# 2.4 Support Vector Machines

SVM 是一个分类算法，它基于以下几点想法：

1. 对于线性可分的数据集，有大量的边际分类；
2. 对于线性不可分的数据集，可以用松弛变量 slack variables 和正则化；
3. 利用特征转换 feature transformation 和核函数 kernel function 来建造高维分类器。

SVM 被认为是深度学习之前的最佳分类方法。

## 线性支持向量机 Linear SVMs

回想 2.2 当中提到的线性分类器: 给定一个数据集$D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$，其中$y_i \in \{-1,1\}$，学习一个线性分类器 linear classifier:

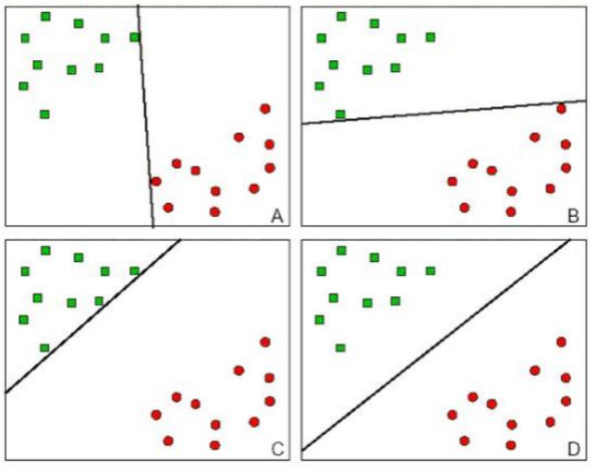
$$\hat{y} = \text{sign}\ (\boldsymbol{w}^T\boldsymbol{x} + b)$$

其中$\boldsymbol{x} = (x_1, x_2, \dots, x_N)^T, \boldsymbol{w} = (w_1, w_2, \dots, w_N)^T$；这里 drop 掉了$x_0 = 1$，同时采用$b$来代替$w_0$；同时，sign函数定义如下：

$$\text{sign}(x) = \begin{cases} 1, & x > 0 \\ 0, & x = 0 \\ -1, & x < 0 \end{cases}$$

其决策边界为$\boldsymbol{w}^T\boldsymbol{x} + b = 0$.

假设数据线性可分，那么可能会有以下多种分法：



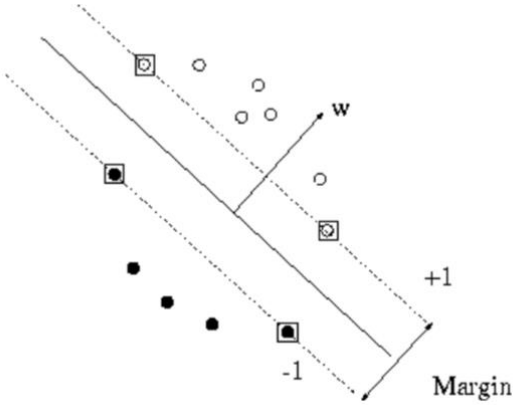

我们常采用置信度最大 maximum confidence 的分法，即将决策边界正好置于两个类别中间，使得

1. 边界到+类别中距离边界最近点的距离=边界到-类别中距离边界最近点的距离；
2. 这个距离是最大的。

这种分类器称为最大边际分类器 maximum margin classifier，又称为线性支持向量机 linear support vector machines，是一种最简单的 SVM。

边际 margin 是指边界到+/-类别中距离边界最近点的距离之和。

支持向量 support vectors 是指决定了边际的点，如下图方框圈住的点。




接下来确定线性支持向量机决策边界当中的$\boldsymbol{w}$和$b$.

边际由三个超平面 hyperplane 决定:

1. 决策边界$\boldsymbol{w}^T\boldsymbol{x}+b=0$;
2. 正平面 Plus-plane：触及+类别的平面，因为其与决策边界平行，所以可以写作$\boldsymbol{w}^T\boldsymbol{x}+b=c$ for some constant $c$ ;
3. 负平面 Minus-plane：触及-类别的平面，根据定义，可以写作$\boldsymbol{w}^T\boldsymbol{x}+b=-c$.

