机器学习：支撑向量机(Support Vector Machine)

---

1. 问题的描述

给定样本 $(x^{(i)}, y^i)$ ，其中 $y^{(i)} \in \{-1, +1\}$ ，假设分类器可参数化：

$$h(w,b) = g(w^T x + b) \quad \text{即}([w^T, b]\begin{bmatrix} X \\ 1 \end{bmatrix})$$

其中 $g(z)$ 为符号函数，即 $g(z)=1$ 当 $z \geqslant 0$； $g(z)=-1$，当 $z<0$。

而支撑向量机的目标就是，希望找到一条线，使得对于分类面的间隔最大。

2. 间隔的定义

定义：函数间隔：

$$\tilde{\gamma}^{(i)}=y^{(i)}(w^T X^{(i)}+b)$$

定义：几何间隔，点 $x^{(i)}$ 到超平面 $w^T x+b=0$ 的距离：

$$\gamma^{(i)}=y^{(i)}\left(\left(\frac{w}{||w||_2}\right)^T x^{(i)}+\frac{b}{||w||}\right)$$

式子中 $y^{(i)}$ 决定符号。

超平面的法向量(单位)： $\frac{w}{||w||_2}$

推导过程：

$$\begin{cases} x_B=x^{(i)}-\frac{w}{||w||_2}\gamma^{(i)} \\ w^T X^B+b=0 \end{cases} \quad w^T x^{(i)}-\frac{w^T w}{||w||_2}\gamma^{(i)}+b=0$$

定义：硬间隔和软间隔。硬间隔问题指数据集可以通过直线完全分开，软间隔问题指数据不能完全由直线分开。

3. SVM 的核心问题

最大化"间隔"(点到超平面的距离)，即希望找到一条线，使得对于分类面的间隔最大。

$$\gamma=\min_{i=1,\cdots,n}\gamma^{(i)}$$

$$\max \gamma$$

$$\text{即} \max_{w,b}\min \gamma^{(i)}$$

即：

$$\max_{w,b}\min_{i=1,2\cdots n}\gamma^{(i)} \Leftrightarrow \begin{cases} \max \gamma \\ s.t \quad y^{(i)}\left(\left(\frac{w}{||w||_2}\right)^T x^{(i)}+\frac{b}{||w||}\right) \geqslant \gamma \end{cases}$$

不失一般性，令 $||w||_2=1$：

$$\Leftrightarrow \begin{cases} \max\limits_{w,b} \gamma \\ s.t \quad ||w||_2=1, \quad y^{(i)}(w^T x^{(i)}+b) \geqslant \gamma \end{cases}$$

$$\Leftrightarrow \begin{cases} \max\limits_{w,b} \frac{\hat{\gamma}}{||w||_2}, \quad \text{令} \hat{\gamma}=\gamma \cdot ||w||_2 \\ s.t \quad ||w||_2=1, \quad y^{(i)}(w^T x^{(i)}+b) \geqslant \hat{\gamma} \end{cases}$$

$\because \hat{w}=\frac{w}{\hat{\gamma}}, \hat{b}=\frac{b}{\hat{\gamma}}$ 。$\therefore$ 不失一般性，令 $\hat{\gamma}=1$ 。最终我们将 $||w||_2$ 这个难解的非线性约束放在了上面的 $\max$ 部分。

$$\Leftrightarrow \begin{cases} \max\limits_{w,b} \frac{1}{||w||_2} \\ s.t \quad y^{(i)}(w^T x^{(i)}+b) \geqslant 1 \end{cases}$$

$$\Leftrightarrow \begin{cases} \min\limits_{w,b} \quad ||w||_2^2 = w^T w \\ s.t \quad y^{(i)}(w^T x^{(i)}+b) \geqslant 1 \end{cases}$$

可以用二次规划(QP)，拉格朗日乘子等方法求解该问题，QP 凸优化方法求解如下问题：

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 \\ s.t \quad y^{(i)}(w^T x^{(i)}+b) \geqslant 1 \end{cases}$$

4. 线性可分 SVM 的优化问题

4.1. 优化公式推导(硬间隔问题)

线性 SVM 本来的优化目标为：

$$\begin{cases} \max \ margin(w,b) \\ s.t \quad every \ y^{(i)}(w^T x^{(i)}+b) > 0 \\ \qquad margin(w,b) = \min\limits_{i=1,\cdots,n} \frac{1}{||w||} y^{(i)}(w^T x^{(i)}+b) \end{cases}$$

即点到超平面的距离公式，为 $margin = \frac{y^{(i)}(w^T x^{(i)}+b)}{||w||_2^2}$ 。优化问题是求 $\max(margin)$ 。而该超平面方程存在冗余，分类面在乘以 k 倍后，实际最优化的分类面是不变的。

$$3w^T x + 3b = 0 \Leftrightarrow w^T x + b = 0$$

所以，可以加上一个约束条件消除改冗余，即约束分子为 1 ，求分母的优化问题。

则线性 SVM 实际的优化问题为：

$$\begin{cases} \max\limits_{w,b} \quad \frac{1}{||w||} \\ s.t \quad \min\limits_{i=1,\cdots,n} y^{(i)}(w^T x^{(i)}+b) = 1 \end{cases}$$

添加松弛条件(扩大值域 $\geqslant 1$)；分类超平面与 2 种样本的间隔 $d(w,b) = \frac{2}{||w||}$ ；转为求极小值的凸优化问题，最终该优化问题转化为：

$$\begin{cases} \min\limits_{w,b} \quad \frac{1}{2} w^T w \\ s.t \quad y^{(i)}(w^T x^{(i)}+b) \geqslant 1, \quad for \ all \ i \end{cases}$$

