SVM原理阐述和实现

赵宇盛

西安电子科技大学
计算机科学与技术学院

SVM（Support Vector Machine）支持向量机
——用于监督学习的可定式的二分类器

通常有三种形式:

· 线性可分SVM
· 线性SVM
· 非线性SVM

· hard-margin SVM
· soft-margin SVM
· kernel SVM

SVM三个关键词

Margin

- 定义：γ为整个数据集 D 中所有样本到分割超平面的最短距离 $\gamma = \min_{n} \gamma^{(n)}$
- SVM的学习策略：**margin maximization**
  - ⇔ convex quadratic programming（with constraint）
  - ⇔ minimizing regularizd hinge loss function（without constraint）
- 函数间隔和几何间隔

Dual

- **拉格朗日乘数法** 优化问题$(\boldsymbol{w}, b) \rightarrow (\alpha)$
- 对偶问题消去了特征空间维度对求解超平面难度的影响
- KTT条件中的互补松弛条件(complementary slackness)

Kernel Trick

- 让SVM在非线性可分场景得到适用
- 通过使用核函数表示特征向量之间的内积，等价于隐式地在高维的特征空间中学习线性支持向量机。

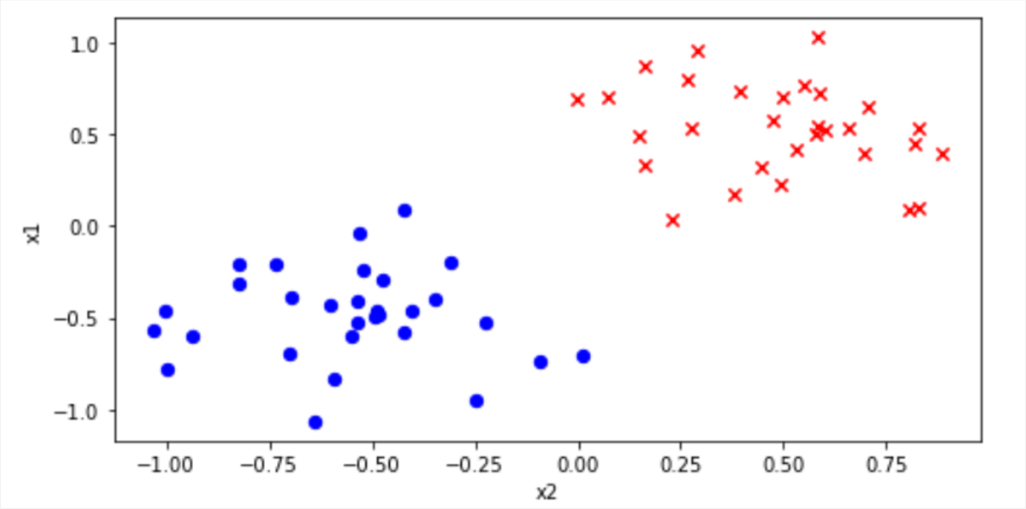
——李航《统计学习方法》

Outline

- Hard-margin SVM
- Dual problem
- Soft-margin SVM
- Kernel SVM

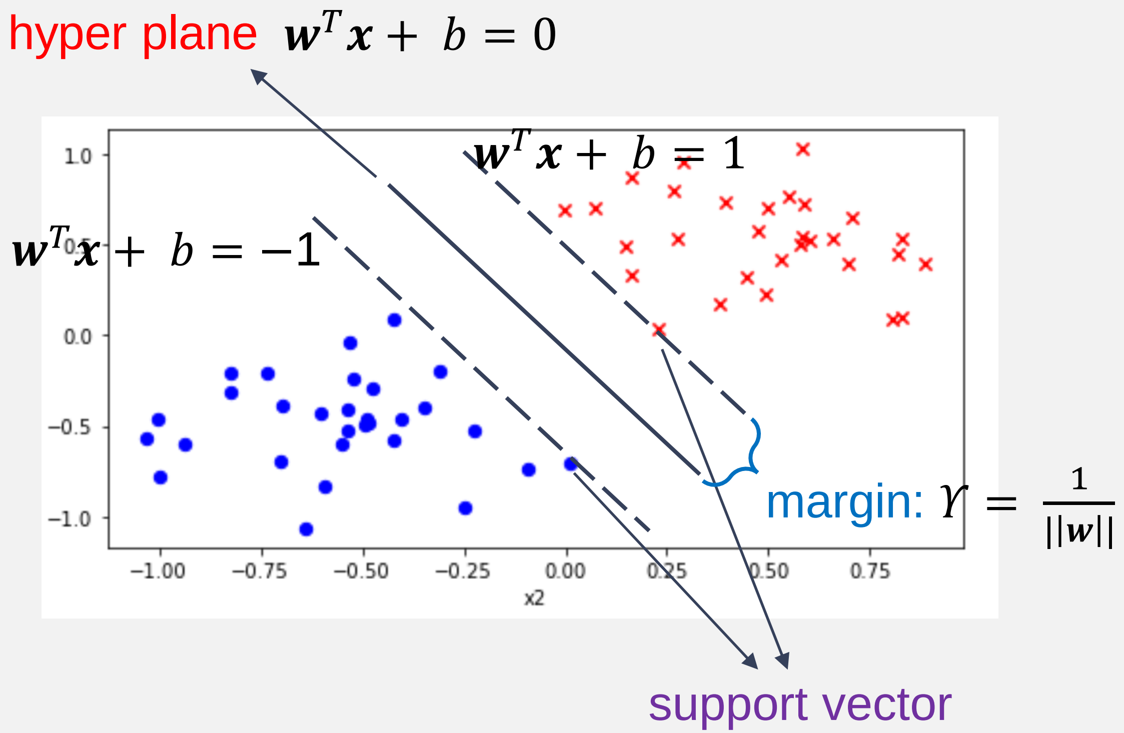
Hard-margin SVM

Example：2-dim 线性可分情况



$\{(\boldsymbol{x}^i, y^i)\}_{i=0}^{n}, \boldsymbol{x}^i \in \mathbb{R}^N, y^i \in \{-1, +1\}$

上图$\mathrm{N} = 2$为例



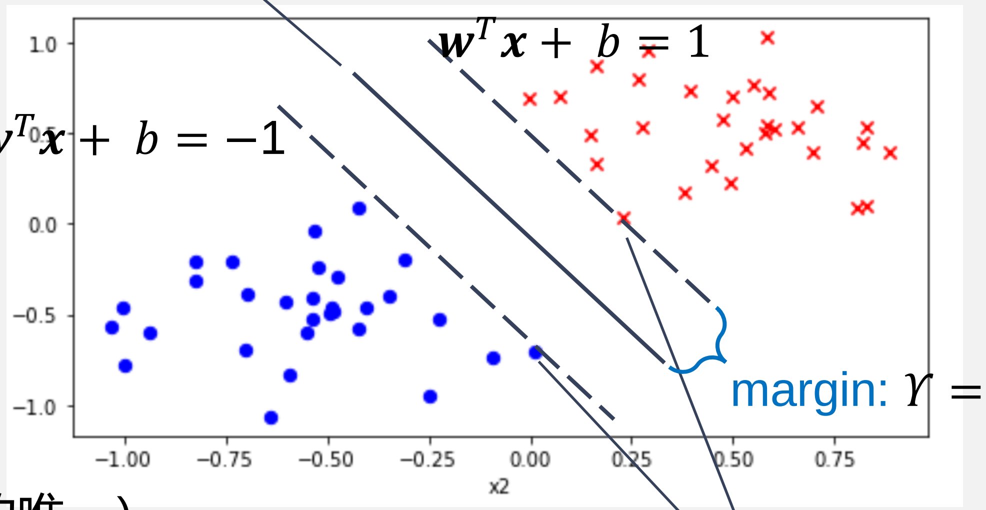
SVM二分类器的**目标**：

找到某一个hyper plane（或者说某一组$\boldsymbol{w}, b$）

使得margin最大化

Hard-margin SVM

hyper plane $\boldsymbol{w}^T\boldsymbol{x} + b = 0$



margin: $\Upsilon = \frac{1}{||\boldsymbol{w}||}$

support vector

$$\max_{\boldsymbol{w},b} \Upsilon = \max_{\boldsymbol{w},b} \min_{n} \Upsilon^{(n)}$$

