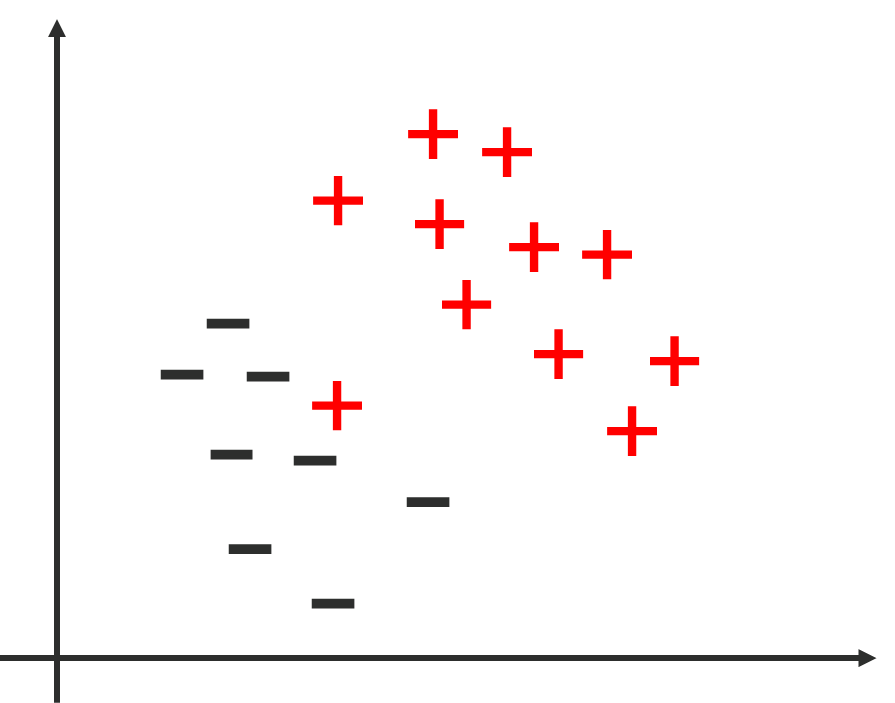
参考答案：

（1）欠拟合，说明该模型容错能力过强，所以应该增大*C*，减小σ2 。

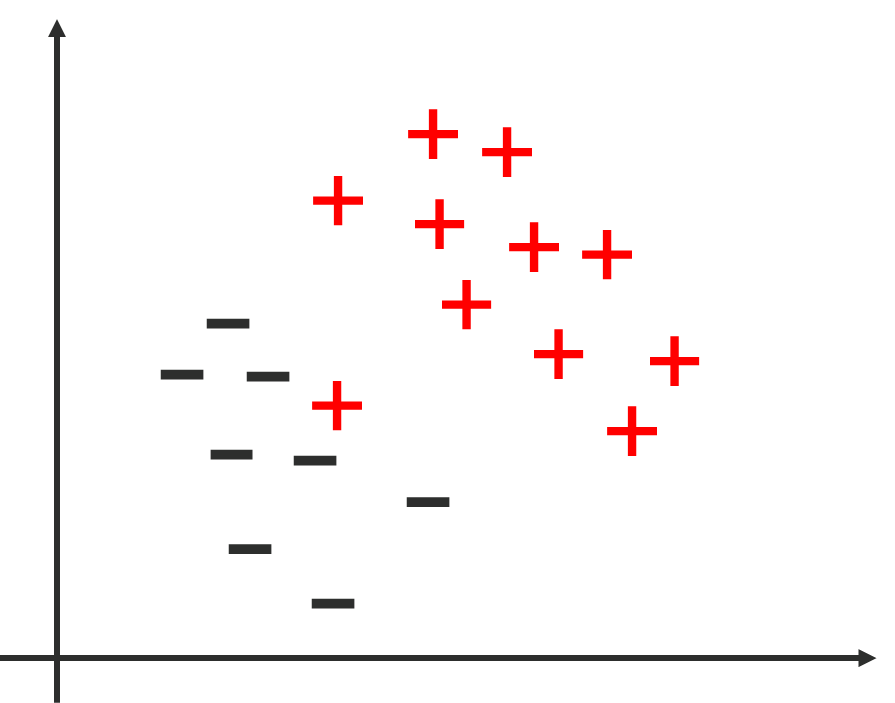
（2）从直观上该决策边界较好的拟合了数据集，但是交叉验证集上的性能较差，说明发生了过拟合，所以应该减小*C*，增大σ2 。