最终求解 $\frac{1}{2}||w||_2^2$ 的优化问题，可以使用梯度下降，$QP \ solver$ ，拉格朗日乘子法等求解。

4.2. 合页损失($Hinge\ Loss$)与随机梯度下降

令 $S_n = w^T x + b$，约束问题 $y_n(w^T x_n + b) \geqslant 1$，转为 $1 - ys \leqslant 0$。

那么可以定义，对于 SVM，分类正确时，损失函数值为 0，错误时，损失函数值为 $1 - ys$。

则损失函数定义为：

$$L_{SVM} = \max(0, 1 - ys)$$

即 $y = 1, err_{SVM} = \max(1, 1 - s), s > 1$，$y = -1, err_{SVM} = \max(1, 1 + s), s < -1$。

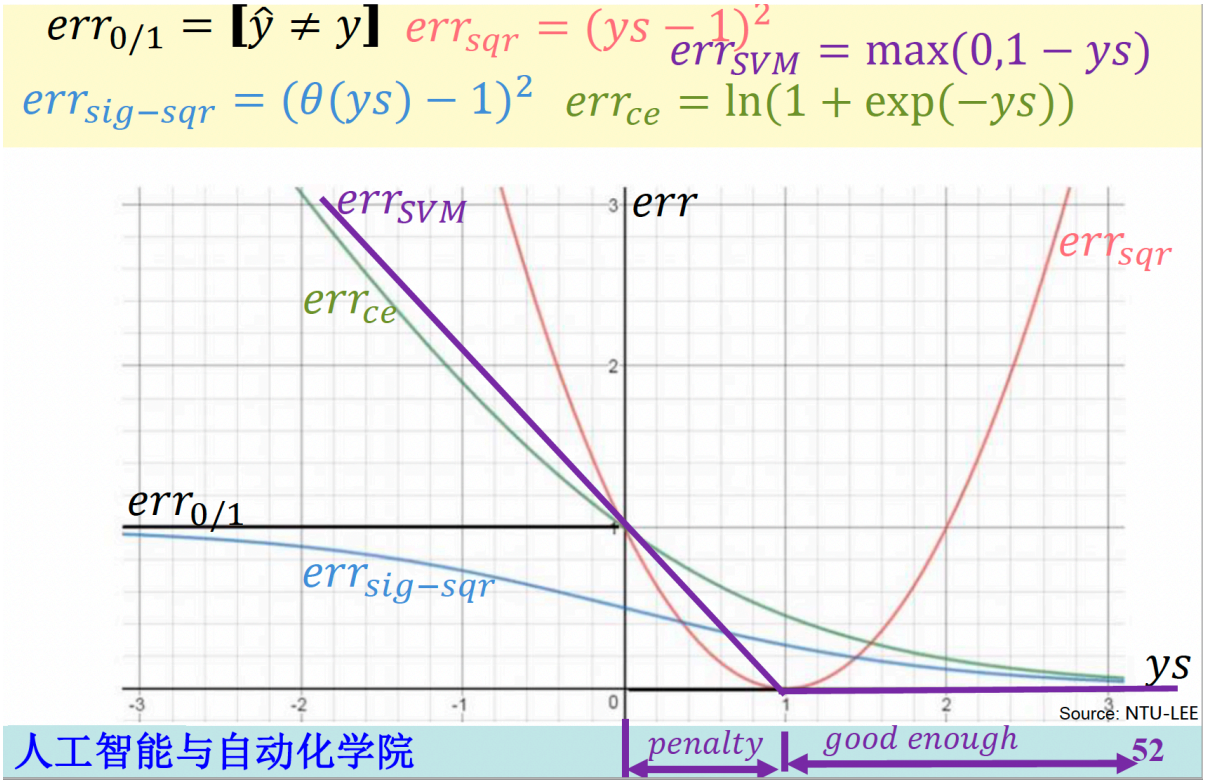
这种函数迫使模型的预测值有大的间隔，距离分类线尽可能远，此即为合页损失 $Hinge\ Loss$，然后求解梯度，并用梯度下降法求解优化问题。

$$\frac{\partial L_{SVM}}{\partial w} = [[1 - y_n(w^T x_n) > 0]](-y_n x_n)$$

线性 SVM 的 $mini - batch$ 梯度下降算法过程如下：

$$\begin{aligned}
&while(until\ \nabla L(w) = 0 \vee 迭代次数到\ n):\\
&\quad for\ batch\ in\ dataloader(shuffle):\\
&\qquad \nabla L(w) = \frac{1}{batch_size} \sum_{i=1}^{batch_size} [[1 - y_i(w_t^T x_i) > 0]](-y_i x_i)\\
&\qquad w_{t+1} = w_t - \eta \cdot \nabla L(w)
\end{aligned}$$

对于感知器 $0/1$ 损失，线性回归(均方误差)，S 函数 $Logistic$ 误差 / 均方误差，$SVM - Hinge\ Loss$ 的对比如上图所示。



4.3. QP 求解

可以用二次规划求解该**线性硬间隔**问题(QP)，QP 凸优化方法求解：

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 \\ s.t \quad y^{(i)}(w^T x^{(i)} + b) \geqslant 1, \quad for\ i = 1, \cdots N \end{cases}$$

转为二次规划问题：

$$\begin{cases} u \leftarrow QP(Q, p, A, c) \\ \min\limits_{u} \frac{1}{2} u^T Q u + p^T u \\ s.t \quad a_m^T u \geqslant c_m, \quad for\ m = 1, \cdots, M \end{cases}$$

对于硬间隔(线性 SVM)问题，Q, p 目标方程和 a, c 不等式约束可以写成矩阵形式：

$$Q = \begin{bmatrix} 0 & 0_d^T \\ 0_d & I_d \end{bmatrix} \quad p = 0_{d+1} \quad a_n^T = y_n \begin{bmatrix} 1 & x_n^T \end{bmatrix} \quad c_n = 1$$

