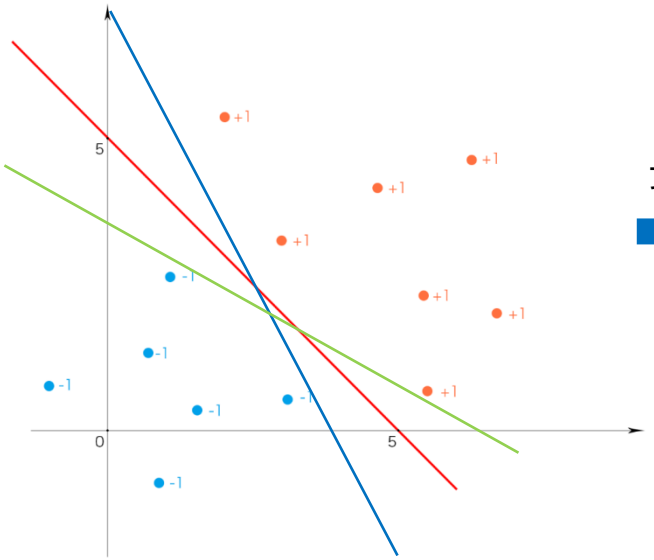
支持向量机(SVM)

01

PART ONE

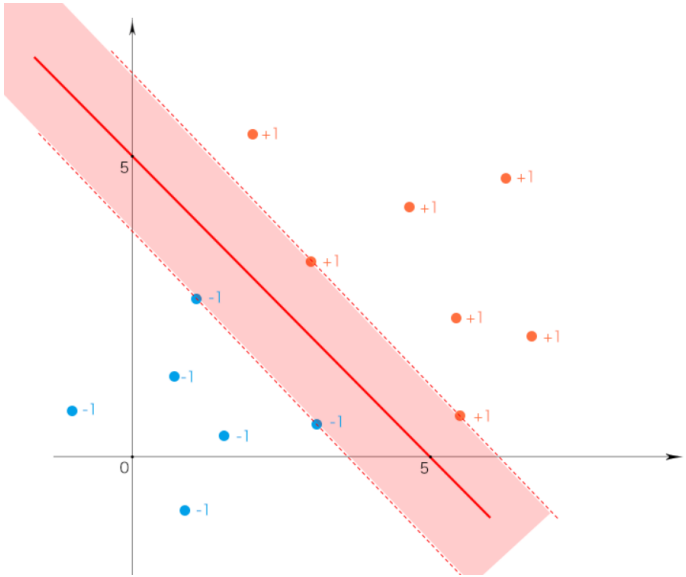
SVM基本原理

最大间隔



最好的一条

最大间隔



普通的支持向量机就是一条直线（超平面），用来完美划分两类。但这又不是一条普通的直线，这是无数条可以分类的直线当中最完美的，因为它恰好在两个类的中间，距离两个类的点都一样远。