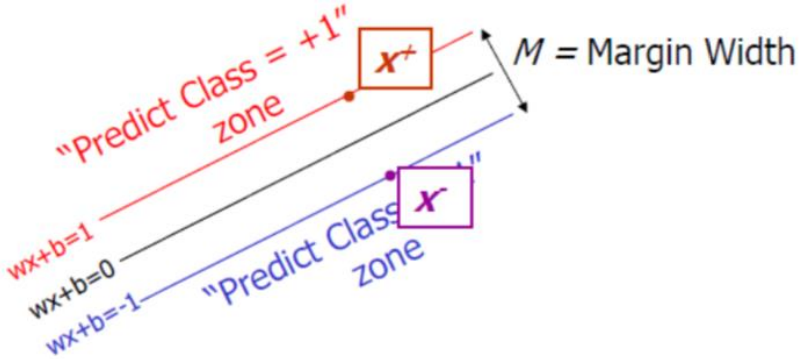
以上三式两边同除以$c$，并重写$\boldsymbol{w}^T/c$和$b/c$为$\boldsymbol{w}^T$和$b$，则有

1. 决策边界$\boldsymbol{w}^T\boldsymbol{x}+b=0$
2. 正平面$\boldsymbol{w}^T\boldsymbol{x}+b=1$
3. 负平面$\boldsymbol{w}^T\boldsymbol{x}+b=-1$

其中$\boldsymbol{w}$垂直于三个平面，对于在决策边界上的两点 u 和 v，有$\boldsymbol{w}^T(\boldsymbol{u}-\boldsymbol{v})=0$

令$\boldsymbol{x}^-$为负平面$\boldsymbol{w}^T\boldsymbol{x}+b=-1$上一点，$\boldsymbol{x}^+$为正平面$\boldsymbol{w}^T\boldsymbol{x}+b=1$上离$\boldsymbol{x}^-$最近一点，则边际$M$是一个从$\boldsymbol{x}^-$指向$\boldsymbol{x}^+$的向量长度，满足：

$$M=||\boldsymbol{x}^+-\boldsymbol{x}^-||$$



从$\boldsymbol{x}^-$指向$\boldsymbol{x}^+$的向量也垂直于三个平面，我们可以假设存在某个常数$\lambda$，使得

$$\boldsymbol{x}^+ - \boldsymbol{x}^- = \lambda \boldsymbol{w}$$

又根据$\boldsymbol{w}^T\boldsymbol{x}^- + b = -1$以及$\boldsymbol{w}^T\boldsymbol{x}^+ + b = 1$，有

$$\lambda = \frac{2}{\boldsymbol{w}^T\boldsymbol{w}}$$

因而

$$M = \left|\left|\boldsymbol{x}^+ - \boldsymbol{x}^-\right|\right| = \left|\left|\lambda \boldsymbol{w}\right|\right| = \lambda\sqrt{\boldsymbol{w}^T\boldsymbol{w}} = \frac{2}{\sqrt{\boldsymbol{w}^T\boldsymbol{w}}} = \frac{2}{\left|\left|\boldsymbol{w}\right|\right|}$$

因而问题变为给定数据集$D = \{\boldsymbol{x}_i, y_i\}_{i=1}^N$，找到合适的$\boldsymbol{w}$和$b$，使得下式最大化：

$$M = \frac{2}{\left|\left|\boldsymbol{w}\right|\right|}$$

约束条件为

$$\boldsymbol{w}^T\boldsymbol{x}_i + b \begin{cases} \geq 1, & \text{if } y_i = 1 \\ \leq -1, & \text{if } y_i = -1 \end{cases}, i \in \{1,2,\dots,N\}$$

以上问题等价于：找到合适的$\boldsymbol{w}$和$b$，使得下式最小化

$$\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2, \qquad \text{subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i \in \{1,2,\dots,N\}$$