$s.t.$

$$\boldsymbol{w}^T\boldsymbol{x}^i + b > 0, y^i = +1$$
$$\boldsymbol{w}^T\boldsymbol{x}^i + b < 0, y^i = -1$$

等价于

$$\max_{\boldsymbol{w},b} \Upsilon \equiv \max_{w} \frac{2}{||\boldsymbol{w}||}$$

$s.t.$

$$y^i(w^T\boldsymbol{x}^i + b) > 0$$

为了保持所得到的hyper plane是**唯一**的（准确地说，参数组的唯一）

需要假设 $\min_{w} |\boldsymbol{w}^T\boldsymbol{x}^i + b| = 1$

规范化：

$$\frac{y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b)}{||\boldsymbol{w}||} \geq \frac{1}{||\boldsymbol{w}||} = \Upsilon$$

因此，以上等价于

$$\max_{\boldsymbol{w}} \frac{2}{||\boldsymbol{w}||^2}$$

$s.t.$

$$y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

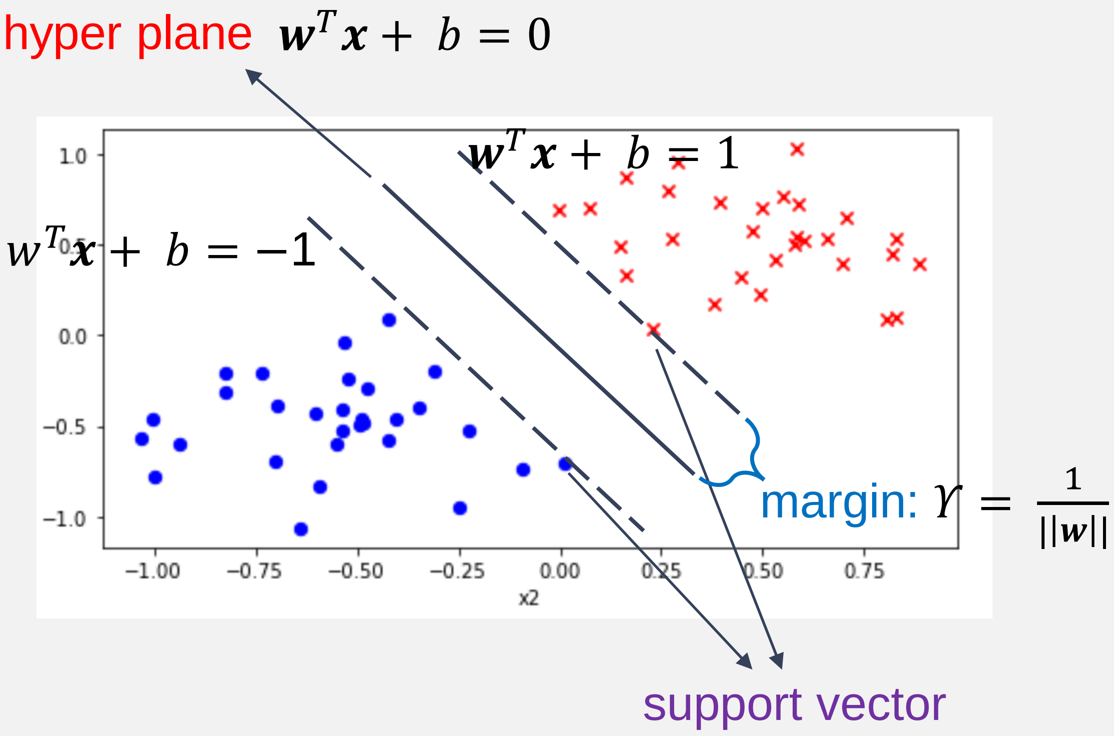
当且仅当$(\boldsymbol{x}^i, y^i)$为support vector时满足等价条件

Hard-margin SVM

$$\max_{\boldsymbol{w}} \frac{2}{||\boldsymbol{w}||^2}$$
$$s.t.$$
$$y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

等价于

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2}$$
$$s.t.$$
$$y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$



SVM： **margin maximization**

convex quadratic programming

如何去做这个凸优化问题呢?

