Lecture 5
支持向量机SVM

内容提要

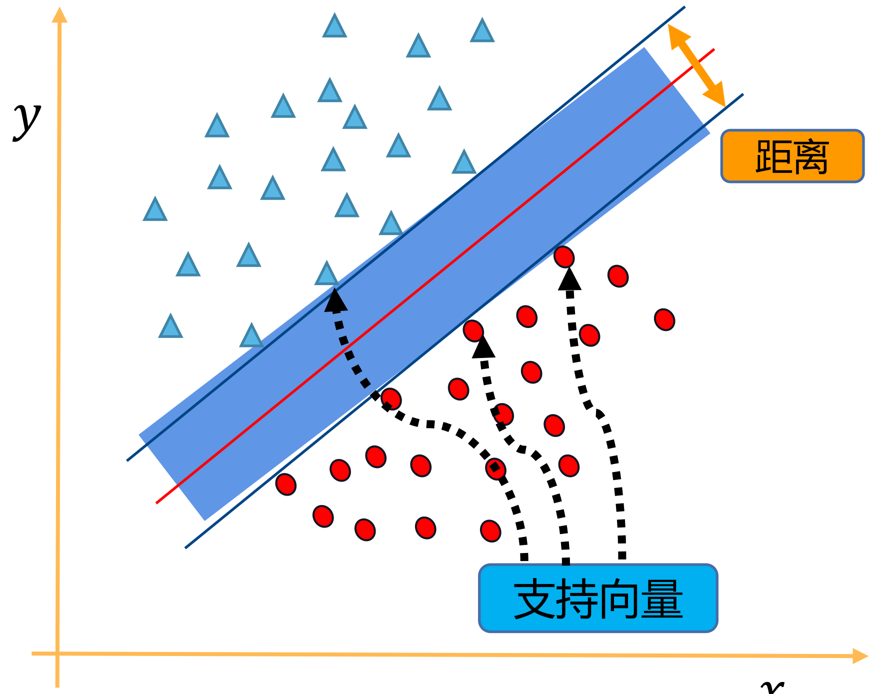
1.支持向量机概述

监督学习 (supervised learning)

对数据进行二元分类的广义线性分类器

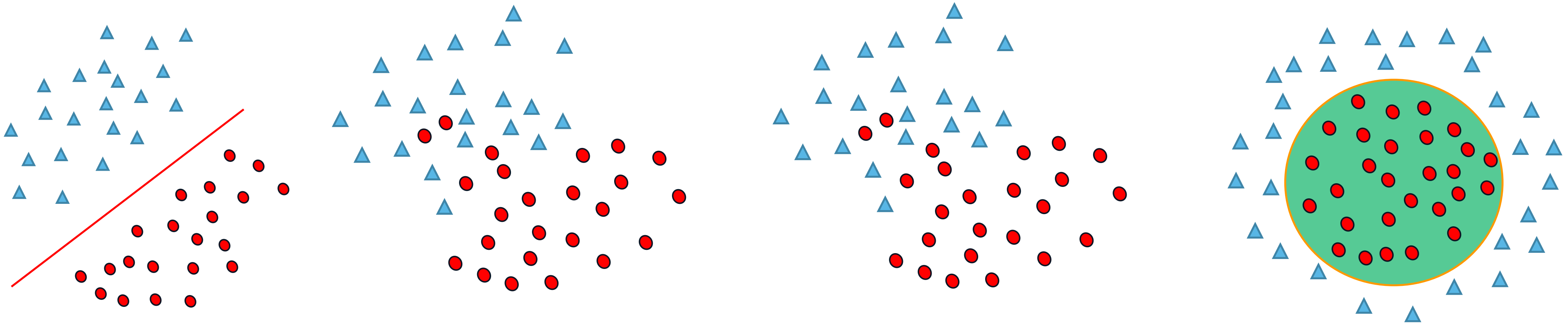
决策边界是对学习样本求解的最大边距超平面

(maximum-margin hyperplane)