这是一个二次规划 quadratic programming/QP 问题，有专业的商业软件可以解决该问题。这里我们仍然会研究对偶优化 dual optimization 问题，它可以让 SVM 高效处理高维数据，这在高维 SVM 当中非常重要。

拉格朗日乘数法是一种寻找多元函数在其变量受到一个或多个条件的约束时的极值的方法，其将一个有 n 个变量与 k 个约束条件的最优化问题转换为一个解有 n + k 个变量的方程组的解的问题。上述问题的拉格朗日函数 Lagrangian 为

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

其中$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_n)$称为拉格朗日乘子，其为约束方程中作为梯度（gradient）的线性组合中各个向量的系数。下面讨论对偶问题。

对于给定的$\boldsymbol{w}$和$b$，定义

$$\mathcal{L}_p(\boldsymbol{w}, b) = \max_{\boldsymbol{\alpha}:\alpha_i \geq 0} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\alpha})$$

若某些$i$违反了约束条件$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1$，则令

$$\mathcal{L}_p(\boldsymbol{w}, b) = \max_{\boldsymbol{\alpha}:\alpha_i \geq 0} \left[\frac{1}{2}\left|\left|\boldsymbol{w}\right|\right|^2 - \sum_{i=1}^{N} \alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)\right] = \infty$$

若所有$i$都满足约束条件$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1$，则令

$$\mathcal{L}_p(\boldsymbol{w},b)=\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\left[\frac{1}{2}||\boldsymbol{w}||^2-\sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1)\right]=\frac{1}{2}||\boldsymbol{w}||^2$$

因而

$$\mathcal{L}_p(\boldsymbol{w},b)=\begin{cases}\frac{1}{2}||\boldsymbol{w}||^2, & \text{if } \boldsymbol{w} \text{ and } b \text{ satisfiey all the constraints}\\ \infty, & \text{otherwise}\end{cases}$$

假若原数据集线性可分，则一定存在一组$\boldsymbol{w}$和$b$满足所有约束条件。
因而原问题可以转化为：$\min_{\boldsymbol{w},b}\mathcal{L}_p(\boldsymbol{w},b)$，使得

1. 满足所有约束条件，否则$\mathcal{L}_p(\boldsymbol{w},b)=\infty$
2. 最小化$\frac{1}{2}||\boldsymbol{w}||^2$

进而原问题转化为

$$\min_{\boldsymbol{w},b}\mathcal{L}_p(\boldsymbol{w},b)=\min_{\boldsymbol{w},b}\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$$

对于给定的$\boldsymbol{\alpha}$，定义

$$\mathcal{L}_d(\boldsymbol{\alpha})=\min_{\boldsymbol{w},b}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$$

对偶优化问题为：

$$\max_{\boldsymbol{\alpha}}\mathcal{L}_d(\boldsymbol{\alpha})=\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\min_{\boldsymbol{w},b}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$$

max 和 min 函数调换了位置。
总体而言，

$$d^*\equiv\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\min_{\boldsymbol{w},b}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})\leq\min_{\boldsymbol{w},b}\max_{\boldsymbol{\alpha}:\alpha_i\geq 0}\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})\equiv p^*$$

当数据集是线性可分的时候，存在$\boldsymbol{w}^*$，$b^*$和$\boldsymbol{\alpha}^*$是拉格朗日函数的解，使得

$$d^*=\mathcal{L}(\boldsymbol{w}^*,b^*,\boldsymbol{\alpha}^*)=p^*$$

同时$\boldsymbol{w}^*$，$b^*$和$\boldsymbol{\alpha}^*$满足以下 KKT 条件 Karush-Kuhn-Tucker conditions:

$$\frac{\partial\mathcal{L}(\boldsymbol{w}^*,b^*,\boldsymbol{\alpha}^*)}{\partial w_j}=0;$$

$$\frac{\partial\mathcal{L}(\boldsymbol{w}^*,b^*,\boldsymbol{\alpha}^*)}{\partial b}=0;$$

$$\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1)=0;\text{(KKT dual complementarity condition)}$$