Dual problem

不等式约束的优化问题 $\xrightarrow{\text{拉格朗日乘数法}}$ 等式约束的优化问题

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2}$$

$$s.t.\, y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

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

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

分别求对w,b的偏导

$$\frac{\partial\Lambda}{\partial\boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{n} \alpha_i y^i \boldsymbol{x}^i \qquad \boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y^i \boldsymbol{x}^i$$

$$\frac{\partial\Lambda}{\partial \mathrm{b}} = -\sum_{i=1}^{n} \alpha_i y^i \qquad 0 = \sum_{i=1}^{n} \alpha_i y^i$$

代入Λ中得

$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y^i y^j \left(\left(\boldsymbol{x}^i\right)^T \boldsymbol{x}^j\right) + \sum_{i=1}^{n} \alpha_i \quad s.t. \sum_{i=1}^{n} \alpha_i y^i = 0$$

Dual problem

不等式约束的优化问题 $\xrightarrow{\text{拉格朗日乘数法}}$ 等式约束的优化问题

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2}$$
$$s.t.\, y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j\left(\left(\boldsymbol{x}^i\right)^T\boldsymbol{x}^j\right) + \sum_{i=1}^{n}\alpha_i$$
$$s.t.\sum_{i=1}^{n}\alpha_i\, y^i = 0,\quad \alpha_i \geq 0, i = 1,2,\ldots,n$$

拉格朗日对偶问题

将**原问题**转换为**对偶问题**的变化：

原问题：含有$N+1$个变量数，n 个约束条件。

对偶问题：含有n个变量数，$n+1$个约束条件。

N：$\boldsymbol{w}$的维度，即特征个数
n：样本数量

当特征数远大于样本数($N >> n$)时——拉格朗日对偶形式简化了原问题。

在约束条件下，最大化该对偶函数依然是一个凸二次规划（QP）问题

Dual problem

不等式约束的优化问题 $\xrightarrow{\text{拉格朗日乘数法}}$ 等式约束的优化问题

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2}$$
$$s.t.\, y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j\left(\left(\boldsymbol{x}^i\right)^T\boldsymbol{x}^j\right) + \sum_{i=1}^{n}\alpha_i$$
$$s.t.\sum_{i=1}^{n}\alpha_i\, y^i = 0,\quad \alpha_i \geq 0, i = 1,2,\ldots,n$$

拉格朗日对偶问题

KTT给出了**互补松弛条件**(complementary slackness)

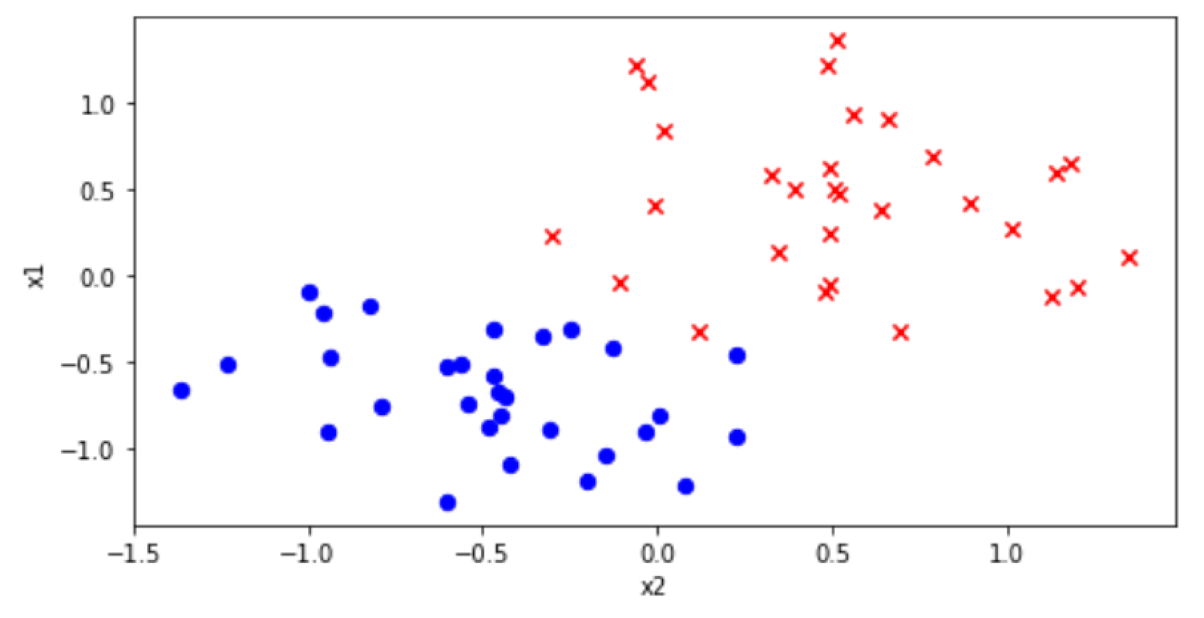
$$\alpha_i\left(1 - y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right)\right) = 0, \alpha_i \geq 0, i = 1,2,\ldots,n$$