1.支持向量机概述

硬间隔、软间隔和非线性 SVM



线性可分　　硬间隔　　软间隔　　线性不可分

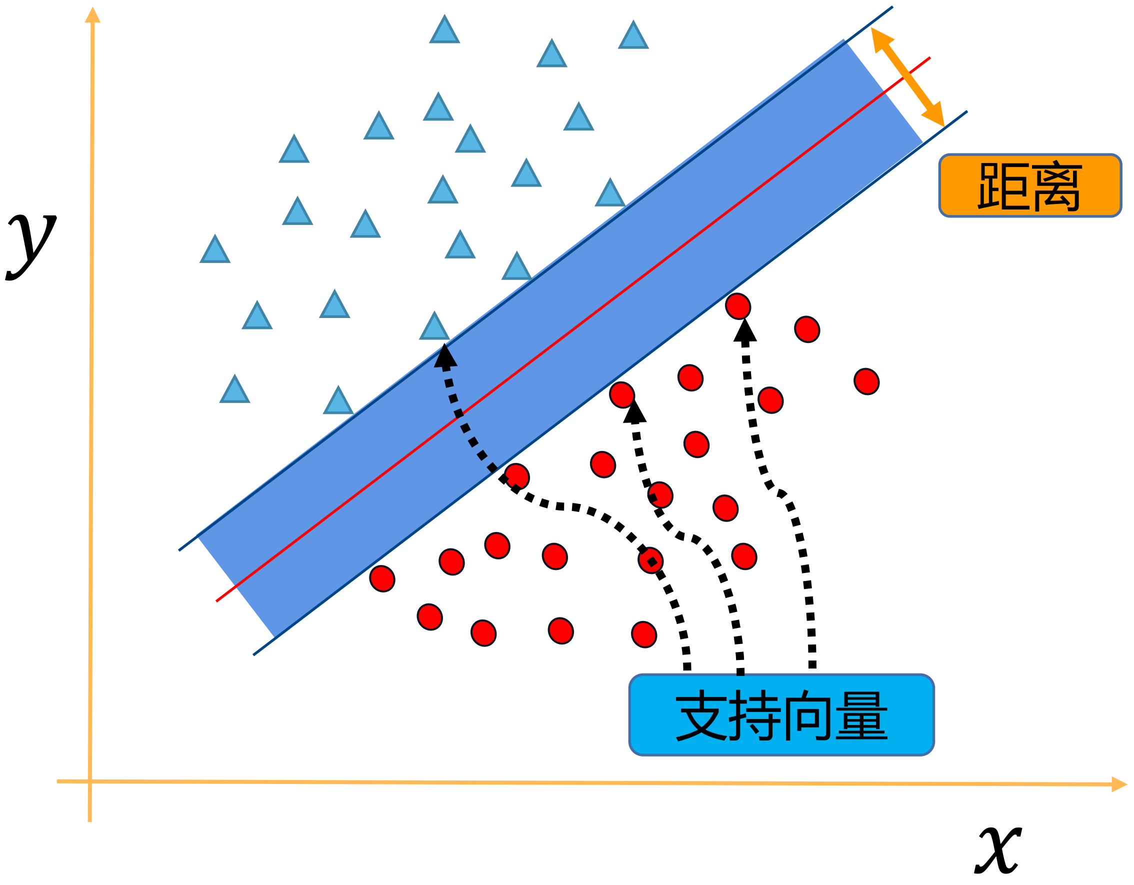
假如数据是完全的线性可分的，那么学习到的模型可以称为硬间隔支持向量机。

硬间隔指完全分类准确，软间隔允许一定量的样本分类错误。

1.支持向量机概述

算法思想

找到集合边缘上的若干数据（称为支持向量（Support Vector）），用这些点找出一个平面（称为决策面），使得支持向量到该平面的距离最大。



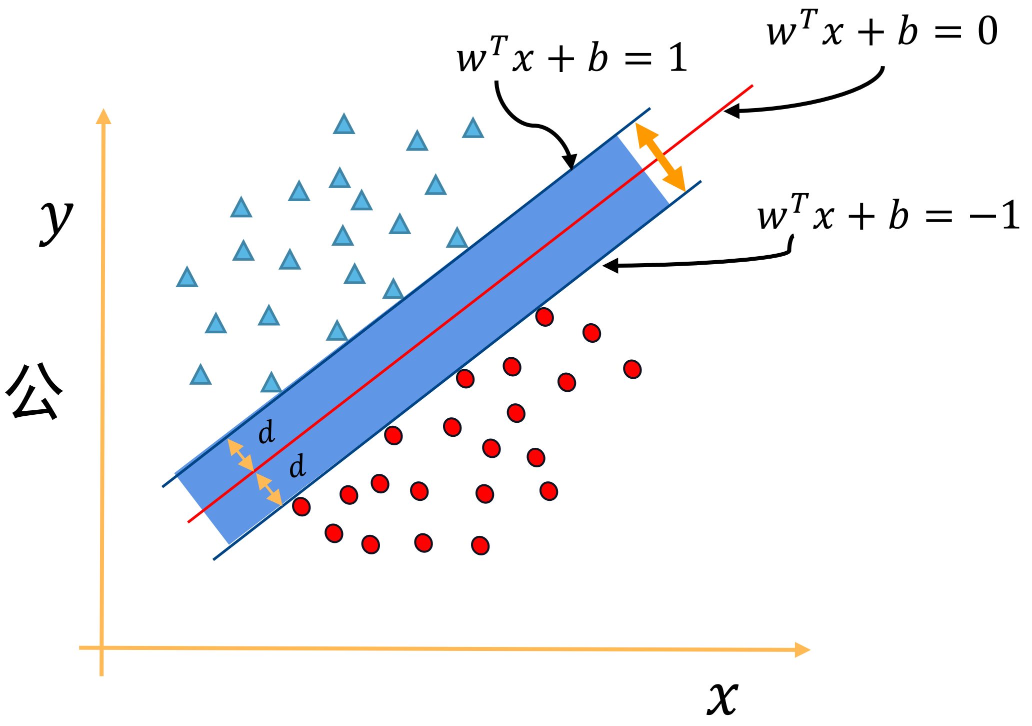
1.支持向量机概述

任意超平面可以用下面这个线性方程来描述：

$$w^T x + b = 0$$

二维空间点 (x, y)到直线$Ax + By + C = 0$的距离公式是：

$$\frac{|Ax + By + C|}{\sqrt{A^2 + B^2}}$$