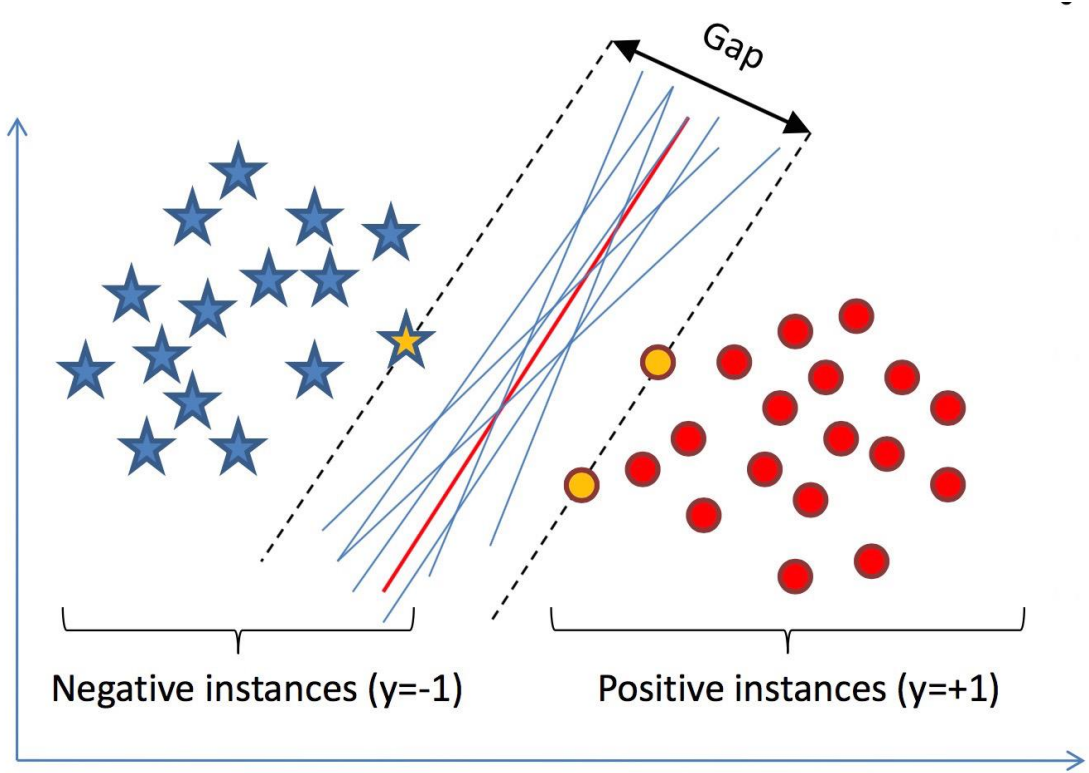
支持向量

所谓的“**支持向量**”就是这些离分界线最近的点。

通过“支持向量”来确定的这个完美的分类决策面，这样就叫**支持向量机**。

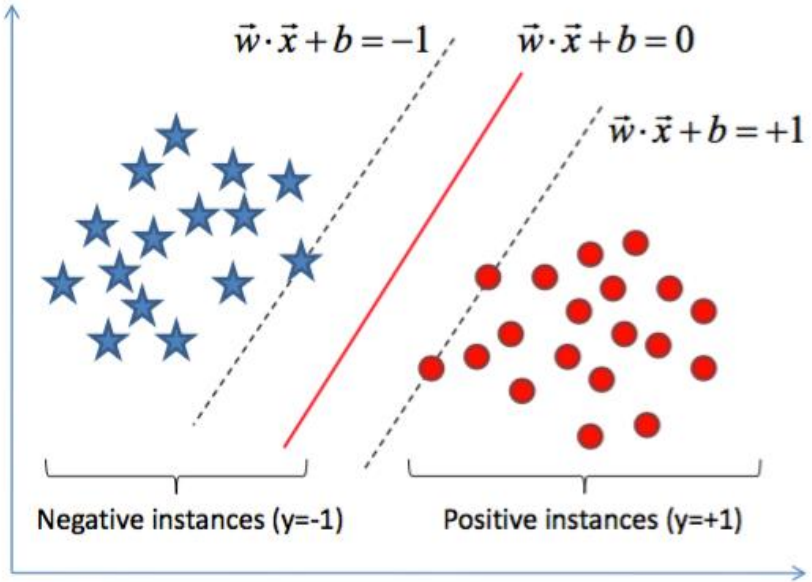
$w^T x + b = 0$：决策面

$$y = \begin{cases} 1, \text{如果} w^T x + b > 0 \\ -1, \text{如果} w^T x + b < 0 \end{cases}$$