互补松弛条件说明当最优解出现在不等式约束的内部，则约束失效。

当$\alpha_i > 0$时，$y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right) = 1$，$\left(\boldsymbol{x}^i, y^i\right)$为决策边界上的support vector

对偶问题的最优解仅仅由support vector决定。

Soft-margin SVM



正例和负例样本在特征空间中不是线性可分的情况

为了能够容忍部分不满足约束的样本，引入松弛变量（Slack Variable）：ξ

Hard-margin SVM

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2}$$

$$s.t.\, y^i(\boldsymbol{w}^T\boldsymbol{x}^i + b) \geq 1$$

优化问题调整为

Soft-margin SVM

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2} + C\sum_{i=1}^{n}\xi_i$$

$$s.t.\, y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right) \geq 1 - \xi_i$$

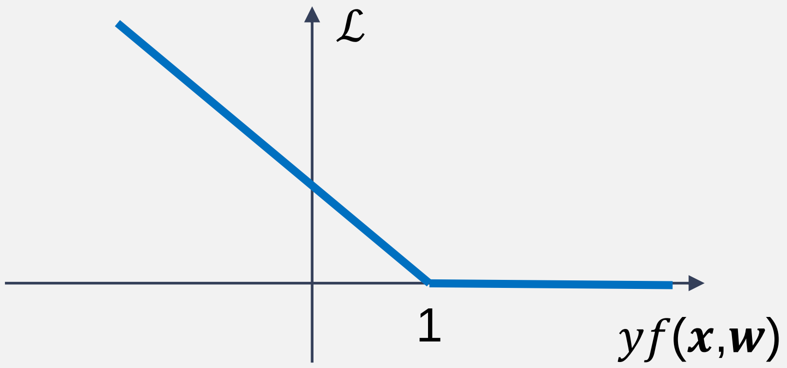
$$\xi_i \geq 0, i = 1,2,\ldots,n$$

参数 $C > 0$ 用来控制间隔和松弛变量惩罚的平衡

Soft-margin SVM

引入hinge loss函数

$$\mathcal{L}_{hinge} = \max\left(0, 1 - y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right)\right), i = 1,2,\ldots,n$$



Soft-margin SVM

$$\min_{w} \frac{||\boldsymbol{w}||^2}{2} + C\sum_{i=1}^{n}\xi_i$$

$$s.t. \; y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right) \geq 1 - \xi_i$$

$$\xi_i \geq 0, i = 1,2,\ldots,n$$

表示为

$$\min_{w}\sum_{i=1}^{n}\max\left(0,1 - y^i\left(\boldsymbol{w}^T\boldsymbol{x}^i + b\right)\right) + \frac{||\boldsymbol{w}||^2}{2C}$$

$$\Leftrightarrow \quad \min_{w}\sum_{i=1}^{n}\mathcal{L}_{hinge} + \boxed{\frac{1}{2C}||\boldsymbol{w}||^2}$$

正则化项

从而，SVM的学习策略可以变为 **minimizing regularizd hinge loss function**

无约束问题，可以用通用的深度学习方法来做！

例子：SVM的一个demo实现

基于 $\min\limits_{w} \sum_{i=1}^{n} \mathcal{L}_{hinge} + \boxed{\frac{1}{2C}||\boldsymbol{w}||^2}$