原问题即为，求解该最小值优化问题：

$$\begin{bmatrix} b \\ w \end{bmatrix} = QP(Q, p, A, c)$$

最终返回 w, b ，得到 SVM 的分类方程：$h(w, b) = g(w^T x + b)$ 。

5. 对偶 SVM (线性 / 非线性)的优化问题

线性和非线性只是自变量(x_n, z_n)不同，分类面一个为直线一个为弧线。优化过程和求解方法，其实没有本质区别，有关对偶理论的 SVM ，这里选择都用 z_n 高维数据来表示自变量，即讨论非线性的情况，线性情况将自变量换为 x_n ，其他推导类似。

对于非线性 SVM ，通过核函数升维，即构造核 SVM ，使得映射 $z_n = \Phi(x_n)$ ，它多种求解方式。

5.1. QP 求解的局限性

和硬间隔 QP 一样，对于非线性问题，依旧可以用二次规划求解该问题(QP)：

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 \\ s.t \quad y^{(i)}(w^T z^{(i)} + b) \geqslant 1, \quad for\ i = 1, \cdots N \end{cases}$$

其中，$z_n = \Phi(x_n)$ ，转为二次规划问题：

$$\begin{cases} u \leftarrow QP(Q, p, A, c) \\ \min\limits_{u} \frac{1}{2} u^T Q u + p^T u \\ s.t \quad a_m^T u \geqslant c_m, \quad for\ m = 1, \cdots, M \end{cases}$$

对于该非线性问题，Q, p 目标方程和 a, c 不等式约束可以写成矩阵形式：

$$Q=\begin{bmatrix}0 & 0_{\tilde{d}}^T\\ 0_d & I_{\tilde{d}}\end{bmatrix}\quad p=0_{\tilde{d}+1}\quad a_n^T=y_n\begin{bmatrix}1 & z_n^T\end{bmatrix}\quad c_n=1$$

问题即为，求解该最小值优化问题：

$$\begin{bmatrix}b\\ w\end{bmatrix}=QP(Q,p,A,c)$$

最终返回 $w\in R^{\tilde{d}},b\in R$，得到 SVM 的分类方程：$h(w,b)=g(w^T\Phi(x)+b)$。

该求解方法有一定问题，QP 问题有 $\tilde{d}+1$ 个变量，如果矩阵 $\tilde{d}$ 过大或者 ∞，会带来该问题的严重求解困难。

所以我们的目标是找到一个没有 $\tilde{d}$ 矩阵依赖的非线性 SVM 求解方法，即拉格朗日对偶求法。

5.2. 拉格朗日乘数法

首先复习一下Lesson 5 监督学习之分类(Logistic, Bayes, MAP)提到的拉格朗日乘数法，以一阶线性的优化函数为例，对于优化问题：

$$\begin{cases}\min\limits_x f(x)\\ s.t\quad h_i(x)\quad =0\quad i=1,\cdots,p\end{cases}$$

构造拉格朗日乘子函数：

$$\mathcal{L}(x,\lambda)=f(x)+\sum_{i=1}^{p}\lambda_i h_i(x)$$

对优化变量 $x,y,\cdots$，乘子变量 λ 分别求偏导并令其为 0，可以得到候选极值点。再求黑塞矩阵，判别它是否正定，判断该候选极值点是否是极值点。

5.3. 原问题和对偶问题

对于支撑向量机的原问题：

$$\begin{cases}\min\limits_w f(w)\\ \quad\quad g_i(w)\leqslant 0\quad i=1,\cdots,k\\ s.t\quad\quad h_i(w)=0\quad i=1,\cdots,l\end{cases}$$

定义拉格朗日函数：

$$\mathcal{L}(w,\alpha,\beta)=f(w)+\sum_{i=1}^{k}\alpha_i g_i(w)+\sum_{i=1}^{l}\beta_i h_i(w)$$

其中 α_i,β_i 为拉格朗日乘子

即，$\mathcal{L}=$ 目标函数 $+$ 不等式约束 $+$ 等式约束。

定义**原问题**：

$$Q_P(w)=\max_{\alpha\geqslant 0,\beta}\mathcal{L}(w,\alpha,\beta)$$

结论：

$$Q_P(w)=\begin{cases}f(w), & \text{如果 } w_i \text{ 满足约束}\\ \infty, & otherwise\end{cases}$$

因此初始问题等价于(对 α,β 求极大值后，消去这两个变量，再对 w 求极小值)：

$$\min_w Q_P(w)=\min_w \max_{\alpha\geqslant 0,\beta}\mathcal{L}(w,\alpha,\beta)$$

定义**对偶问题**：

$$Q_D(\alpha,\beta)=\min_w \mathcal{L}(w,\alpha,\beta)$$

即(对 w 求极小值后，对 α,β 求函数极大值)：

$$\max_{\alpha\geqslant 0,\beta}\min_w \mathcal{L}(w,\alpha,\beta)=\max_{\alpha\geqslant 0,\beta} Q_D(\alpha,\beta)$$

定义对偶问题和原问题的**最优取值**：

$$d^*\triangleq \max_{\alpha\geqslant 0,\beta} Q_D(\alpha,\beta)$$

$$p^*\triangleq \min_w Q_P(w)$$

$$\max_w\min_{\alpha\geqslant 0,\beta}\mathcal{L}(w,\alpha,\beta)=d^*$$

$$\min_{\alpha\geqslant 0,\beta}\max_w\mathcal{L}(w,\alpha,\beta)=p^*$$

5.4. 定理：弱对偶定理

如果原问题和对偶问题都存在最优解，则对偶问题的最优值不大于原问题的最优值。

根据数学推导可知，对于二元函数：

$$\max_x\min_y f(x,y)=\max_x f(x,y^*)$$

$$\min_y\max_x f(x,y)=\min_y f(x^*,y)$$

结论： $d^*\leqslant p^*$ 。

定义 $[$对偶间隔$]$： $p^*-d^*\geqslant 0$ ，而我们关心的是 $p^*=d^*$ 时的情况，也就是强对偶成立的时候。

5.5. 定理：$[Slater]$ 条件

$[Slater]$ 条件用于将原问题转为对偶问题，它是强对偶的充分条件($Slater$ -> 强对偶)