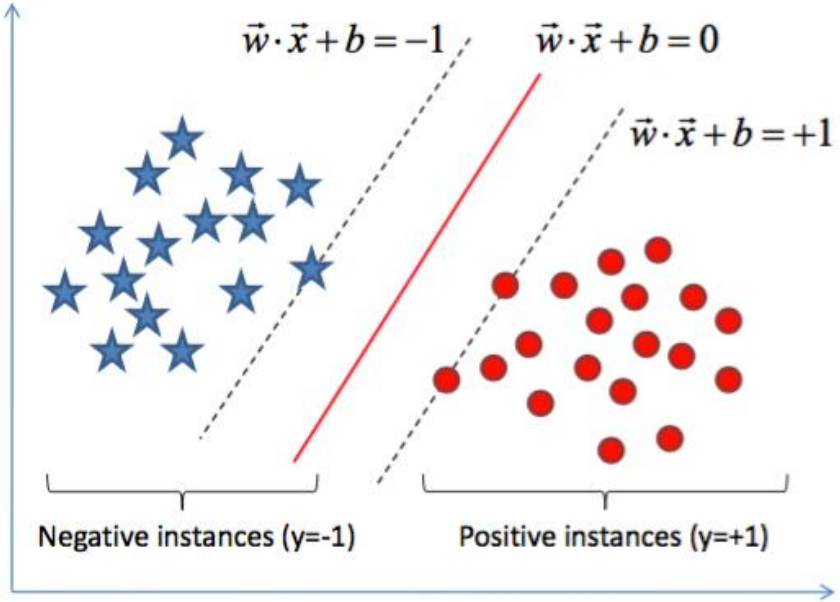


计算间隔

- 固定w的方向，可以找到一个超平面： $w^T x + b_{+1} = 0$，使得所有正样本都在平面的一侧；
- 同样可以找到一个超平面： $w^T x + b_{-1} = 0$，使得所有负样本都在平面的一侧；
- 调整w的大小，我们一定可以得到$b_{+1} - b_{-1} = 2$，因此重新取$b = b_{+1} - 1$
- 因此得到图中的三个平面，中间的平面可以用来预测新数据。



最大间隔



The gap is distance between parallel hyperplanes:

$\vec{w}\cdot\vec{x}+b=-1$ and
$\vec{w}\cdot\vec{x}+b=+1$

Or equivalently:

$\vec{w}\cdot\vec{x}+(b+1)=0$
$\vec{w}\cdot\vec{x}+(b-1)=0$

We know that

$D=|b_1-b_2|/\|\vec{w}\|$

Therefore:

$D=2/\|\vec{w}\|$

最大间隔： $\max\frac{2}{\|w\|}$ $\Longleftrightarrow$ $\min\frac{\|w\|^2}{2}$

转化后，我们的问题成为了一个凸优化问题。

最优化目标

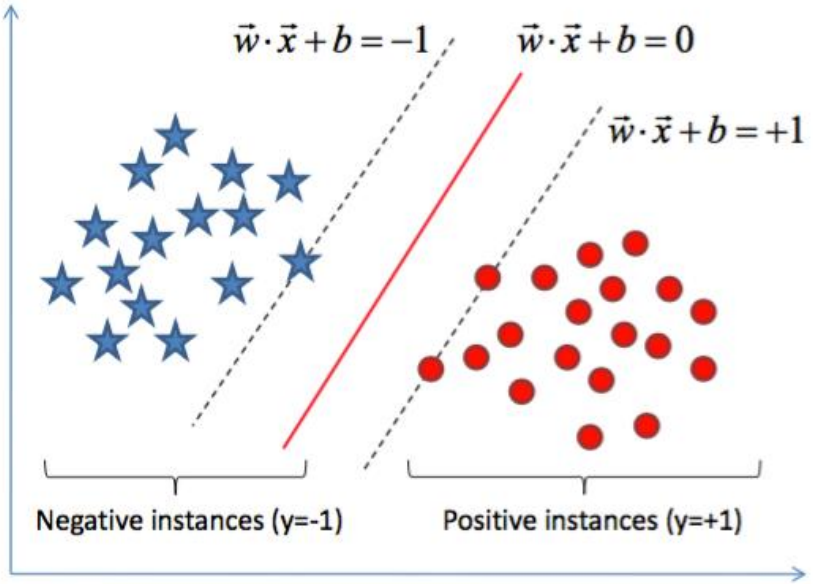
- 最小化：

$$\min \frac{\|w\|^2}{2}$$

$$\vec{w}\cdot\vec{x}_i + b \le -1 \text{ if } y_i = -1$$
$$\vec{w}\cdot\vec{x}_i + b \ge +1 \text{ if } y_i = +1$$