扩展到n 维空间后，点 $x = (x_1, x_2 \dots x_n)$ 到超平面

$w^T x + b = 0$ 的距离为：$\frac{|w^T x+b|}{||w||}$

其中 $||w|| = \sqrt{w_1^2 + \cdots w_n^2}$

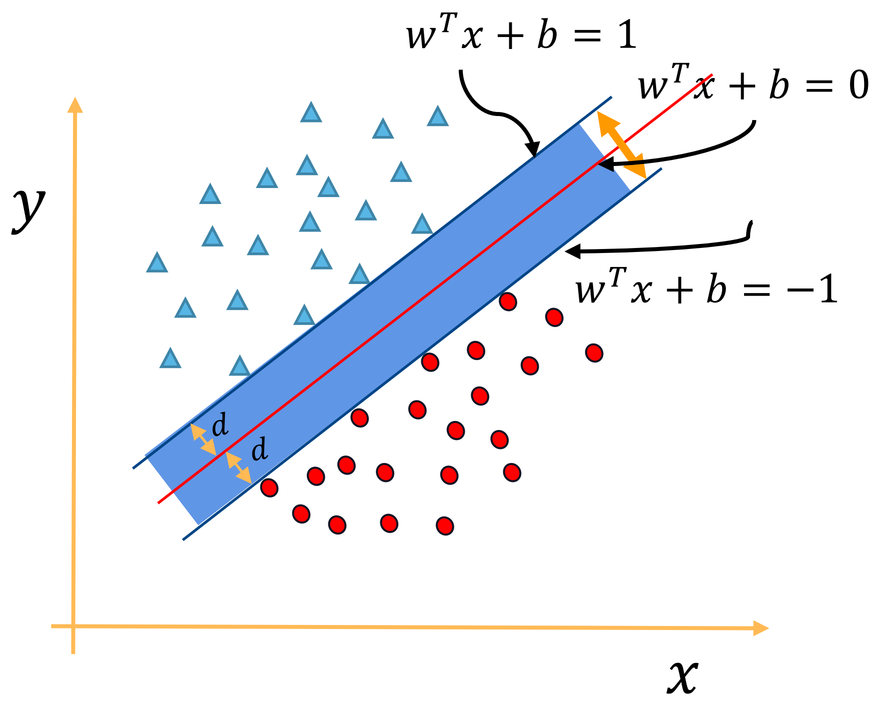
支持向量到超平面的距离：

$$d = \frac{|w^T x+b|}{||w||}$$

1.支持向量机概述

支持向量到超平面的距离为 d，其他点到超平面的距离大于 d：

$$\begin{cases} \dfrac{w^T x + b}{\| w \|} \geq d & y = 1 \\ \dfrac{w^T x + b}{\| w \|} \leq -d & y = -1 \end{cases}$$