假设 f 和 g_i 是凸的，h_i 是仿射。同时假设 g_i 是严格可行的，即存在一个候选 w ，使得对于 $\forall i$ ，满足 $g_i(w)<0$ 。

则存在 (x^*,α^*,β^*) ，使得它们同时为原问题和对偶问题的最优解：

$$p^* = d^* = \mathcal{L}(x^*, \alpha^*, \beta^*)$$

$[Slater]$ 条件将 SVM 的问题转为一个凸优化的问题，方便求解。

5.6. 定理：KKT 条件

KKT 条件用于求解带等式和不等式约束的最优化问题，给出了一阶必要条件。

假设如下优化问题：

$$\begin{cases} \min\limits_{w} f(w) \\ s.t \quad g_i(w) \leqslant 0 \quad i = 1, \cdots, k \\ \qquad h_i(w) = 0 \quad i = 1, \cdots, l \end{cases}$$

之后构造拉格朗日乘子函数消去等式和不等式约束：

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^{k} \alpha_i g_i(w) + \sum_{i=1}^{l} \beta_i h_i(w)$$

在上述条件和假设下， w^*, α^*, β^* 可由 KKT 条件解出：

$$\begin{aligned} &\frac{\partial \mathcal{L}(w^*, \alpha^*, \beta^*)}{\partial w_i} = 0, \quad i = 1, \cdots, n \\ &\frac{\partial \mathcal{L}(w^*, \alpha^*, \beta^*)}{\partial \beta_i} = 0, \quad i = 1, \cdots, l \\ &\alpha_i^* \geqslant 0, \quad g_i(w^*) \leqslant 0, \quad h_i(w^*) = 0 \\ &\alpha_i^* \cdot g_i(w^*) = 0, \quad i = 1, \cdots, k \end{aligned}$$

KKT 条件只是取得极值的必要条件(极值 -> KKT)，如果一个最优化问题是凸优化问题，则 KKT 条件为优化问题满足解的充分必要条件，因此可以通过求解 KKT 条件求解 SVM 。

5.7. 线性 / 非线性可分 SVM 的拉格朗日对偶优化求解

回到 SVM 问题本身，前面我们已经解释过，线性和非线性 SVM 只是自变量(x_n, z_n)不同，优化过程其实没有区别，都可以用拉格朗日对偶来求解。

这里讨论非线性的情况，假设这是一个非线性(高维空间)可分的 SVM ，总共有 m 个样本，x_i 是 n 维向量，$y = \pm 1$ 为监督学习的标签，高维核函数映射 $z_n = \Phi(x_n)$ ：

$$\begin{cases} \min\limits_{w,b} \frac{1}{2}||w||^2 \triangleq f(w) \\ s.t \quad y^{(i)}(w^T z^{(i)} + b) \geqslant 1, \quad i = 1, \cdots, m \\ \qquad g_i(w) \triangleq 1 - y^{(i)}(w^T z^{(i)} + b) \leqslant 0 \end{cases}$$

目标函数的黑塞矩阵是 n 阶单位矩阵 I ，是严格正定矩阵，因此目标函数是凸函数。可行域是线性不等式围成的区域，因此是一个凸集。这个最优化问题是一个凸优化问题。

根据线性 SVM 的优化问题中，超平面方程存在冗余，即对于一组解 w, b 存在 $y^{(i)}(w^T x^{(i)} + b) \geqslant 1$ ， 则另一组解 $2w, 2b$ 同样成立 $y^{(i)}(2w^T x^{(i)} + 2b) \geqslant 2 > 1$ 。因此 $[Slater]$ 条件也成立，所以 $p^* = d^*$ 。

所以可以将原问题构造拉格朗日乘子函数：

$$\mathcal{L}(w,\alpha,b)=\frac{1}{2}||w||^2-\sum_{i=1}^{m}\alpha_i[y^{(i)}(w^Tz^{(i)}+b)-1]$$

并转为考虑对偶问题：

$$\max_{\alpha\geqslant 0}\min_{w,b}\mathcal{L}(w,\alpha,b)$$

这是一个无约束的优化问题，因此通过求导来解决。先固定 α ，求解式子对于 w,b 的偏导，使得 $\mathcal{L}$ 函数取极小值：

$$\nabla_w\mathcal{L}(w,\alpha,b)=w-\sum_{i=1}^{m}\alpha_i\ y^{(i)}z^{(i)}=0\quad(\text{根据线性回归部分的引理})$$

$$w^*=\sum_{i=1}^{m}\alpha_i^*\ y^{(i)}z^{(i)}\quad\cdots\cdots(1)$$

$$\frac{\partial\mathcal{L}(w,\alpha,b)}{\partial b_i}=0$$

$$\sum_{i=1}^{m}\alpha_i^*\ y^{(i)}=0\quad\cdots\cdots(2)$$

将上面 $(1)(2)$ 式带入原拉格朗日乘子函数，消掉 w,b ，并省略 $*$ 号：

$$\mathcal{L}(w,\alpha,b)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}(z^{(i)})^Tz^{(j)}\alpha_i\alpha_j-b\sum_{i=1}^{m}\alpha_i\ y^{(i)}$$

化简后，可得到如下的求极大值的优化问题：

$$\max_{\alpha}w(\alpha)\triangleq\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}(z^{(i)})^Tz^{(j)}\alpha_i\alpha_j$$