Equivalently:

$$y_i(\vec{w}\cdot\vec{x}_i + b) \ge 1$$



- 约束条件：

$$y_i(w^T x_i + b) \ge 1, i = 1,2,\cdots,n$$

- 综合起来表示：

$$\min \frac{\|w\|^2}{2} \quad s.t., y_i(w^T x_i + b) \ge 1, i = 1,2,\cdots,n$$

问题求解

$$\min_{w,b}\theta(w) = \min\frac{\|w\|^2}{2} \quad s.t., y_i(w^T x_i + b) \geq 1, i = 1,2,\cdots,n$$

- 拉格朗日乘子

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

$$\theta(w) = \max_{\alpha_i \geq 0} L(w,b,\alpha)$$

$$\min_{w,b}\theta(w) = \min_{w,b}\max_{\alpha_i \geq 0} L(w,b,\alpha)$$

- 对偶问题

对偶问题

$$\max_{\alpha_i \geq 0}\min_{w,b} L(w,b,\alpha)$$

问题求解

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

对L(w,b,a)分别对w,b求偏导，求极小值：

$$\frac{\partial\mathcal{L}}{\partial w}=0\Rightarrow w=\sum_{i=1}^{n}\alpha_i y_i x_i$$

如果得到了α的值就可以计算w

$$\frac{\partial\mathcal{L}}{\partial b}=0\Rightarrow\sum_{i=1}^{n}\alpha_i y_i=0$$

得到了α的约束条件

继续求解

$$
\begin{aligned}
\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha}) &= \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{i=1}^{n} \alpha_i \left[y_i \left(\boldsymbol{w}^{\mathrm{T}} \boldsymbol{x}_i + b \right) - 1 \right] \\
&= \frac{1}{2} \boldsymbol{w}^{\mathrm{T}} \boldsymbol{w} - \boldsymbol{w}^{\mathrm{T}} \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i - b \sum_{i=1}^{n} \alpha_i y_i + \sum_{i=1}^{n} \alpha_i \\
&= \frac{1}{2} \boldsymbol{w}^{\mathrm{T}} \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i - \boldsymbol{w}^{\mathrm{T}} \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i - b \cdot 0 + \sum_{i=1}^{n} \alpha_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \left(\sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \right)^{\mathrm{T}} \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i \\
&= \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^{\mathrm{T}} \boldsymbol{x}_j
\end{aligned}
$$

重新定义问题

- 最大化：

$$\sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j$$

- s.t. $\alpha_i \geq 0, i = 1,2...n$

$$\sum_{i=1}^{n} a_i y_i = 0$$

- 最小化：

$$\frac{1}{2} \sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n} \alpha_i$$

- s.t. $\alpha_i \geq 0, i = 1,2...n$

$$\sum_{i=1}^{n} a_i y_i = 0$$

- 输出 W', b'

$$w' = \sum_{i=1}^{n} \alpha_i y_i x_i$$

$$b' = y_j - w' x_j = y_j - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_j \,, \alpha_j \neq 0$$