正则化项

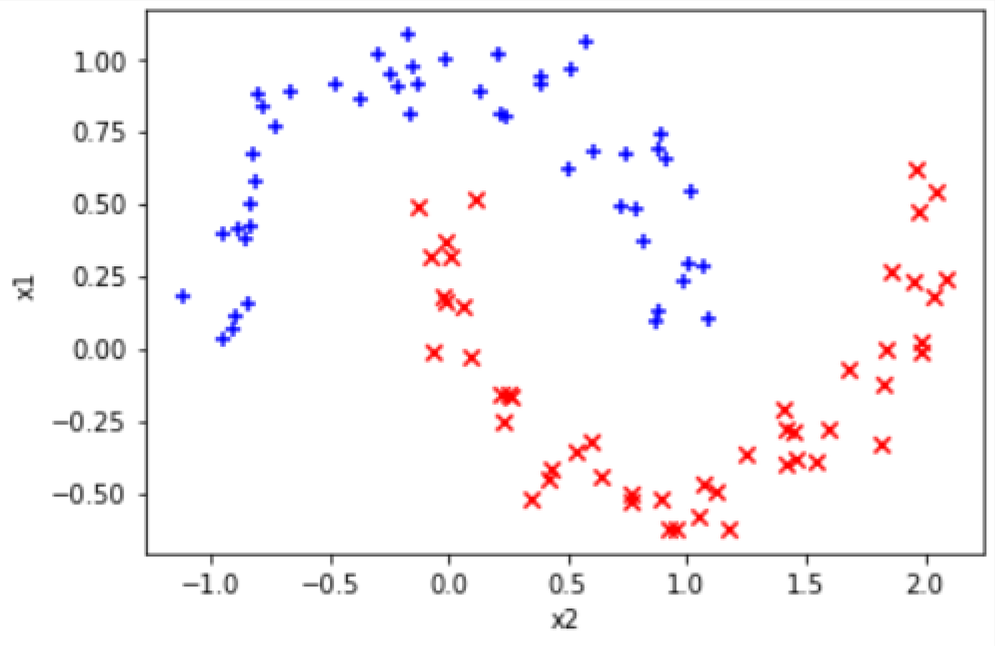
——来自台湾大学李宏毅机器学习课程

我的简单的小demo：（在线性可分的$\mathbb{R}^2$ 数据集上）

- Step 1：未知参数 $\boldsymbol{w}$
- Step 2：损失函数 $\mathcal{L}_{hinge}$ + L2的正则项
- Step 3：优化器：mini-batch SGD

Kernel SVM

非线性可分的情况，譬如☟



$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j \left(\left(\boldsymbol{z}^i\right)^T \boldsymbol{z}^j\right) + \sum_{i=1}^{n}\alpha_i$$

$$s.t. \sum_{i=1}^{n}\alpha_i\, y^i = 0, \quad \alpha_i \geq 0, i = 1,2,\ldots,n$$

$$\Downarrow$$

$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) + \sum_{i=1}^{n}\alpha_i$$

$$s.t. \sum_{i=1}^{n}\alpha_i\, y^i = 0, \quad \alpha_i \geq 0, i = 1,2,\ldots,n$$

使用核函数（Kernel Function）**隐式**地将样本从原始特征空间映射到更高维的空间

$$k\left(\boldsymbol{x}_i, \boldsymbol{x}_j\right) = \left(\Phi(\boldsymbol{x}_i)\right)^T \Phi(\boldsymbol{x}_j)$$