转为最小化问题，并添加约束条件：

$$\begin{cases}\min\limits_{\alpha}w(\alpha)\triangleq-\sum\limits_{i=1}^{m}\alpha_i+\frac{1}{2}\sum\limits_{i,j=1}^{m}y^{(i)}y^{(j)}(z^{(i)})^Tz^{(j)}\alpha_i\alpha_j\\ s.t\quad\alpha_i\geqslant 0\quad i=1,\cdots,m\\ \qquad\sum\limits_{j=1}^{m}\alpha_jy^{(j)}=0\end{cases}$$

通过求解上述问题可得 α^* 。

带入 (1) 式可得 w^* 。

带入原问题可得 b^* 。

$$b^*=-\frac{1}{2}\left(\max_{i:\ y^{(i)}=1}w^{*T}z^{(i)}+\min_{i:\ y^{(i)}=-1}w^{*T}z^{(i)}\right)$$

最后，带入 w^*, b^* ，带入 x 的高维核函数映射，SVM 的分类面为：

$$g_{SVM} = sign(\sum_{i=1}^{m} \alpha_i y^{(i)} z^{(i)T} z + b^*)$$

其中 $sign$ 表示符号函数。

6. 候选支撑向量与支撑向量的确定

在 SVM 的原问题中，如果一个向量满足在分类面的边界上，即 $y^{(i)}(w^T z^{(i)} + b) = 1$ ，则该特征向量 $z^{(i)}$ 为候选支撑向量。使用原问题求解是无法确定哪些是支撑向量的。

在 SVM 的硬间隔对偶问题中，如果 $\alpha_i = 0$ ，则表示样本在预测函数中不起作用。那么如果一个向量其对应的 $\alpha_i > 0$ ，则该特征向量 $z^{(i)}$ 为支撑向量。可以证明，这些向量刚好满足在分类面的边界上(满足 $y^{(i)}(w^T z^{(i)} + b) = 1$)，且

在后面的学习中我们将看到， SVM 的软间隔对偶问题中，只要 $\alpha_i > 0$ ，那么这些向量就都是支撑向量，它们有的可能不一定在最大的分类面上，但是都共同决定了最优的 w^* ($w^* = \sum_i \alpha_i y_i z_i$)

求解 w^* 时，运用核方法，实际上只需要全部的支撑向量求解即可：

$$w = \sum_{support\ vector} \alpha_i\ y^{(i)} z^{(i)}$$

求解 b^* 时，实际上只需要任意一个支撑向量 ($y^{(i)}, z^{(i)}$) 即可求解：

$$b^* = y^{(i)} - w^{*T} z^{(i)}$$

最后得到分类方程：$g_{SVM} = sign(w^{*T} \Phi(x) + b^*)$ 。

7. 关于 SVM 的预测

如果输入是一个矩阵，则映射为： $X^{(m+1)} \to \hat{y}^{(m+1)}$ 。

$$\begin{aligned}
\hat{y}^{(m+1)} &= w^{*T} X^{(m+1)} + b^* \\
&= \left(\sum_{i=1}^{m} \alpha_i^* y^{(i)} (X^{(i)})^T \right) X^{(m+1)} + b^* \\
&= \sum_{i=1}^{m} \alpha_i^* y^{(i)} < X^{(i)}, X^{(m+1)} > + b^*
\end{aligned}$$

8. 不可分的 SVM

即两类点集之间没有明显的间隔，这里只讨论线性情况 x_n ，非线性 z_n 的情况类似，使用核函数映射，替换自变量即可。

8.1. 修改目标函数和约束(软间隔问题)

构造软间隔，即通过松弛变量和惩罚因子对违反不等式的约束进行惩罚：

$$\begin{aligned}&\min \frac{1}{2}||w||^2+\mathbf{C}\sum_{\mathbf{i=1}}^{\mathbf{m}}\xi_{\mathbf{i}} \quad \mathbf{(C>0)}\\ s.t \quad &y^{(i)}(w^Tx^{(i)}+b)\geqslant 1-\xi_{\mathbf{i}}\\ &\xi_i\geqslant 0 \quad i=1,\cdots,m\end{aligned}$$

ξ_i 是松弛变量，如果它不为 0，表示样本违反了不等式的约束条件。C 为惩罚因子，是人工设定的 >0 的常数，用来对违反不等式约束条件的样本进行惩罚。

将原问题的等式和不等式约束都写为标准形式，然后构造拉格朗日乘子函数：

$$\mathcal{L}(w,b,\alpha,\xi,\beta)=\frac{1}{2}||w||^2+C\sum_{i=1}^{m}\xi_i-\sum_{i=1}^{m}\alpha_i[y^{(i)}(w^Tx^{(i)}+b)-1+\xi_i]-\sum_{i=1}^{m}\beta_i\xi_i$$

并转为考虑对偶问题：

$$\max_{\alpha\geqslant 0}\min_{w,b}\mathcal{L}(w,b,\alpha,\xi,\beta)$$

这是一个无约束的优化问题，因此通过求导来解决。先固定 α，求解 w,b，使得 $\mathcal{L}$ 函数取极小值：

$$\begin{aligned}&\nabla_w\mathcal{L}(w,\alpha,b)=w-\sum_{i=1}^{m}\alpha_i\ y^{(i)}x^{(i)}=0 \quad \text{(根据线性回归部分的引理)}\\ &w^*=\sum_{i=1}^{m}\alpha_i^*\ y^{(i)}x^{(i)} \quad \cdots\cdots(1)\\ &\frac{\partial\mathcal{L}(w,b,\alpha,\xi,\beta)}{\partial b_i}=\sum_{i=1}^{m}\alpha_i^*\ y^{(i)}=0 \quad \cdots\cdots(2)\\ &\frac{\partial\mathcal{L}(w,b,\alpha,\xi,\beta)}{\partial \xi_i}=C-\alpha_i-\beta_i=0 \quad \cdots\cdots(3)\end{aligned}$$