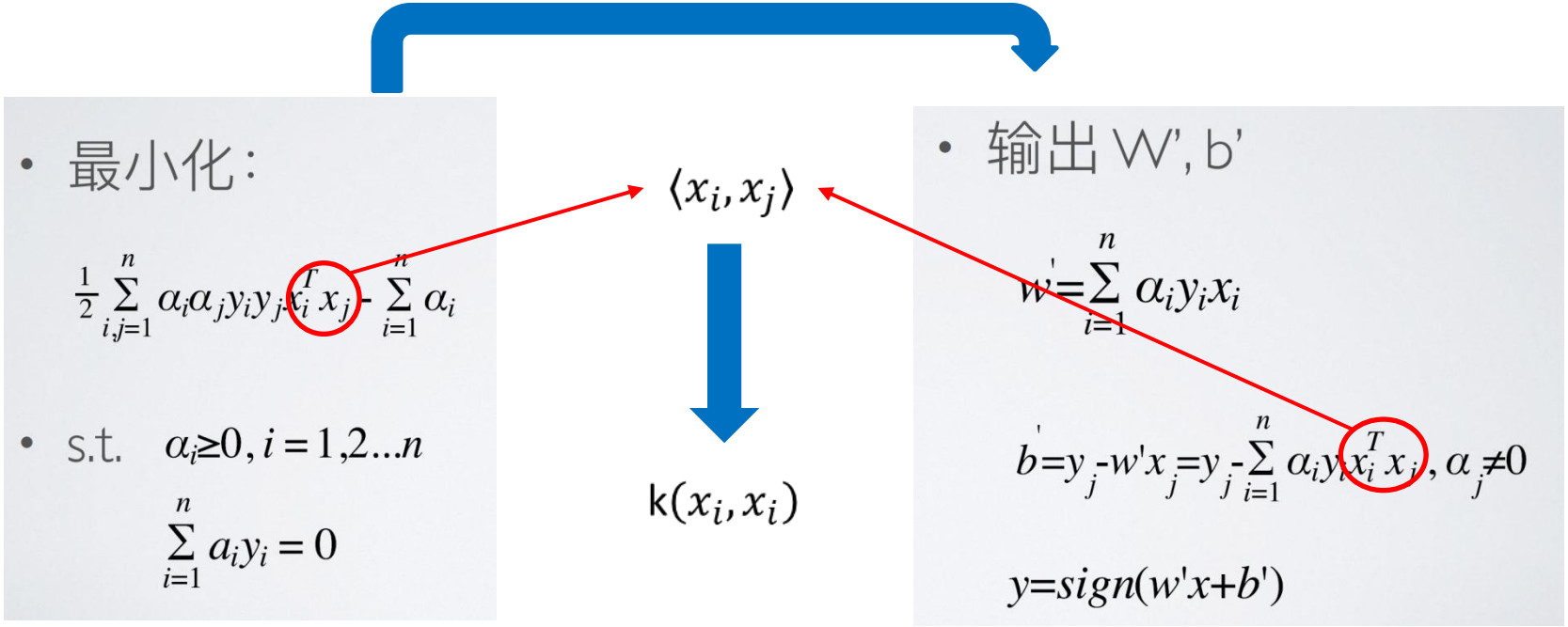
$$y = sign(w'x + b')$$

02

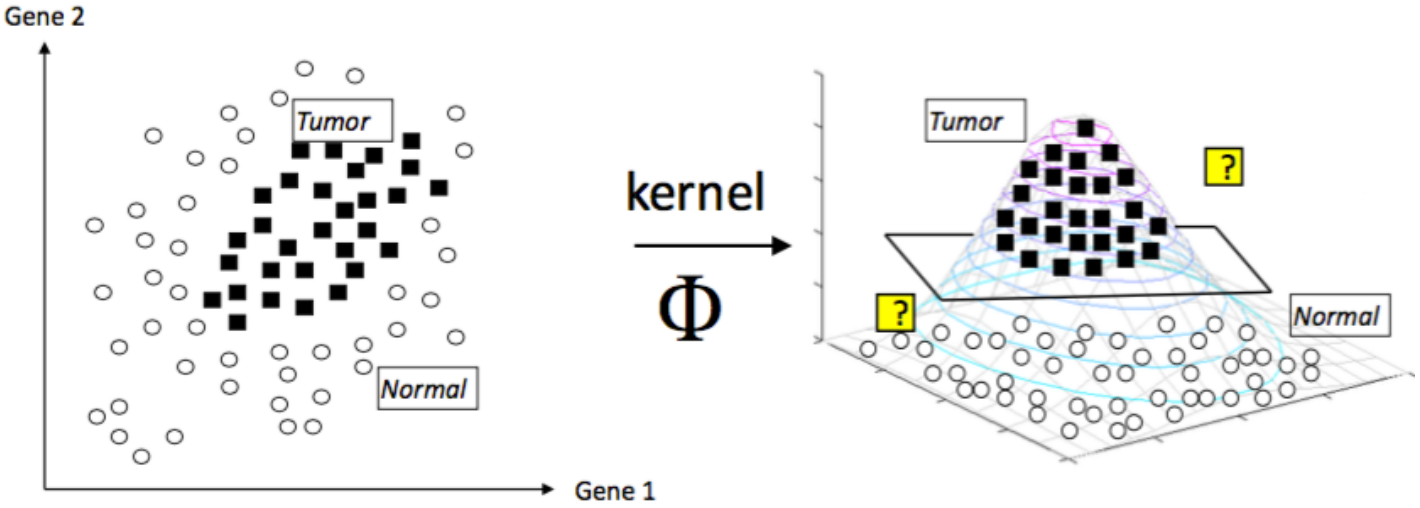
PART TWO

核方法

引入核方法


• 最小化：
$\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{n}\alpha_i$
• s.t. $\alpha_i \geq 0, i = 1,2...n$
$\sum_{i=1}^{n} a_i y_i = 0$
$\langle x_i, x_j \rangle$
$k(x_i, x_i)$
• 输出 W', b'
$w' = \sum_{i=1}^{n} \alpha_i y_i x_i$
$b' = y_j - w'x_j = y_j - \sum_{i=1}^{n} \alpha_i y_i x_i^T x_j, \alpha_j \neq 0$
$y = sign(w'x + b')$

