有监督学习-支持向量机

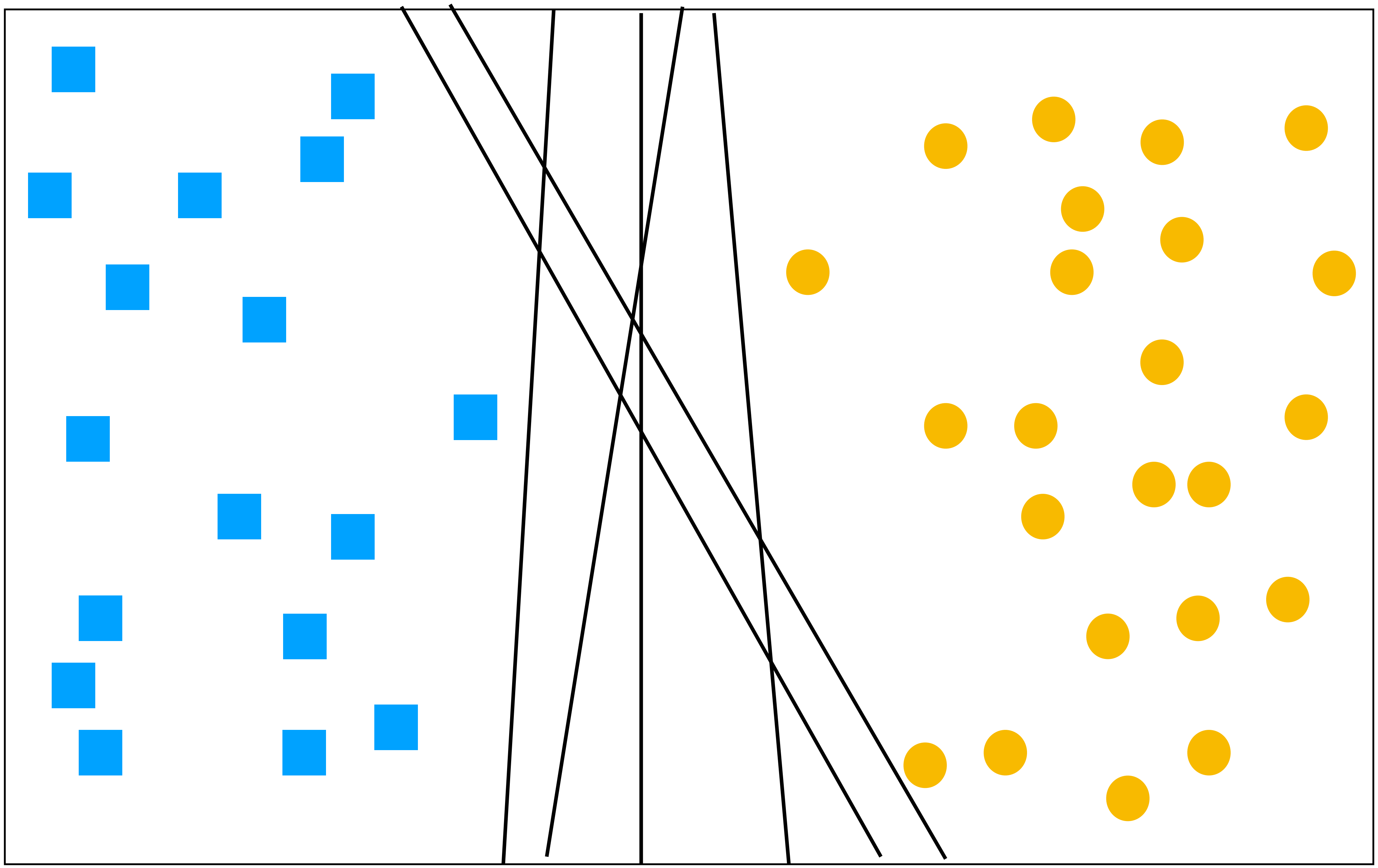
Supervised learning Support Vector Machine