将上面 $(1)(2)(3)$ 式带入原拉格朗日乘子函数，消掉 w,b，并省略 $*$ 号：

$$\mathcal{L}(w,\alpha,b)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}(x^{(i)})^Tx^{(j)}\alpha_i\alpha_j-b\sum_{i=1}^{m}\alpha_i\ y^{(i)}$$

化简可得到如下的求极大值的优化问题：

$$\max_{\alpha}w(\alpha)\triangleq\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i,j=1}^{m}y^{(i)}y^{(j)}(x^{(i)})^Tx^{(j)}\alpha_i\alpha_j$$

$\alpha_i+\beta_i=C$ 且 $\beta_i\geqslant 0$，因此有 $\alpha_i<C$。

转为最小化问题，并添加约束条件：

$$\begin{cases}\min\limits_{\alpha} w(\alpha) \triangleq -\sum\limits_{i=1}^{m}\alpha_i + \frac{1}{2}\sum\limits_{i,j=1}^{m} y^{(i)}y^{(j)}(x^{(i)})^T x^{(j)}\alpha_i\alpha_j \\ s.t \quad 0 \leqslant \alpha_i \leqslant C \quad i=1,\cdots,m \\ \qquad \sum\limits_{i=1}^{m}\alpha^{(i)}y^{(i)} = 0\end{cases}$$

通过求解上述问题可得 α^* ； 带入 (1) 可得 w^* ； 带入原问题可得 b^* 。

$$g_{SVM} = sign\big(\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)T}x + b\big)$$

8.2. 三种样本对应的约束条件

定义矩阵 Q ， 其元素为 $Q_{ij} = y^{(i)}y^{(j)}x^{(i)T}x^{(j)}$ 。

对偶问题可以写成矩阵和向量的形式：

$$\begin{cases}\min \frac{1}{2}\alpha^T Q\alpha - e^T\alpha \\ 0 \leqslant \alpha_i \leqslant C \\ y^T\alpha = 0\end{cases}$$

其中 e 是分量全为 1 的向量， y 是样本的类别标签向量。可以证明 Q 是一个半正定矩阵(任意非 0 向量 x ， 满足 $x^TQx \geqslant 0$)， 令 $Q = X^TX, X = (y_1x_1,\cdots,y_mx_m)$ 。

在最优点必须满足 KKT 条件，将条件应用于原问题：

$$\begin{aligned}\min \frac{1}{2}||w||^2 + \mathbf{C}\sum_{\mathbf{i=1}}^{\mathbf{m}}\xi_{\mathbf{i}} \quad \mathbf{(C>0)} \\ s.t \quad y^{(i)}(w^Tx^{(i)}+b) \geqslant 1 - \boldsymbol{\xi_{\mathbf{i}}} \\ \xi_i \geqslant 0 \quad i=1,\cdots,m\end{aligned}$$

上述两个不等式约束必须满足：

$$\begin{aligned}\alpha_i y^{(i)}\big(y^{(i)}(w^Tx^{(i)}+b) - 1 + \xi_i\big) = 0, \quad i=1,\cdots,m \\ \beta_i\xi_i = 0, \quad i=1,\cdots,m\end{aligned}$$

如果 $\alpha_i = 0$ ， 则 $y^{(i)}(w^Tx^{(i)}+b) - 1 + \xi_i = 0$ 没有约束。但是存在 $\alpha_i + \beta_i = C$ ， 因此 $\beta_i = C$ ； 又因为有第二个方程 $\beta_i\xi_i = 0$ 的约束，则必须有 $\xi_i = 0$ 。因此代回原问题的第一个约束条件，则有：

$$y^{(i)}(w^Tx^{(i)}+b) \geqslant 1$$

如果 $\alpha_i > 0$ ， 则有 $y^{(i)}(w^Tx^{(i)}+b) - 1 + \xi_i = 0$ 。而又因为 $\xi_i > 0$ ， 则有：

$$y^{(i)}(w^Tx^{(i)}+b) \leqslant 1$$

对于 $\alpha_i > 0$ 可以分为两种情况：

如果 $0 < \alpha_i < C$，由于 $\alpha_i + \beta_i = C$ 的约束，因此 $\beta_i > 0$；又因为有第二个方程 $\beta_i \xi_i = 0$ 的约束，则必须有 $\xi_i = 0$。由于原问题约束要求：$y^{(i)}(w^T x^{(i)} + b) \geqslant 1 - \xi_i = 0$，但是上面又要满足 $y^{(i)}(w^T x^{(i)} + b) \leqslant 1$，则有：

$$y^{(i)}(w^T x^{(i)} + b) = 1$$

如果 $\alpha_i = C$，由于 $\alpha_i + \beta_i = C$ 的约束，因此 $\beta_i = 0$；前文提到，需满足 $y^{(i)}(w^T x^{(i)} + b) = 1 - \xi_i$，此时 $\xi_i > 0$。如果 $\xi_i < 1$，则满足 $y^{(i)}(w^T x^{(i)} + b) < 1$，此时样本在软间隔内，能正确分类但不满足最大间隔约束；如果 $\xi_i > 1$，则 $y^{(i)}(w^T x^{(i)} + b) < 0$，此时样本被错误分类。如果 $\xi_i = 1$，则 $y^{(i)}(w^T x^{(i)} + b) = 0$，样本刚好落到 $\mathcal{H}: w^T x^{(i)} + b = 0$ 的分类面。这三种样本也都属于支撑向量。

总结三个条件，在最优点处满足：

$$\begin{cases} \alpha_i = 0 \Leftrightarrow y^{(i)}(w^T x^{(i)} + b) \geqslant 1 \\ 0 < \alpha_i < C \Leftrightarrow y^{(i)}(w^T x^{(i)} + b) = 1 \\ \alpha_i = C \Leftrightarrow y^{(i)}(w^T x^{(i)} + b) \leqslant 1 \end{cases}$$