通过核函数提升空间维度



Data is not linearly separable in the input space

Data is linearly separable in the feature space obtained by a kernel

$$\Phi : \mathbf{R}^N \rightarrow \mathbf{H}$$

常见核函数

通常人们会从一些常用的核函数中选择，根据问题和数据的不同，选择不同的参数，实际上就是得到了不同的核函数。

- 多项式核：

$$k(x_1, x_2) = \left(x_1^T x_2 + 1\right)^m$$

- 多项式核：会将原始空间映射为无穷维空间

σ太大会衰减比较快，相当于低维空间

σ太小可能会出现严重过拟合

$$\kappa(x_1, x_2) = \exp\left(-\frac{\|x_1 - x_2\|^2}{2\sigma^2}\right)$$

- 线性核：就是线性分类，主要是为了和前面的进行形式统一。

$$\kappa(x_1, x_2) = \langle x_1, x_2 \rangle$$

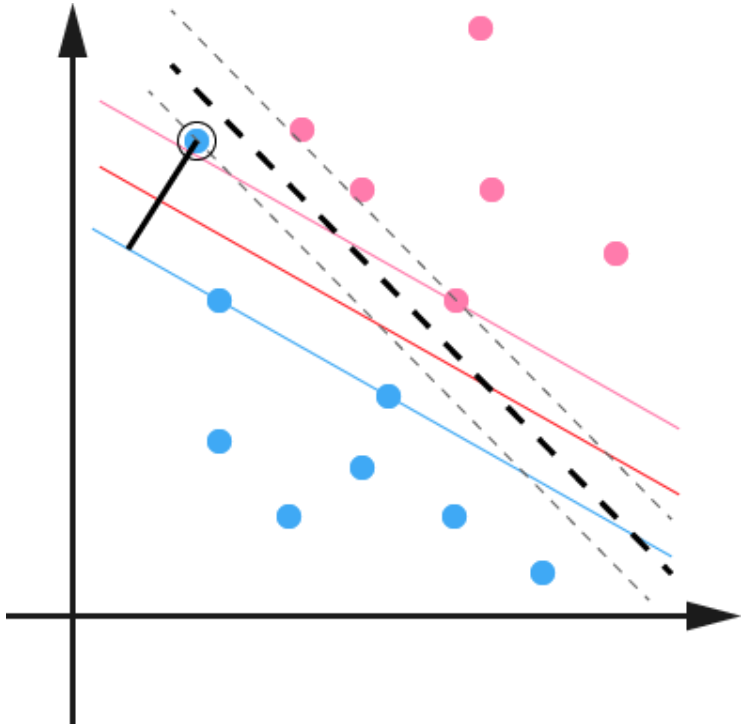
03

PART THREE

松弛因子

异常点的问题

- 可能因为数据有噪音，对于这种偏离正常位置很远的数据点，我们称之为 异常点。
- 在SVM 模型里，异常点的存在有可能造成很大的影响，因为超平面本身就是只有少数几个支持向量组成的。
- 如果这些支持向量里又存在 outlier 的话，其影响就很大了。