Kernel SVM

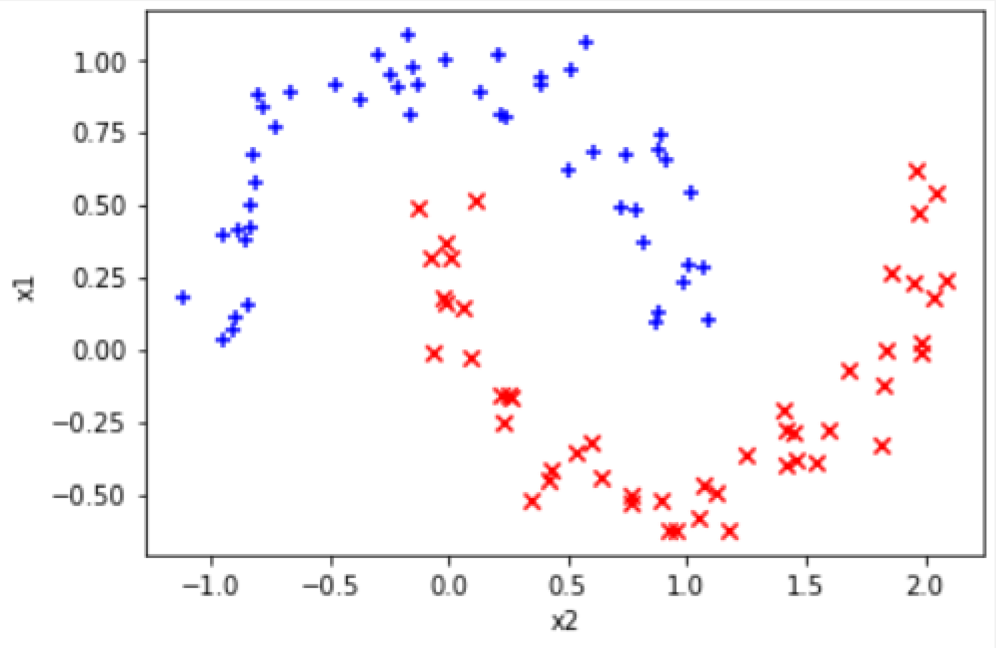
$$\Gamma(\alpha) = -\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y^i y^j k(\boldsymbol{x}_i, \boldsymbol{x}_j) + \sum_{i=1}^{n}\alpha_i$$

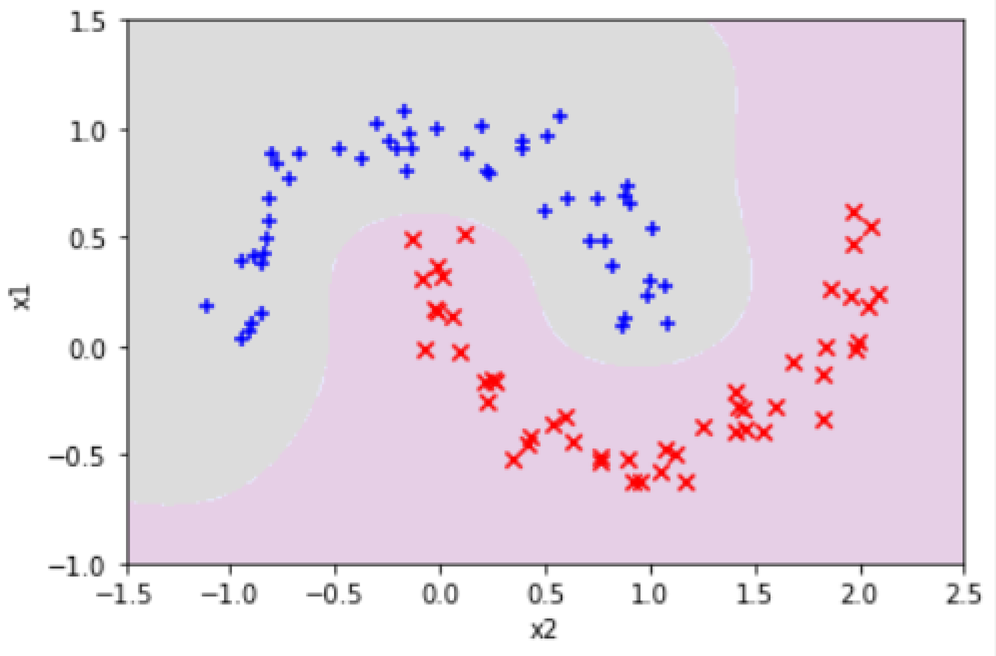
$$s.t. \sum_{i=1}^{n}\alpha_i\, y^i = 0\,,\quad \alpha_i \geq 0, i = 1,2,\ldots,n$$

核函数 $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left(\Phi(\boldsymbol{x}_i)\right)^T \Phi(\boldsymbol{x}_j)$

不同的核函数其VC维也不同，对于$\mathbb{R}^N$大小的特征空间

- 对于线性核的分类器，其超平面是N-1维的，而VC维是N+1维
- 对于高斯核的分类器，其VC维是无穷





我的分享完毕

谢谢大家！