第一种对应自由变量，即非支撑向量。

第二种对应距离超平面刚好为最大间隔 $\frac{1}{||w||_2}$ 的支撑向量。前两种情况都满足 $\xi_i = 0$。

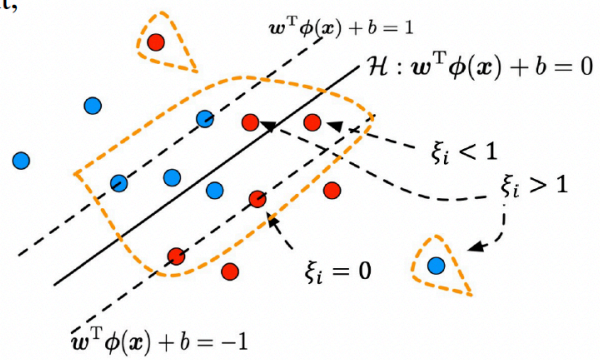
第三种对应违反不等式约束的样本($\xi_i > 0$)，也是支撑向量($\alpha > 0$)。如果 $\xi_i < 1$，样本在软间隔内，如果 $\xi_i = 1$，样本刚好落到 $\mathcal{H}: w^T x^{(i)} + b = 0$ 的分类面上，如果 $\xi_i > 1$，样本将被错误分类。

在 SMO 分治算法中，会应用这个条件来选择性优化变量。

Support vectors are $\boldsymbol{\phi}(\boldsymbol{x}_i)$ such that $\alpha_i^* > 0$.

They are the set of points which satisfy one of the following:
(1) they are tight with respect to the large margin contraint,
(2) they do not satisfy the large margin contraint,
(3) they are misclassified.

- when $\xi_i^* = 0$, $y_i(\boldsymbol{w}^{*\mathrm{T}}\boldsymbol{\phi}(\boldsymbol{x}_i) + b^*) = 1$,
  and thus the point is $1/\|\boldsymbol{w}^*\|_2$ away from the hyperplane.
- when $\xi_i^* < 1$, the point is classified correctly
  but does not satisfy the large margin constraint.
- when $\xi_i^* > 1$, the point is misclassified.



Support vectors (circled with the orange line) are the only points that matter!

对于线性不可分的情况，除了惩罚因子，也可以通过构造核函数的方法升维来解决：

9. 非线性 SVM - 核 SVM 与核方法($Kernel\ methods$)

9.1. 核函数与核方法

前面我们已经提到了核方法对 x_n 进行数据升维，可以解决非线性的 SVM 的问题。除了非线性，核 SVM 也可以解决数据线性不可分的情况。

$Mercer$ 定理(核函数的充分条件)：**任何半正定对称函数都可以作为核函数**(核矩阵 K 的特征值均为非负)。

研究对偶 SVM 的目的是找到一个没有 $\tilde{d}$ 矩阵依赖的非线性 SVM 求解方法。但是在对偶 SVM 中，α^* 的极小值优化问题求解，存在大量的 $z_n z_m^T = \Phi(x_n)^T \Phi(x_m)$ ，如何提高效率呢?

定义：给定一个特征映射 Φ ，定义其对应的核为：

$$\forall x, z \quad K(x,z) = \Phi^T(x)\Phi(z)$$

核函数关心的是二者的内积 (inner product)。

举个例子，假设 $x, z \in R^n$, $K(x,z) \triangleq (x^T z)^2$ 。

定义，假设 $n = 3$ ，三维映射到九维：

$$\Phi\left(\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}\right) = \begin{pmatrix} x_1 & x_1 \\ x_1 & x_2 \\ x_1 & x_3 \\ x_2 & x_1 \\ x_2 & x_2 \\ x_2 & x_3 \\ x_3 & x_1 \\ x_3 & x_2 \\ x_3 & x_3 \end{pmatrix} \in R^{9\times 1}$$

核函数：Φ 变换+内积计算

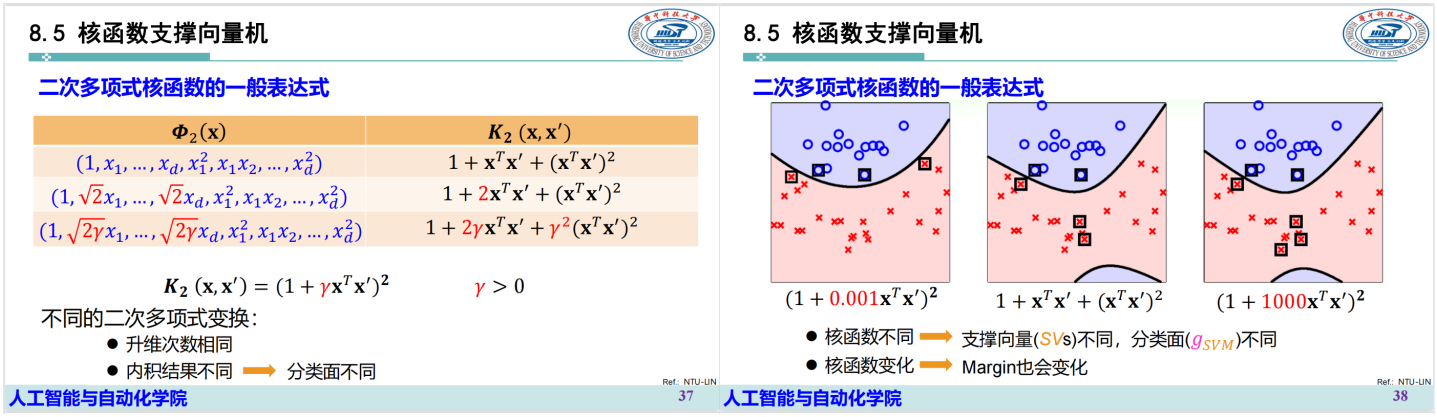
$$\forall x, z \in R^3 \quad K(x,z) = \Phi(x)^T\Phi(z)$$
$$\text{计算复杂度：} \quad O(n) \quad O(n^2)$$