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1\geq 0;$$

$$\alpha_i\geq 0.$$

求解对偶优化问题：

$$\max_{\boldsymbol{\alpha}} \mathcal{L}_d(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}:\alpha_i \geq 0} \min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$$

对于固定的$\boldsymbol{\alpha}$，首先求$\min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha})$，令梯度为 0：

$$\nabla_{\boldsymbol{w}} \mathcal{L} = \boldsymbol{w} - \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i = 0$$

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

得到

$$\boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i, \qquad \text{and} \sum_{i=1}^{N} \alpha_i y_i = 0$$

结合拉格朗日函数

$$\mathcal{L}(\boldsymbol{w},b,\boldsymbol{\alpha}) = \frac{1}{2}||\boldsymbol{w}||^2 - \sum_{i=1}^{N} \alpha_i (y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) - 1)$$

对偶优化问题变为：找到合适的$\boldsymbol{\alpha}$，使得以下式子最大化，其中$\boldsymbol{x},\boldsymbol{y}$已知

$$\mathcal{L}_d(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

约束条件为

$$\alpha_i \geq 0 \text{ and } \sum_{i=1}^{N} \alpha_i y_i = 0$$

该问题的求解需要用到顺次最小优化 Sequential Minimal Optimization/SMO 算法，最坏时间复杂度为$O(n^2)$.

通过 SMO 算法我们可以求解以上方程得到$\boldsymbol{\alpha}^*$，进而代入以下方程，可以得到$\boldsymbol{w}^*$

$$\boldsymbol{w}^* = \sum_{i=1}^{N} \alpha_i^* y_i \boldsymbol{x}_i$$

原拉格朗日函数的解$\boldsymbol{w}^*$，$b^*$和$\boldsymbol{\alpha}^*$剩下$b^*$未知。可以再次使用 KKT 对偶补充条件：找到任意属于+类别的支持向量$\boldsymbol{x}^+$，满足

$$\boldsymbol{w}^{*\mathrm{T}} \boldsymbol{x}^+ + b^* = 1$$

找到任意属于-类别的支持向量$\boldsymbol{x}^-$，满足

$$\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}^- + b^* = 1$$

综上两式

$$b^* = -\frac{1}{2}(\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}^+ + \boldsymbol{w}^{*\mathrm{T}}\boldsymbol{x}^-)$$

一般来说，取所有支持向量进行上式计算后取均值的效果更好。

从 KKT 对偶补充条件 KKT dual complementarity condition 当中我们看到，

$$\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1) = 0; \text{(KKT dual complementarity condition)}$$
$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 \geq 0;$$
$$\alpha_i \geq 0.$$

当$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 > 0$时，$\alpha_i = 0$; 当$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 = 0$时，$\alpha_i > 0$，说明训练样本处在 margin 边缘时其算子才会被激活，而处在 margin 边缘的训练样本被称为支持向量，进而说明只有支持向量对新进样本的判决起了作用。

判定：对于新进样本$\boldsymbol{x}$, 其所属类别为

$$y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b)$$

因为非支持向量的$\alpha_i^* = 0$, $\boldsymbol{w}^*$实际上可以写成

$$\boldsymbol{w}^* = \sum_{i=1}^{N}\alpha_i^* y_i \boldsymbol{x}_i = \sum_{i=1}^{N_s}\alpha_i^* y_i \boldsymbol{x}_i$$

其中$N_s$为支持向量的个数。进而有

$$y = \text{sign}(\boldsymbol{w}^T\boldsymbol{x} + b) = \text{sign}\left(\sum_{i=1}^{N_s}\alpha_i^* y_i \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{x} + b\right)$$