令d为 1（方便推导和优化且不影响目标函数的优化），

将两个方程合并，可以简写为：

$$d = \frac{|w^T x + b|}{||w||}$$

得到最大间隔超平面的上下两个超平面：

$$y(w^T x + b) \geq 1$$

2.线性可分支持向量机

2.线性可分支持向量机

背景知识

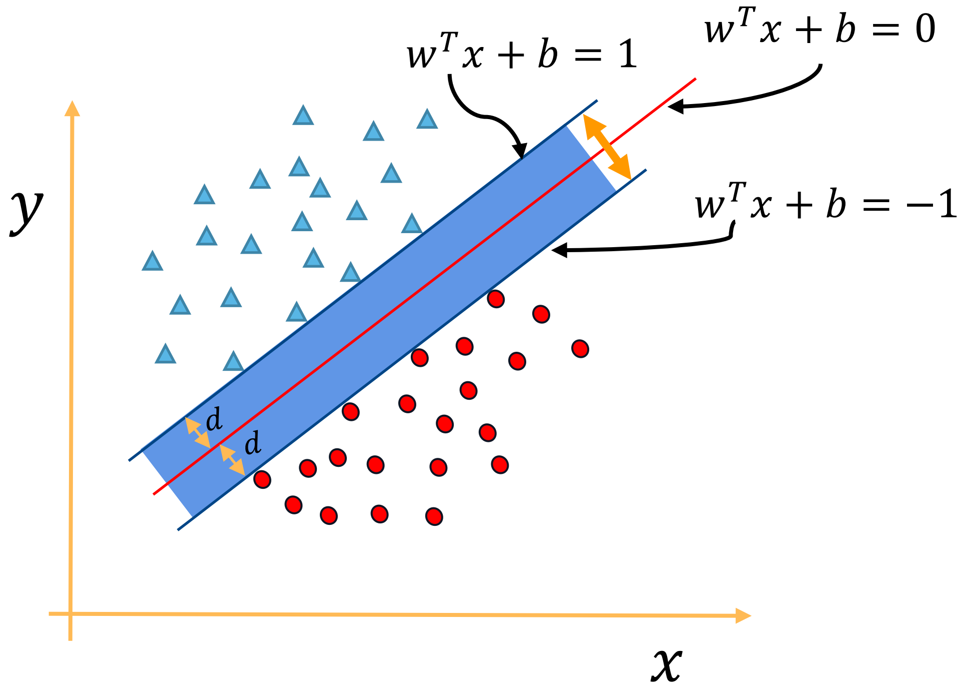
点到面的距离公式

$$d = \frac{|Ax_0 + By_0 + Cz_0 + D|}{\sqrt{A^2 + B^2 + C^2}}$$

$$y\left(w^T x + b\right) \geq 1$$

$$y\left(w^T x + b\right) = \left|w^T x + b\right| \qquad d = \frac{\left|w^T x + b\right|}{\|w\|}$$

支持向量机的最终目的是最大化d



2. 支持向量机求解

函数间隔：$d^* = y_i(w^T x + b)$

几何间隔：$d = \frac{y(w^T x+b)}{||w||}$，当数据被正确分类时，几何间隔就是点到超平面的距离；为了求几何间隔最大，SVM基本问题可转化为求解：($\frac{d^*}{||w||}$为几何间隔，d^*为函数间隔)

$$\max_{w,b} \frac{d^*}{\| w \|}$$

$$s.t. \ \ y_i \left(w^T x_i + b \right) \geq d^*, i = 1, 2, \ldots, m$$

2.支持向量机求解

①转化为凸函数：

先令$d^* = 1$，方便计算（参照衡量，不影响评价结果）

$$\max_{w,b} \frac{1}{\|w\|}$$

$$\text{s.t. } y_i\left(w^T x_i + b\right) \geq 1, i = 1, 2, \ldots, m$$

再将$\max\limits_{w,b} \frac{1}{||w||}$转化成$\min\limits_{w,b} \frac{1}{2}||w||^2$求解凸函数，1/2是为了求导之后方便计算。

$$\min_{w,b} \frac{1}{2}\|w\|^2$$

$$\text{s.t. } y_i\left(w^T x_i + b\right) \geq 1, i = 1, 2, \ldots, m$$

2.支持向量机求解

②用拉格朗日乘子法和KKT条件求解最优值：

$$\min_{w,b}\frac{1}{2}\|w\|^2$$

$$\text{s.t.}\quad -y_i\left(w^T x_i+b\right)+1\leq 0, i=1,2,\ldots,m$$

整合成：$L(w,b,\alpha)=\frac{1}{2}||w||^2+\sum_{i=1}^{m}\alpha_i\left(-y_i(w^T x_i+b)+1\right)$其中$\alpha$为拉格朗日乘子