支持向量机

支持向量机

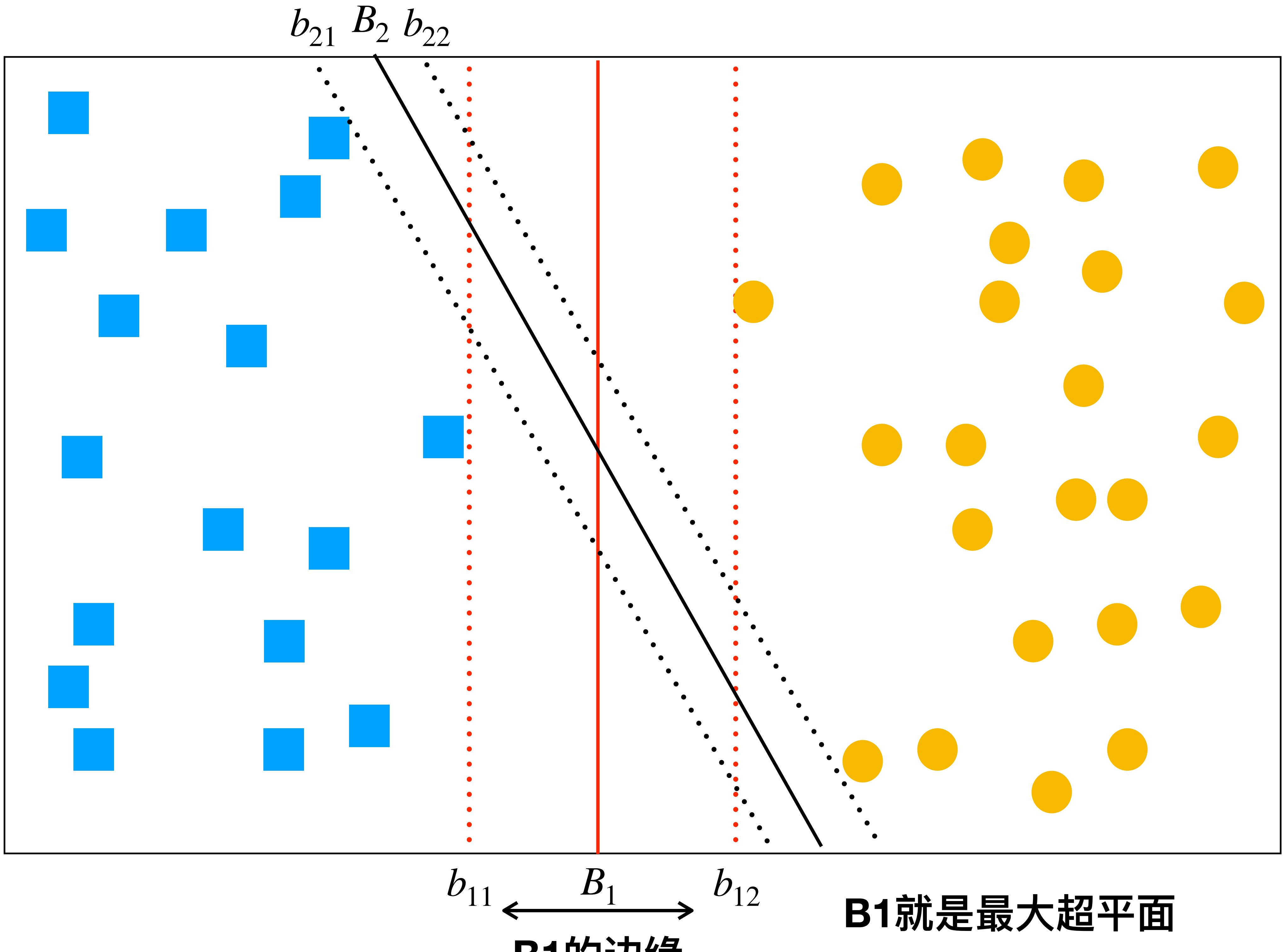
- 支持向量机(Support Vector Machine)具有坚实的统计学理论基础，并在许多实际应用（手写字体识别、文本分类等）中展示了大有可为的实践应用。