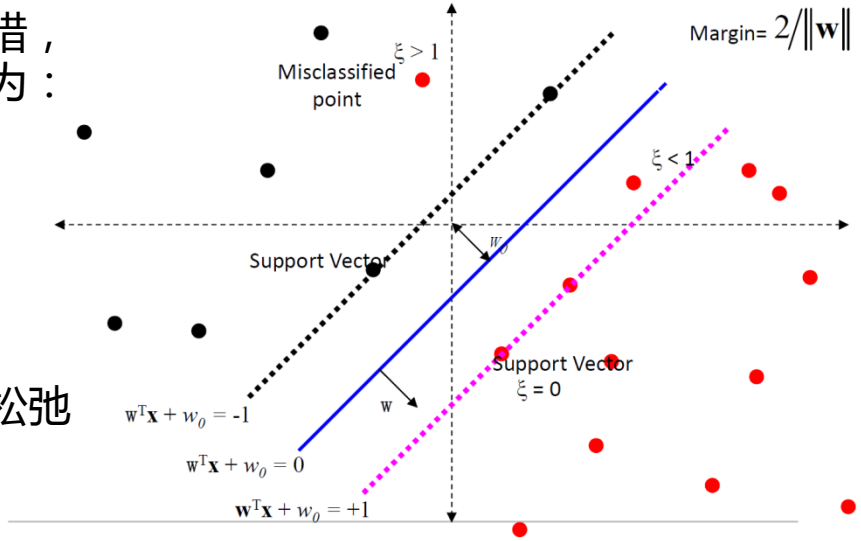


松弛变量

- 在实际问题中，总会出现异常点，数据不一定线性可分。
- 因此解决方案引入软间隔允许一些样本出错，即允许某些样本不满足约束，将约束放松为：

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

松弛变量



- 但不满足约束的样本越少越好，因此增加松弛量限制：

$$\xi_i \geq 0$$

$$\sum_i \xi_i \leq C$$

多分类问题

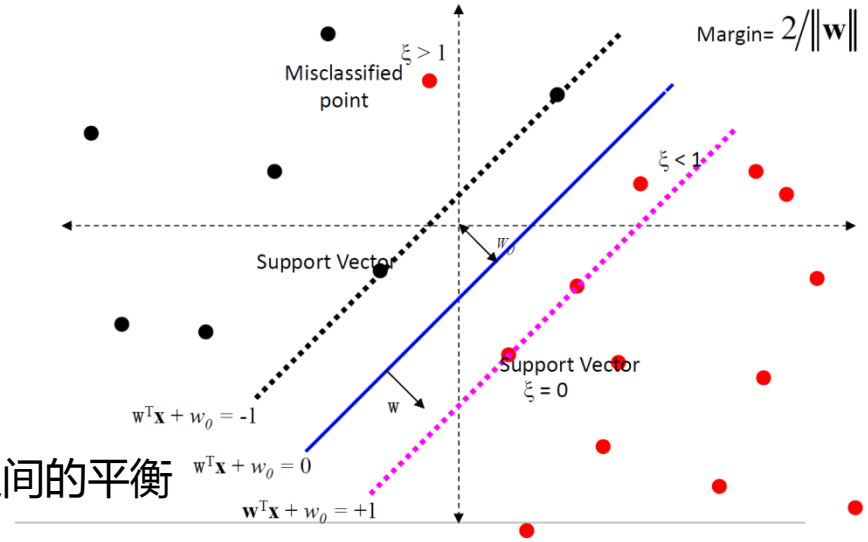
- 目标函数：

$$\min\left(\frac{1}{2}\|w\|^2 + C\sum_i^n \xi_i\right)$$

- 约束条件：

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$



- 其中参数C控制间隔和松弛变量惩罚项之间的平衡

- 被误分的点的$\xi_i > 1$，因此$\sum_i \xi_i$为被误分点的数目的上界，可视为训练误差。

04

PART FOUR

Scikitlearn 中的SVM实现

Scikitlearn中提供的SVC实现

- *class* sklearn.svm.SVC(*C=1.0*, *kernel='rbf'*, *degree=3*, *gamma='auto'*, *coef0=0.0*, *shrinking=True*, *probability=False*, *tol=0.001*, *cache_size=200*, *class_weight=None*,*verbose=False*, *max_iter= -1*, *decision_function_shape='ovr'*, *random_state=None*)
  - C：C-SVC的惩罚参数，默认值是1.0。C越大，即对误分类的惩罚增大，趋向于对训练集全分对
  - probability ：是否采用概率估计，默认为False
  - cache_size ：核函数cache缓存大小，默认为200。核缓存的大小对较大问题求解的运行时间有非常强的影响。如果有足够内存，建议将cache_size设置为更高的值
  - decision_function_shape ：多类分类任务中用到，可为‘ovo’， ‘ovr’ or None,default=None