推导:

根据Karush-Kuhn-Tucker（KKT）条件:

$$\frac{\partial}{\partial w}L(w,b,\alpha)=w-\sum_{i=1}^{n}\alpha_i y_i x_i=0, w=\sum_{i=1}^{m}\alpha_i y_i x_i$$

$$\frac{\partial}{\partial b}L(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i y_i=0$$

2.支持向量机求解

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i \quad \sum_{i=1}^{m}\alpha_i y_i=0$$

代入$L(w,b,\alpha)$

$$\min_{w,b} L(w,b,\alpha)=\frac{1}{2}\|w\|^2+\sum_{i=1}\alpha_i\left(-y_i\left(w^T x_i+b\right)+1\right)$$

$$=\frac{1}{2}w^T w-\sum_{i=1}^{m}\alpha_i y_i w^T x_i-b\sum_{i=1}^{m}\alpha_i y_i+\sum_{i=1}^{m}\alpha_i$$

$$=\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y_i x_i-\sum_{i=1}^{m}\alpha_i y_i w^T x_i+\sum_{i=1}^{m}\alpha_i$$

$$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i w^T x_i$$

$$=\sum_{i=1}^{m}\alpha_i-\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)$$

2.支持向量机求解

$$w=\sum_{i=1}^{m}\alpha_i y_i x_i \quad \sum_{i=1}^{m}\alpha_i y_i=0$$

再把max问题转成min问题：添加负号

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)=\min_{\alpha}\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\left(x_i\cdot x_j\right)-\sum_{i=1}^{m}\alpha_i$$

$$\text{s.t.}\ \sum_{i=1}^{m}\alpha_i y_i=0$$

$$\alpha_i\geq 0, i=1,2,\ldots,m$$

得到最优解： $\alpha^*=(\alpha_1^*,\alpha_2^*,\ldots,\alpha_m^*)^T$

解出后，代入超平面模型也就是：

$y=w^{*T}x+b^*=\sum_{i=1}^{m}\alpha_i^* y_i\left(x_i\cdot x_j\right)+b^*$，可得$b^*=y-\sum_{i=1}^{m}\alpha_i^* y_i\left(x_i\cdot x_j\right)$， $w^*=\sum_{i=1}^{m}\alpha_i^* y_i x_i$

以上为SVM对偶问题的对偶形式

3.线性支持向量机

3.线性支持向量机

若数据线性不可分，则可以引入松弛变量$\xi \geq 0$，使函数间隔

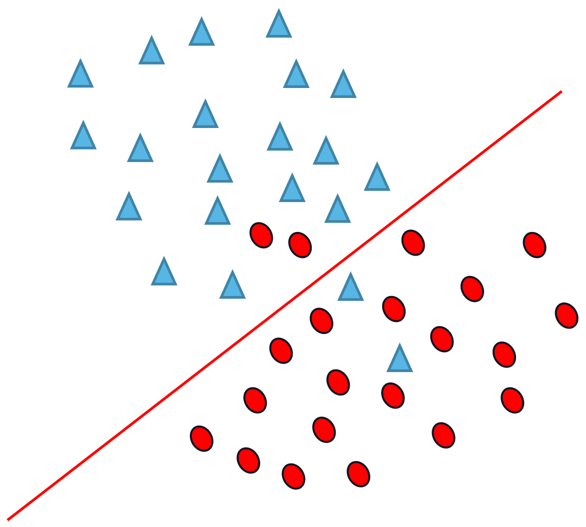
加上松弛变量大于等于1 ，则目标函数：

$$\min_{w,b,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{s.t. } y_i\left(w^T x_i + b\right) \geq 1 - \xi_i$$