线性分类器



一个线性可分的数据集上的可能决策边界

线性分类器



最大超平面原理

- 具有较大边缘的决策边界比那些具有较小边缘的决策边界具有更好的泛化误差。

- 直觉上，较小的决策边缘，决策边缘任何轻微的扰动都可能对分类产生显著的影响，因此，决策边缘小的分类器对模型拟合更加敏感，泛化能力很差。

- 因此，需要设计最大化决策边界的线性分类器，以确保泛化误差最小。线性SVM（Linear SVM）就是这样的分类器。

线性支持向量机：可分情况

- 考虑一个包含N个训练样本D的二元分类器。
- 每个样本表示为 $(x_i, y_i)(i=1,2,...,N)$ ，其中 $x_i = (x_{i1}, x_{i2}, \ldots, x_{id})$ ，

  $y_i \in \{-1,1\}$
- 线性分类器的决策边界可以写为： $w \cdot x + b = 0$
- 则空间中任意一点到决策边界的距离可以写为

$$d = \frac{|wx + b|}{||w||}$$

线性支持向量机：可分情况

- 假设超平面（w, b）能将训练样本正确分类，则对于任意

  $(x_i, y_i) \in D$ 若 $y_i = 1$ 则有 $wx + b > 0$，若 $y_i = 0$ 则有 $wx + b < 0$

- 令 $\begin{cases} wx_i + b >= 1 & , y_i = 1 \\ wx_i + b <= -1 & , y_i = -1 \end{cases} <=> y_i(wx_i + b) >= 1$

- 距离超平面最近的这几个点使等号成立，他们称之为支持向量，所以两个异类支持向量到超平面的间隔为 $d = \dfrac{2}{||w||}$

- 我们的目标是找到最大的间隔，也就是最大化d

线性支持向量机：可分情况

- 最大化d，等价于最小化 $d = \frac{||w||^2}{2}$

- 所以，SVM的学习任务可以被形式化的描述为以下被约束的优化

$$\begin{cases} min\frac{||w||^2}{2} \\ y_i(wx_i + b) >= 1 \end{cases} \quad i = 1,2,...,N$$

- 在SVM中我们采用拉格朗日算子来进行优化

拉格朗日算子

- 拉格朗日算子基本型

$$\begin{cases} min_x f(x) \\ s.t.\ h_i(x)=0, i=1,2,...,m \end{cases} \ldots\ldots(1)$$

- 拉格朗日乘子法做的就是将约束条件添加到目标函数当中，使其变成一个无约束优化问题。可以这么做的原因是，只要满足hj(x)=0，那么不管加多少个都是不改变目标值的。原问题转换为：

$$min_{x,\lambda} = f(x) + \sum_{i=1}^{m} \lambda_i h_j(x) \ldots\ldots(2)$$

拉格朗日算子

- 假如x有p个特征，则(2)的最优解满足

$$\begin{cases} \frac{\partial L}{x_k} = \frac{\partial}{x_k} + \sum_{i=1} m\lambda_i \frac{\partial h_i(x)}{\partial x_k} = 0, k = 1,2,...,p \\ \frac{\partial L}{\lambda_i} = 0, i = 1,2,...,m \end{cases} \ldots\ldots (3)$$

- 当约束条件包含不等式的时候，我们可以使用KTT条件来求最优解，KTT条件是对拉格朗日算子的一个扩张。其问题可以描述为：

$$\begin{cases} min_x f(x) \\ s.t.\ h_i(x) = 0, i = 1,2,...,m \\ g_j(x) <= 0, j = 1,2,...,n \end{cases} \ldots\ldots (4)$$

KTT条件

- 在KTT条件下，(4)可以转换为

$$min_{x,\lambda,\mu} = f(x) + \sum_{i=1}^{m} \lambda_i h_j(x) + \sum_{j=1}^{n} \mu_j g_j(x) \ldots\ldots (5)$$