只有支持向量对新进样本的判决起了作用。

## 数据线性不可分时：松弛变量与正则化

将以上线性 SVM 扩展至更一般的情况：

数据线性可分时，我们寻找一个可以以零误差切割数据集的平面；

数据线性不可分时，我们寻找一个可以以最小误差切割数据集的平面。

数据线性可分时，我们要解决的问题是找到合适的$\boldsymbol{w}$和$b$, 使得下式最小化

$$\frac{1}{2}||\boldsymbol{w}||^2, \qquad \text{subject to } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1, i \in \{1,2,\dots,N\}$$

数据线性不可分时，我们引入松弛变量 slack variables $\xi_i \geq 0$, 约束条件变为：

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in \{1,2,\dots,N\}$$

同时在目标函数当中引入惩罚系数$C$,

$$\text{minimize} \quad \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_i \xi_i$$

$$\text{subject to} \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i, i \in \{1,2,\dots,N\}$$

$\sum_i \xi_i$为训练误差的上确界。$C$越大，证明我们越不希望在最终结果中看到误差，决策边界会更复杂，从而过拟合的可能性越大。

以上问题的拉格朗日函数为

$$\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r}) = \frac{1}{2}||\boldsymbol{w}||^2 + C\sum_{i=1}^{N}\xi_i - \sum_{i=1}^{N}\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1 + \xi_i) - \sum_{i=1}^{N} r_i\xi_i$$

上述式子的最后一项来源于约束条件$\xi_i \geq 0$.

原问题等价于

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \max_{\boldsymbol{\alpha}:\alpha_i\geq 0;\boldsymbol{r}:r_i\geq 0} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})$$

对偶问题为

$$\max_{\boldsymbol{\alpha}:\alpha_i\geq 0;\boldsymbol{r}:r_i\geq 0} \min_{\boldsymbol{w},b,\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})$$

先求解以下问题

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{r})$$

有

$$\nabla_{\boldsymbol{w}}\mathcal{L} = 0 \Rightarrow \boldsymbol{w} = \sum_{i=1}^{N}\alpha_i y_i \boldsymbol{x}_i$$

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

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - r_i = 0$$

从而对偶问题变为：找到合适的$\boldsymbol{\alpha}$，其中$\boldsymbol{x}, \boldsymbol{y}$已知，使得

$$\text{maximize} \quad \sum_{i=1}^{N}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{N}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

$$\text{subject to} \quad C \geq \alpha_i \geq 0 \text{ and } \Sigma_{i=1}^{N}\alpha_i y_i = 0$$

## 非线性 SVM：核函数

核心思想：将低维线性不可分的数据转化为高维线性可分的数据

假设某个线性不可分的数据集$D = \{\boldsymbol{x}_i, y_i\}_{i=1}^{N} \subset \mathbb{R}^{\mathrm{m}}$，采用两个映射

$$\mathbb{R}^{\mathrm{m}} \to \mathcal{H}, \qquad \boldsymbol{x} \to \boldsymbol{\phi}(\boldsymbol{x})$$

其中$\mathbb{R}^{\mathrm{m}}$为原始数据空间，是输入的属性 attributes;

$\mathcal{H}$为特征空间，是属性组合而成的特征 features。

同时将原始数据空间的$\boldsymbol{x}$映射至特征空间$\boldsymbol{\phi}(\boldsymbol{x})$。

非线性 SVM 的对偶优化问题为

$$\text{maximize} \quad \mathcal{L}_d(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j)$$

$$\text{subject to} \quad C \geq \alpha_i \geq 0 \ \text{ and } \ \Sigma_{\mathrm{i=1}}^{\mathrm{N}} \alpha_i y_i = 0$$

定义核函数$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$, iff 满足 Mercer Condition:

1. $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$是一个对称矩阵
2. $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$是一个半正定矩阵 Positive Semi-definite Matrix:

$$\boldsymbol{x}^{\mathrm{T}} \boldsymbol{K} \boldsymbol{x} \geq 0, \qquad \text{subject to } \boldsymbol{x} \subset \mathbb{R}^n \text{ and } \boldsymbol{x} \neq \boldsymbol{0}$$