对偶问题：

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \left(x_i \cdot x_j\right) = \min_{\alpha} \frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j \left(x_i \cdot x_j\right) - \sum_{i=1}^{m}\alpha_i$$

$$\text{s.t. } C \geq \alpha_i \geq 0, i = 1, 2, \ldots, m \quad \sum_{i=1}^{m}\alpha_i y_i = 0$$



软间隔

C为惩罚参数， C 值越大 ，对分类的惩罚越大。跟线性可分求解的思路一致，同样这里先用拉格朗日乘子法得到拉格朗日函数，再求其对偶问题。

3.线性支持向量机

ξ 为"松弛变量"

$$\xi_i = \max(0,1 - y_i(w \cdot x_i + b))$$

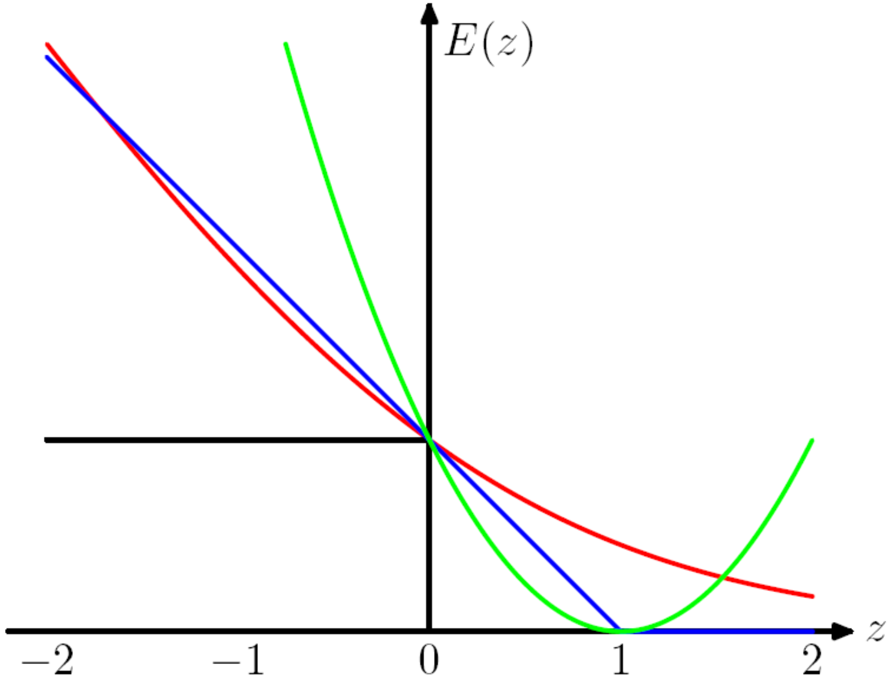
即hinge损失函数。每一个样本都有一个对应的松弛变量，表征该样本不满足约束的程度。

hinge损失：$\ell_{hinge}(z) = \max(0,1-z)$

指数损失：$\ell_{exp}(z) = \exp(-z)$

对率损失：$\ell_{log}(z) = \log(1+\exp(-z))$

绿色的线为 指数损失
蓝色的线为 hinge损失
红色的线为 对率损失
黑色的线为 0/1损失



损失函数图像

3.线性支持向量机

求解原始最优化问题的解w^*和b^*，得到线性支持向量机，其分离超平面为

$$w^{*T}x + b^* = 0$$

分类决策函数为：$f(x) = \mathrm{sign}(w^{*T}x + b^*)$

线性可分支持向量机的解w^*唯一，但b^*不唯一。对偶问题是

$$\min_{\alpha} \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j \left(x_i \cdot x_j\right) - \sum_{i=1}^{m}\alpha_i$$