- KKT条件认为最优解满足条件

$$\begin{cases} \dfrac{\partial L}{x_k} = 0, k = 1,2,\ldots,p \\ \dfrac{\partial L}{\lambda_i} = 0, i = 1,2,\ldots,m \\ \mu_j g_j(x) = 0, j = 1,2,\ldots,n \\ \mu_j >= 0, j = 1,2,\ldots,n \end{cases} \ldots\ldots (6)$$

线性支持向量机：可分情况

- SVM的优化问题是满足KTT条件的，将

$$\begin{cases} minx\frac{||w||^2}{2} \\ y_i(wx_i+b)>=1 \end{cases} \ldots\ldots(7)$$

- 转换为 $$minL=\frac{1}{2}||w||^2-\sum_{i=1}^{N}\lambda_i(y_i(wx_i+b)-1)\ldots\ldots(8)$$

- 满足优化条件

$$\begin{cases} \frac{\partial L}{w}=w-\sum_{i=1}^{N}\lambda_i y_i x_i=0 \\ \frac{\partial L}{b}=\sum_{i=1}^{N}\lambda_i yi=0 \\ \lambda_i(y_i f(x_i)-1))=0 \\ \lambda_i>=0 \end{cases} \quad i=1,2,...,N\ldots\ldots(9)$$

线性支持向量机：可分情况

- 将（9）中的前两项带入（8），消去w与b转换为

$$minL = \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^t x_j - \sum_{i=1}^{N}\lambda_i \ldots\ldots (9)$$

- 仍然需要满足

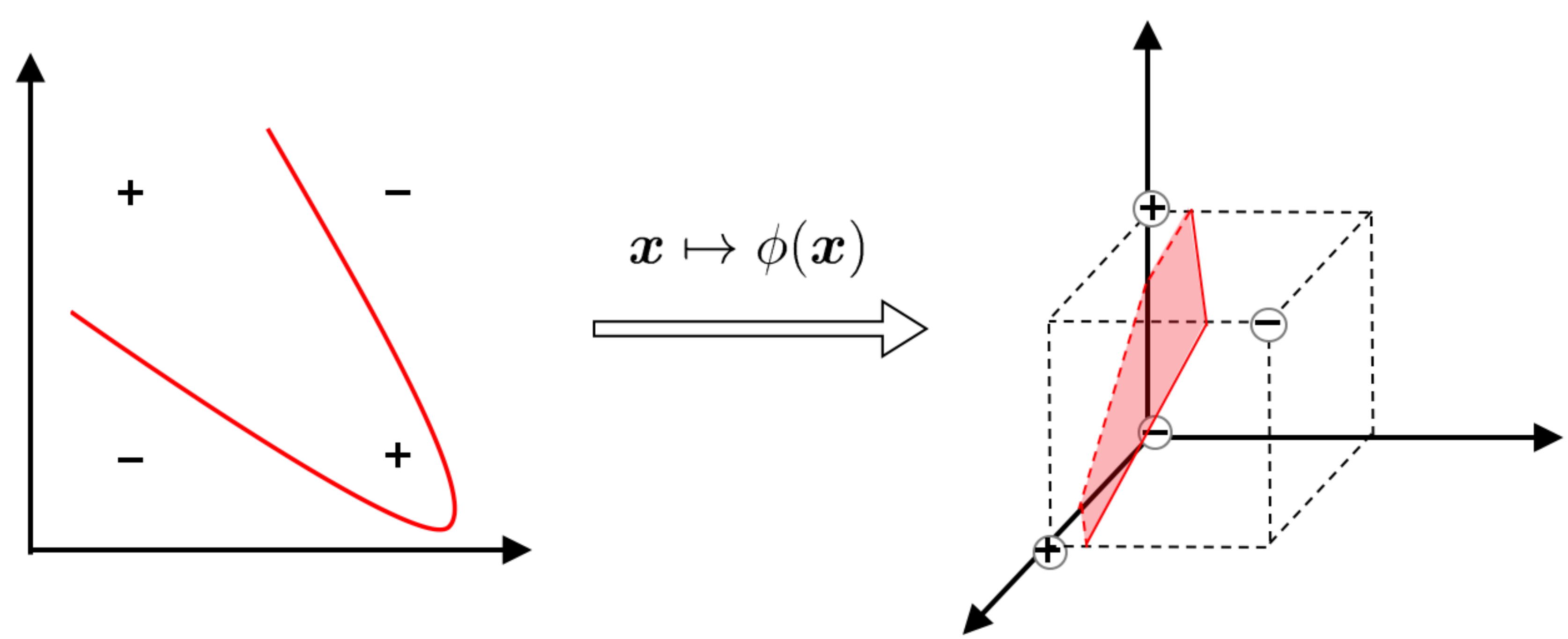
$$\begin{cases} \sum_{i=1}^{N}\lambda_i y_i = 0 \\ \lambda_i(y_i f(x_i) - 1)) = 0 \\ \lambda_i > = 0 \end{cases} i = 1,2,...,N \ldots\ldots (10)$$