则有

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \boldsymbol{\phi}(\boldsymbol{x}_i)^T \boldsymbol{\phi}(\boldsymbol{x}_j)$$

从而对偶优化问题变为

$$\text{maximize} \quad \mathcal{L}_d(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$\text{subject to} \quad C \geq \alpha_i \geq 0 \ \text{ and } \ \Sigma_{\mathrm{i=1}}^{\mathrm{N}} \alpha_i y_i = 0$$

进而解出$\boldsymbol{w}^*$和$b^*$，再根据以下式子判决新进样本所属类型：

$$y = \text{sign}\left( \sum_{i=1}^{N_s} \alpha_i^* y_i K(\boldsymbol{x}_i, \boldsymbol{x}_j) + b^* \right)$$

## 核技巧和核函数

不需要定义$\boldsymbol{\phi}(\boldsymbol{x})$再计算内积，可以直接定义核函数$K(\boldsymbol{x}_i, \boldsymbol{x}_j)$，可以较易计算高维特征。

常见的核函数：

1. 多项式核 polynomial kernel:

$$K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}^{\mathrm{T}} \boldsymbol{z})^d$$

当$d=2$时，

$$\phi(\boldsymbol{x})=(x_1x_1,x_1x_2,x_1x_3,\dots,x_nx_n)$$

衍生形式：不均匀多项式核 Inhomogeneous polynomial

$$K(\boldsymbol{x},\boldsymbol{z})=(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}+1)^d$$

2. 高斯核 Gaussian:

$$K(\boldsymbol{x},\boldsymbol{z})=\exp(-\frac{\left||\boldsymbol{x}-\boldsymbol{z}|\right|^2}{2\sigma^2})$$

以上核函数又称为径向基函数核 Radial Basis Function/RBF

根据泰勒展开式

$$e^n=\sum_{i=0}^{\infty}\frac{n^i}{i!}=1+n+\frac{n^2}{2!}+\frac{n^3}{3!}+\cdots$$

有

$$\begin{aligned}K(\boldsymbol{x},\boldsymbol{z})&=\exp\left(-\frac{\left||\boldsymbol{x}-\boldsymbol{z}|\right|^2}{2\sigma^2}\right)\\&=\exp\left(-\frac{|\boldsymbol{x}|^2+|\boldsymbol{z}|^2-2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}}{2\sigma^2}\right)\\&=\exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right)\exp\left(-\frac{|\boldsymbol{z}|^2}{2\sigma^2}\right)\exp\left(\frac{2\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z}}{2\sigma^2}\right)\\&=\exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right)\exp\left(-\frac{|\boldsymbol{z}|^2}{2\sigma^2}\right)\left(\sum_{i=0}^{\infty}\frac{(\boldsymbol{x}^{\mathrm{T}}\boldsymbol{z})^i}{i!\,\sigma^2}\right)\quad\text{Taylor Expansion}\\&=\sum_{i=0}^{\infty}\left[\exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right)\exp\left(-\frac{|\boldsymbol{z}|^2}{2\sigma^2}\right)\sqrt{\frac{1}{i!\,\sigma}}\sqrt{\frac{1}{i!\,\sigma}}(\boldsymbol{x}^T)^i(\boldsymbol{z})^i\right]\\&=:\phi(\boldsymbol{x})^{\mathrm{T}}\phi(\boldsymbol{z})\end{aligned}$$