$$\text{s.t.} \sum_{i=1}^{m}\alpha_i y_i = 0$$

$$0 \le \alpha_i \le C, \quad i = 1, 2, \cdots, m$$

3.线性支持向量机

解出后，代入超平面模型：

$$w^{*T}x + b^* = 0$$

可得

$$b^* = y - \sum_{i=1}^{m} \alpha_i^* y_i \left(x_i \cdot x_j \right)$$

$$w^* = \sum_{i=1}^{m} \alpha_i^* y_i x_i$$

其中：$0 < \alpha_i^* < C$

4.线性不可分支持向量机

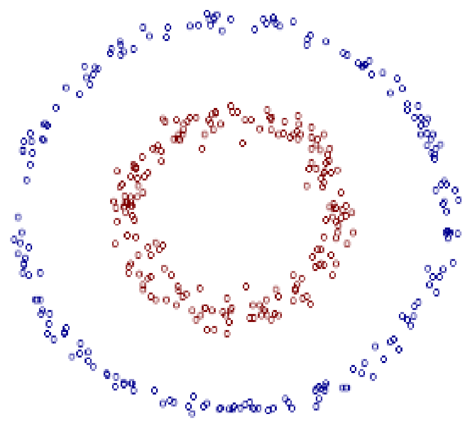
4.线性不可分支持向量机

核技巧

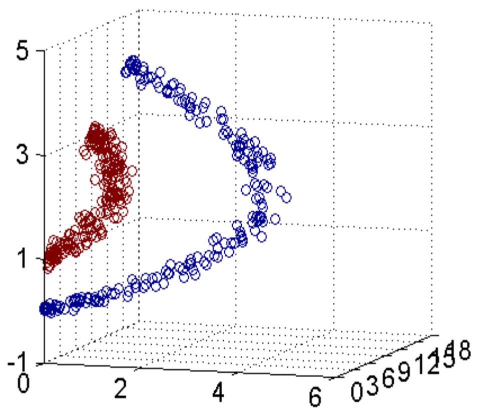
若不存在一个能正确划分两类样本的超平面，需要引入：**核函数**

将样本从原始空间映射到更高维的特征空间，使得样本在新的空间中线性可分

可以使用原来的推导计算，只是推导是在新的空间，用核函数替换当中的内积



线性不可分



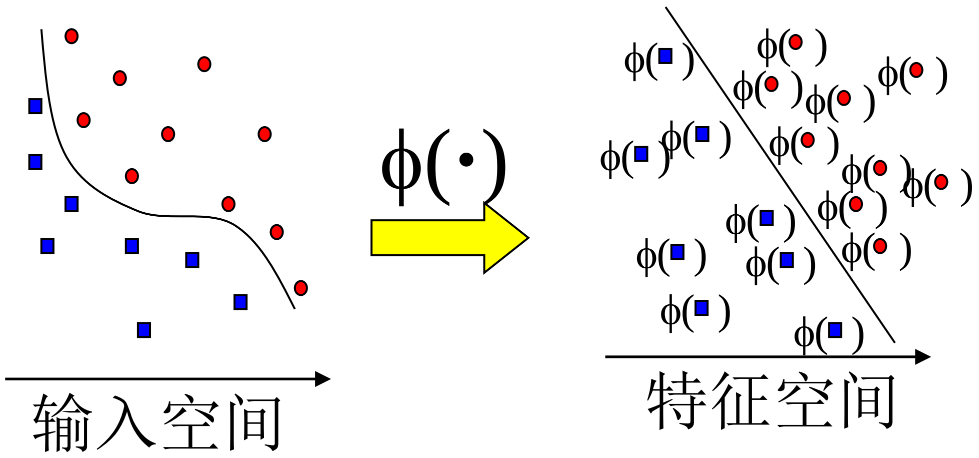
高维下线性可分

4.线性不可分支持向量机

令$\phi(x)$表示x映射后的特征向量，在特征空间中划分超平面的模型为：

$f(x) = w^T\phi(x)$+b

通过一个非线性转换后的两个样本间的内积



具体地，$k(x,z)$是一个核函数或正定核，存在一个从输入空间到特征空间的映射，对于任意空间输入的x,z 有：

$$k(x_i, z_i) = \phi(x_i)^T \cdot \phi(z_i)$$

4.线性不可分支持向量机

在线性支持向量机学习的对偶问题中，用核函数 $k(x,z)$ 替代内积，求解得到的是非线性支持向量机

$$\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j\phi(x_i)^T\phi(x_j)=\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}\alpha_i\alpha_j y_i y_j k(x_i,x_j)$$