6.9 考虑如图所示的数据集，当C = 0和C→∞时，定性的作出决策边界，并说明理由。

（1）C = 0



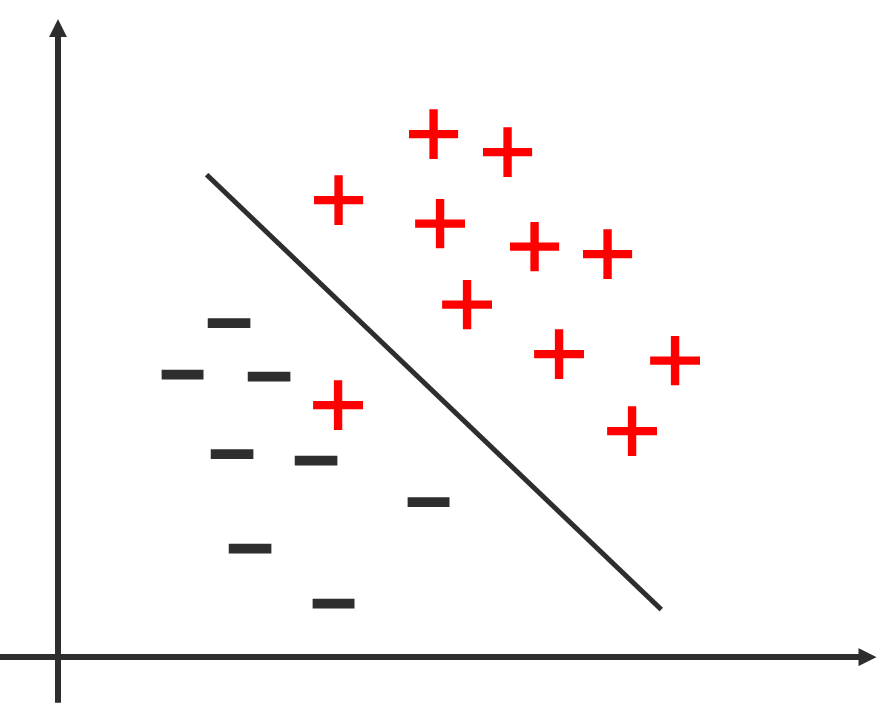
（2）C→∞



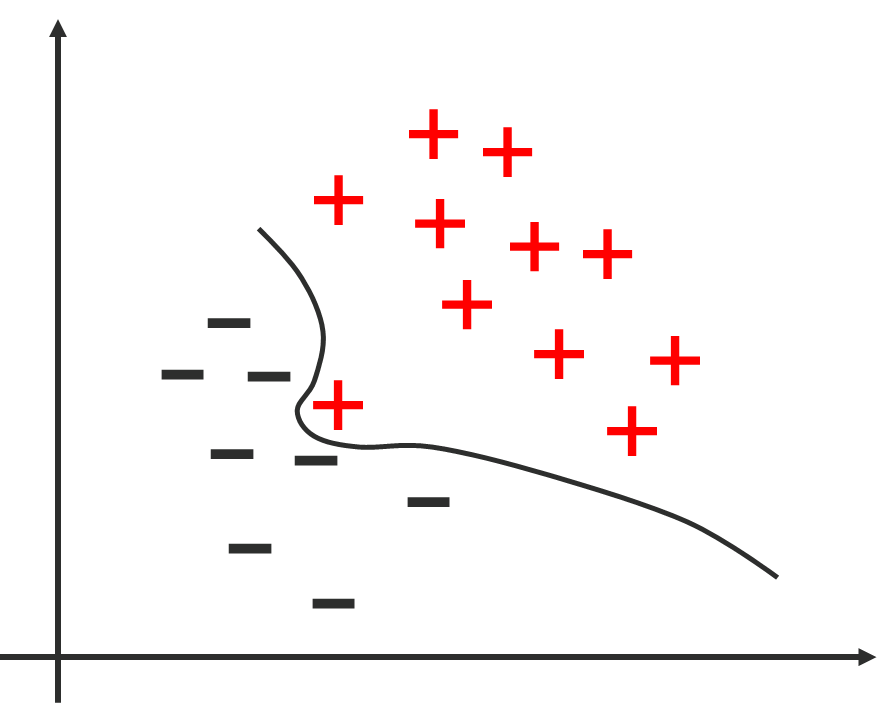
参考答案：

C = 0时，模型有较强的容错能力; C→∞相当于一个完全的硬间隔模型。

（1）C = 0



(2) C→∞

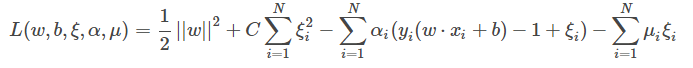


6.10 线性支持向量机还可以定义为以下形式：

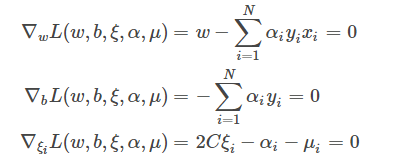
试求其对偶形式。

参考答案：

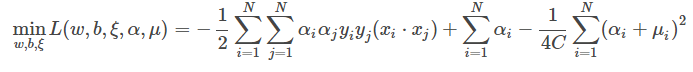
拉格朗日函数：



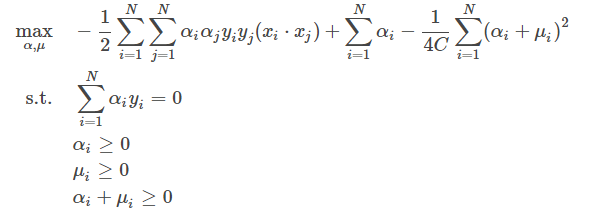
对***w***, *b*, 分别求导，并令其值等于零，得到



将上述等式代入，得到



得到对偶问题



6.11 结合logistic回归章节内容，提出几种SVM实现多分类的方案。

参考答案：

SVM实现多分类主要有三种方法。一是直接修改SVM的目标函数，使其适应多分类的情况。但这种方法往往复杂度较高，适合类别较少的情况。二是采用前序章节提到的OvR和OvO策略，训练多个SVM二分类器。

6.12 利用SMO算法解决SVM的优化问题时，对*α*2的更新有如下规则：

试证明上下界*L*, *H*的确定规则如下：

参考答案：

由SMO算法，选择两个参数，后，假定其他参数已知，则原优化问题可改写如下：

对*y*1 = *y*2和*y*1≠*y*2约束条件可分别改写为：, 。

*y*1≠*y*2时，，对有0 < < C; 更新前，对*k*，有。

则对，有，再结合原约束条件可得：

*y*1=*y*2时，同理可证。

6.13 利用SMO算法解决SVM的优化问题时，对参数，的选择采用启发式的选择方法，其中外层循环选择第一个变量时，需要检验样本点(***x***i, *y*i)是否满足KKT条件，简述可进行如下等价的理由。