- 通过SMO方法对(9)求解

线性不可分

- 现实问题中，很难出现线性可分的情况，那么如果不存在一个超平面可以将两类样本正确的划分怎么办?

- 将原始样本投影到更高维的空间中，使得样本在高维空间中变得线性可分



线性不可分

- 通过 ϕ将x映射到更高的空间中，则在新的空间中，超平面所对应的模型是

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

- 对于 $\begin{cases} min x \frac{||w||^2}{2} \\ y_i(wx_i + b) >= 1 \end{cases} i = 1,2,\ldots,N$ 我们可以变换为

$$\begin{cases} min x \frac{||w||^2}{2} \\ y_i(w\phi(x_i) + b) >= 1 \end{cases} i = 1,2,\ldots,N$$

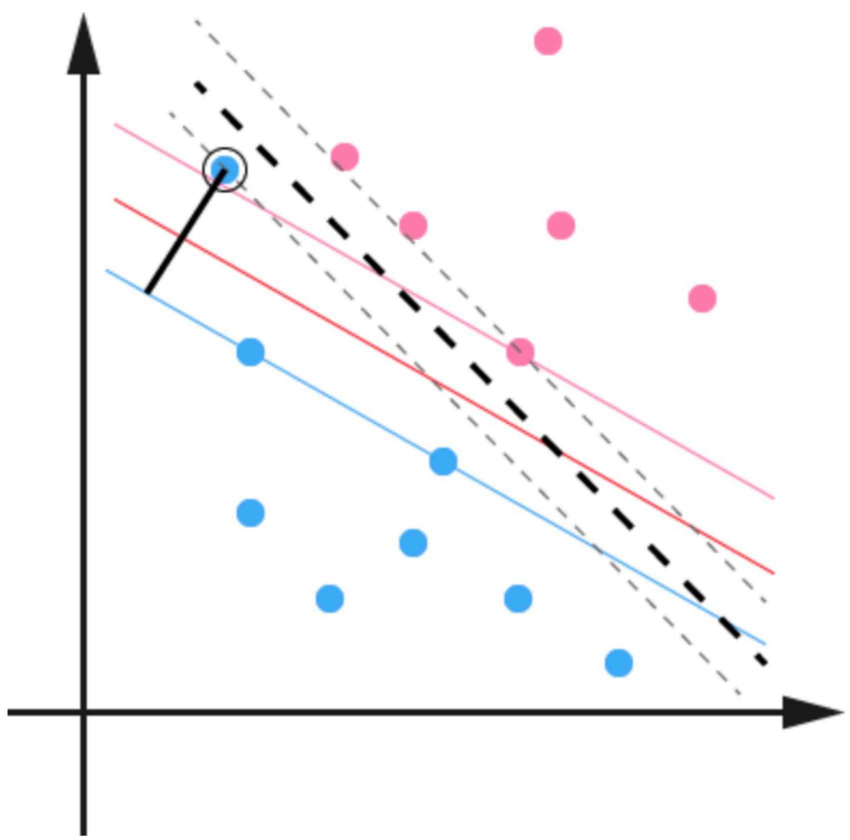
- 同样可以使用拉格朗日算子进行求解

常用的核函数

名称	表达式	参数
线型核	$\kappa(x_i, x_j) = x_i^T x_j$	
多项式核	$\kappa(x_i, x_j) = (x_i^T x_j)^d$	$d \geqslant 1$ 为多项式次数
高斯核	$\kappa(x_i, x_j) = \exp(-\frac{\\|x_i - x_j\\|^2}{2\sigma^2})$	$\sigma > 0$为高斯核的带宽
拉普拉斯核	$\kappa(x_i, x_j) = \exp(-\frac{\\|x_i - x_j\\|}{\sigma})$	$\sigma > 0$