- 需要调节模型复杂度参数有：C、kernel、degree、gamma、coef0

核函数的参数

- 核函数的参数有degree(M)、gamma(γ)、coef0(θ)，核函数可以有以下几种选择：
  - 线性核：$k(\mathbf{x},\mathbf{x}')=\mathbf{x}^T\mathbf{x}'$
  - 多项式核，缺省为3阶多项式 $k(\mathbf{x},\mathbf{x}')=\left(\gamma\mathbf{x}^T\mathbf{x}'+\theta\right)^M$
  - 径向基函数（RBF）：$k(\mathbf{x},\mathbf{x}')=\exp\left(-\gamma\left(\mathbf{x}-\mathbf{x}'\right)^2\right)$
  - sigmoid函数：$k(\mathbf{x},\mathbf{x}')=\tanh\left(\gamma\mathbf{x}^T\mathbf{x}'+\theta\right)$
  - 也可以用户自定义核
- 在初始化时通过参数kernel指定用什么核

RBF核的参数

- 使用径向基函数（RBF）核训练SVM时，需要考虑正则参数C和核函数宽度参数gamma。
  - 参数C会被所有SVM核用到，用来在样本误分类和决策平面的复杂性之间做出权衡。C越小，决策边界越平滑； C越大，要求更多样本倍分正确。
  - gamma定义单个训练样本能影响多大范围。gamma越大，对应RBF的标准差σ越小，影响的范围更小； gamma越小，决策边界越平滑。
- 这两个值的选择会极大影响SVM的性能。建议使用sklearn.grid_search.GridSearchCV来在C和gamma广阔的指数空间里进行选择，得到最合适的值。

05

PART FIVE

项目实战

案列分析

Otto Group Product Classification Challenge

- 竞赛官网：https://www.kaggle.com/c/otto-group-roduct-classification-challenge
- 电商商品分类：
  – Target：共9个商品类别
  – 93个特征：整数型特征
- 扩展阅读：

  https://blog.csdn.net/v_july_v/article/details/7624837