$$\text{s.t. }\sum_{i=1}^{m}\alpha_i y_i=0$$

$$\alpha_i\geq 0, i=1,2,\ldots,m$$

求解后得到： $f(x)=w^T\phi(x)+b$

$$=\sum_{i=1}^{m}\alpha_i y_i\phi(x_i)^T\phi(x)+b$$

$$=\sum_{i=1}^{\mathrm{m}}\alpha_i\, y_i k\left(x,x_i\right)+b$$

4.线性不可分支持向量机

常用核函数有：

线性核函数

$$k(x_i, x_j) = x_i^T x_j$$

多项式核函数

$$k(x_i, x_j) = (x_i^T x_j)^d$$

高斯核函数

$$k(x_i, x_j) = exp(-\frac{||x_i - x_j||}{2\delta^2})$$

总结

一些SVM普遍使用的准则：

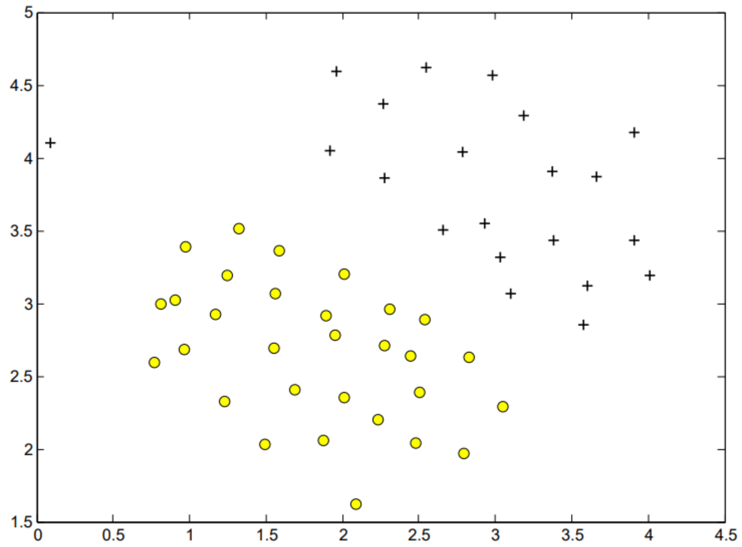
n为特征数，m为训练样本数。

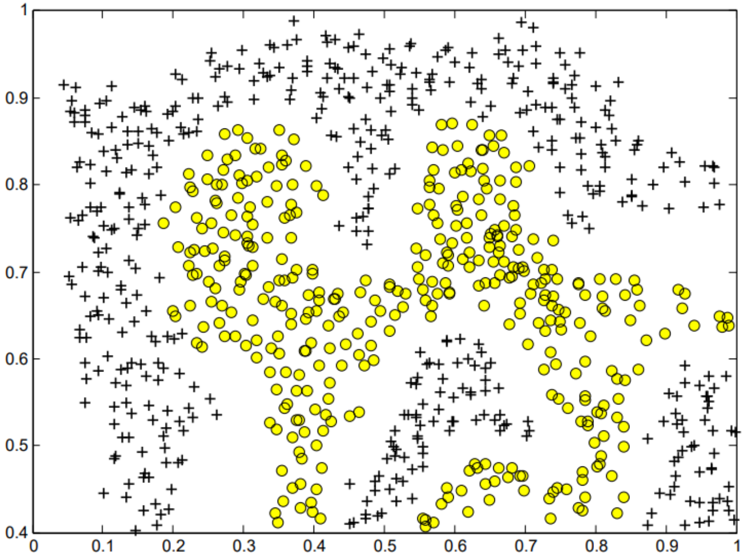
(1) 如果相较于m而言，n要大许多，即训练集数据量不够支持训练一个复杂的非线性模型，选用逻辑回归模型或者不带核函数的支持向量机。

(2) 如果n较小，而且m大小中等，例如n在 1-1000 之间，而m在10-10000之间，使用高斯核函数的支持向量机。

(3) 如果n较小，而m较大，例如n在1-1000之间，而m大于50000，则使用支持向量机会非常慢，解决方案是创造、增加更多的特征，然后使用逻辑回归或不带核函数的支持向量机。

作业4:基于SVM的垃圾邮件分类





- 推荐编程环境:Anaconda+Jupyter notebook

  安装教程: [点这]

- 用SVM解决线性可分和非线性可分数据集

Sources

- https://www.youtube.com/watch?v=bfmFfD2RIcg

  Neural Network In 5 Minutes | What Is A Neural Network? 5mins
- https://www.youtube.com/watch?v=CqOfi41LfDw

  Neural Networks Part. 1: Inside the Black Box 19mins
- https://www.youtube.com/watch?v=oJNHXPs0XDk

  Neural Network Architectures & Deep Learning 9mins
- https://www.youtube.com/watch?v=3JQ3hYko51Y

  Neural Network 3D Simulation 3mins
- https://www.youtube.com/watch?v=f0t-OCG79-U

  Convolutional Neural Network Visualization 2mins