Sigmoid核	$\kappa(x_i, x_j) = \tanh(\beta x_i^T x_j + \theta)$	$\beta > 0, \theta < 0$ tanh为双曲正切函数

松弛变量

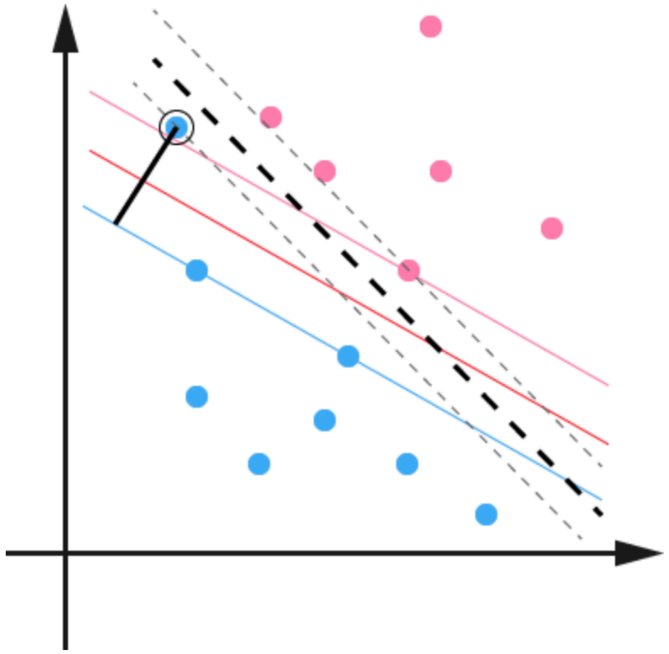
- 现实问题中，有非常大的可能即使使用了核函数，仍然是不可分的状态，或者即使可分，但是泛化误差非常高，并不是我们想要的。例如下图，这个时候我们要引入松弛变量，也就是允许某些点是可以被错误分类的



松弛变量

- 我们对这种偏移的点添加向回”拉”一些，让他返回到原有的超平面上。

- 对于这些离群点有对应的松弛变量，其他的点是没有松弛变量的。



松弛变量

- 我们将松弛变量记为 $\xi(\xi_i >= 0)$
- 则原约束条件变换为

$$\begin{cases} wx_i + b >= 1 - \xi & , y_i = 1 \\ wx_i + b <= -1 + \xi & , y_i = -1 \end{cases} <=> y_i(wx_i + b) >= 1 - \xi$$

- 原问题转换为

$$\begin{cases} minx \frac{||w||^2}{2} + C\sum_{i=1}^{N} \xi_i \\ y_i(wx_i + b) >= 1 - \xi_i \\ \xi_i >= 0 \end{cases} \quad i = 1,2,...,N$$

松弛变量

- C为惩罚项
  - 如果C为无穷大的时候，就会变成硬间隔（没有松弛变量的状态），因为为了使整体求最小 ξ_i 只能无限趋近于0
  - C很小的情况下，位于两条线之间错分的样本变多，对样本的拟合能力下降，容易出现前拟合的状态

实践

实践

- 使用支持向量机进行人脸识别