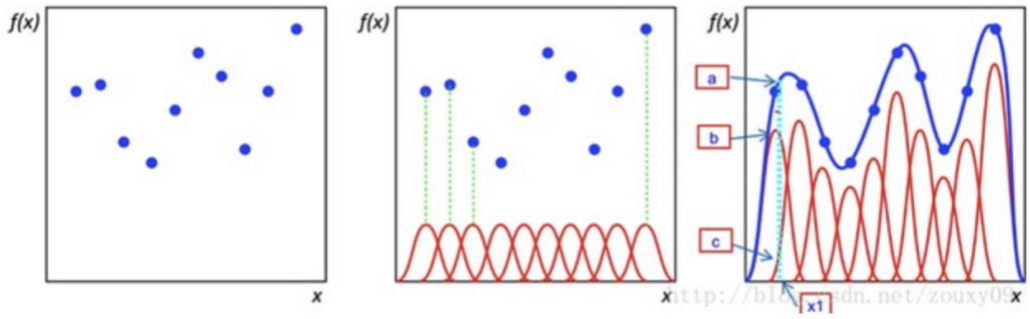
其中

$$\phi(\boldsymbol{x})^{\mathrm{T}}=\exp\left(-\frac{|\boldsymbol{x}|^2}{2\sigma^2}\right)\cdot\left(1,\sqrt{\frac{1}{\sigma}}\boldsymbol{x}^T,\sqrt{\frac{1}{2!\,\sigma}}(\boldsymbol{x}^T)^2,\dots\right)$$

$$\phi(\boldsymbol{z})=\exp\left(-\frac{|\boldsymbol{z}|^2}{2\sigma^2}\right)\cdot\left(1,\sqrt{\frac{1}{\sigma}}\boldsymbol{z},\sqrt{\frac{1}{2!\,\sigma}}(\boldsymbol{z})^2,\dots\right)$$

因此高斯核函数拥有将任意多的属性映射至无限多维的能力。

相关应用有径向基函数网络 RBF Network.




以上图示较好的说明了径向基函数的作用。在决定线性不可分数据集的决策边界时，对于样本当中的每个点生成一个高斯函数（如图 2 中的红色函数），用径向基与其对应（如图 2 中的绿色虚线）。之后将高斯函数与该径向基相乘，得到一个拉伸的高斯函数，每个高斯函数才自己的中心点附近为决策边界做出贡献，从而得到决策边界（图 3 当中的蓝色曲线）。

高斯核函数当中$\sigma$的作用：较大的$\sigma$代表新进样本所属类别由较多的训练样本点决定；较小的$\sigma$代表新进样本所属类别由最靠近的训练样本点决定。

3. 其他核函数

如何根据已知的核函数构建新的核函数：

Given valid kernels $k_1(\mathbf{x}, \mathbf{x}')$ and $k_2(\mathbf{x}, \mathbf{x}')$, the following new kernels will also be valid:

$$
\begin{aligned}
k(\mathbf{x}, \mathbf{x}') &= c k_1(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= f(\mathbf{x}) k_1(\mathbf{x}, \mathbf{x}') f(\mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= q\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \exp\left(k_1(\mathbf{x}, \mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') + k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_1(\mathbf{x}, \mathbf{x}') k_2(\mathbf{x}, \mathbf{x}') \\
k(\mathbf{x}, \mathbf{x}') &= k_3\left(\phi(\mathbf{x}), \phi(\mathbf{x}')\right) \\
k(\mathbf{x}, \mathbf{x}') &= \mathbf{x}^{\mathrm{T}} \mathbf{A} \mathbf{x}' \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') + k_b(\mathbf{x}_b, \mathbf{x}_b') \\
k(\mathbf{x}, \mathbf{x}') &= k_a(\mathbf{x}_a, \mathbf{x}_a') k_b(\mathbf{x}_b, \mathbf{x}_b')
\end{aligned}
$$

where $c > 0$ is a constant, $f(\cdot)$ is any function, $q(\cdot)$ is a polynomial with nonnegative coefficients, $\phi(\mathbf{x})$ is a function from $\mathbf{x}$ to $\mathbb{R}^M$, $k_3(\cdot, \cdot)$ is a valid kernel in $\mathbb{R}^M$, $\mathbf{A}$ is a symmetric positive semidefinite matrix, $\mathbf{x}_a$ and $\mathbf{x}_b$ are variables (not necessarily disjoint) with $\mathbf{x} = (\mathbf{x}_a, \mathbf{x}_b)$, and $k_a$ and $k_b$ are valid kernel functions over their respective spaces.