其中

参考答案：

原始问题的对偶问题，

设对***w***\*, *b*\* = ，问题转化为：

(1)时，要使(\*)式最大，

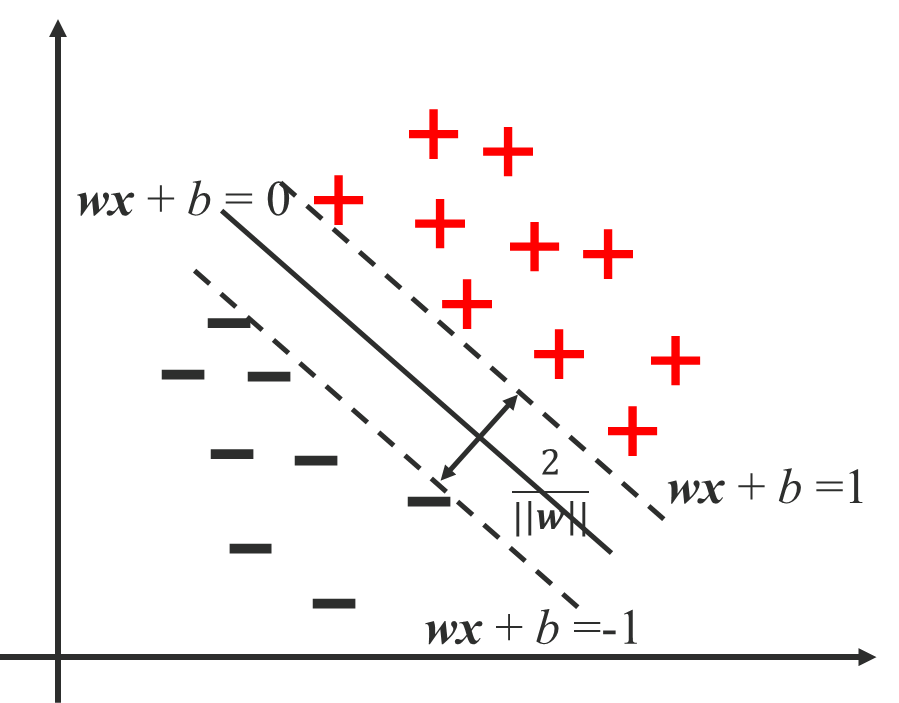
(2) 时，取值对(\*)式的值无影响，故可取(0, *C*)间任一值。

(3)时，，要使(\*)式最大，

6.14 参考SVM的示意图及优化目标，简要画出SVR的示意图并给出优化目标。

（1）SVM

优化目标函数：



优化目标：使分隔超平面两侧最近的点的距离和最大。

(2)SVR

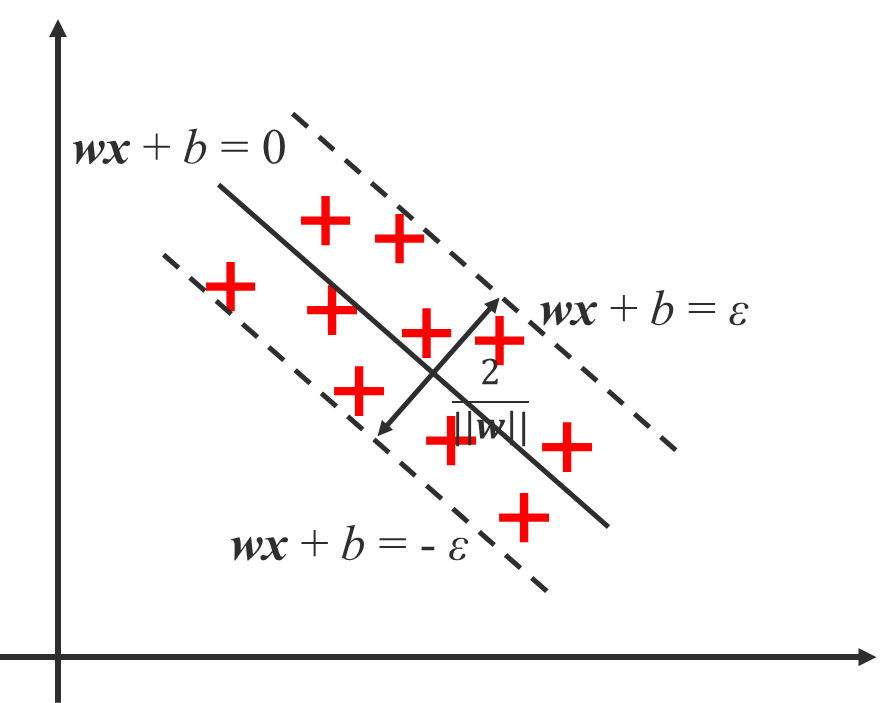
优化目标函数：

示意图：

优化目标：

参考答案：

示意图：



优化目标：使超平面两侧最远的点到超平面距离和最小。

6.15 归纳总结利用SVM处理分类问题的一般步骤。

参考答案：

（1）收集数据。

（2）数据预处理。从收集到的原始数据中提取最有价值的特征集，并分别标注正例和反例。

（3）分析数据，确定模型。可采用可视化、降维的方法，对特征总体或某些维度进行初步分析，确定是利用线性SVM模型还是非线性模型、是否需要引入软间隔以及核函数的类型等。

（4）训练算法。将数据集引入模型进行训练，在此过程中还需要进行参数的调优。

（5）测试算法。可以利用前面章节的知识，在交叉验证集上进行性能测试，评估模型的泛化能力。

（6）模型应用。