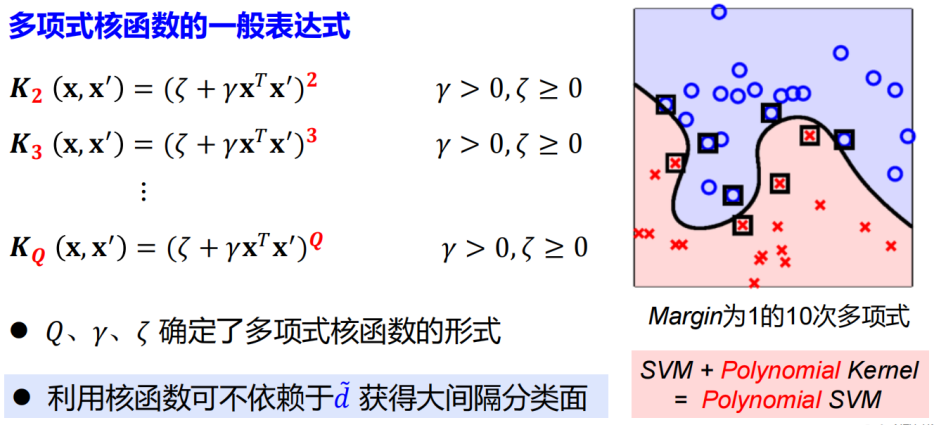
比如说二次多项式的某一标准核函数的计算如下：

$$K_{\Phi_2(x)}(x,x') = \Phi(x)^T\Phi(x') = 1 + x^T x' + (x^T x')(x^T x')$$

$$\begin{aligned} K(x,z) &= \Big(\sum_{i=1}^n x_i z_i\Big)\Big(\sum_{i=1}^n x_i z_i\Big) \\ &= \sum_{i=1}^n \sum_{j=1}^n x_i z_i x_j z_j \\ &= \sum_{i,j=1}^n \Big(x_i x_j\Big)\Big(z_i z_i\Big) \end{aligned}$$

不同核函数的升维方式会导致分类面，支撑向量均发生变化。





9.2. 高斯核函数 SVM

高斯核 SVM 的主要解决的是无穷维的分类问题。它通过求解 α_n 确定所有的支撑向量 z_n ，构造以支撑向量为中心的高斯函数的线性组合，能实现在无穷维中获得最大间隔分类面。

高斯核函数也被称为径向基核函数(RBF):

$$K(x,z)=\exp-\frac{||x-x'||^2}{2\sigma^2},\quad \sigma>0$$

$$g_{SVM}=sign\Big(\sum_{support\ vector}\alpha_n y_n\exp\Big(-\frac{||z-z_n||^2}{2\sigma^2}\Big)+b\Big)$$

该径向基函数也常用于构造径向基神经网络，用于求解局部感受野，在控制领域的神经网络中用的非常多。

10. 非线性 SVM -> SMO 分治算法($Sequential\ Minimal\ Optimization$)

它是求解 SVM 对偶问题的高效算法，核心思想为每次从优化变量中挑选两个分量进行优化，固定其他分量，这样能保证满足等式约束条件。

对于非线性且不可分的 SVM ，我们假设这里使用惩罚因子 C 解决不可分样本问题，同时使用核函数升维解决非线性问题。最后运用拉格朗日对偶法，最终要求解的优化问题可以表示为：

$$\begin{cases} \min\limits_{\alpha} w(\alpha) \triangleq -\sum\limits_{i=1}^{m}\alpha_i + \frac{1}{2}\sum\limits_{i,j=1}^{m} y_i y_j K(x_i, x_j)\alpha_i\alpha_j \\ s.t \quad 0 \leqslant \alpha_i \leqslant C \quad i = 1, \cdots, m \\ \qquad \sum\limits_{i=1}^{m}\alpha^{(i)} y^{(i)} = 0 \end{cases}$$

核函数之前的表示为 $\Phi(x_n)$，这里表示为 $K(x_i, x_j) = K_{ij}$，和 $Mean - Shift$ 算法的表示一致。

这可以看做一个二次规划问题，前面我们讲了 QP 求解方法的局限性，即当样本数 m 很大时，面临着效率低和存储空间太大的问题。所以采用其他方法求解。

SMO 算法每次选择两个变量优化，假设选取的分量为 α_i, α_j，其他分量都固定，当成常数，又因为 $y_i y_i = 1, y_j y_j = 1$，则目标函数写为：

$$f(\alpha_i, \alpha_j) = \frac{1}{2}K_{ii}\alpha_i^2 + \frac{1}{2}K_{jj}\alpha_j^2 + sK_{ij}\alpha_i\alpha_j + y_i v_i \alpha_i + y_j v_j \alpha_j - \alpha_i - \alpha_j + c$$

其中，c 是一个常数，$s = y_i y_j$，$v_i = \sum\limits_{k=1, k\neq i,j}^{m} y_k \alpha_k^* K_{ik}$，$v_j = \sum\limits_{k=1, k\neq i,j}^{m} y_k \alpha_k^* K_{jk}$。

这里的 α^* 为 α 在上一轮迭代后的值。该子问题的目标函数是二元二次函数，可以直接给出最小值的解析解。

约束条件为：$0 \leqslant \alpha_i \leqslant C, 0 \leqslant \alpha_j \leqslant C$，以及 $y_i\alpha_i + y_j\alpha_j = -\sum\limits_{k=1, k\neq i,j}^{m} y_k\alpha + k = \xi$。

利用上面的等式约束可以消掉目标函数里的 α_i，从而只剩下变量 α_j，从而进一步直接求得目标函数的解析解。

11. 总结(不同的 SVM 优化选择)

线性和非线性取决于分类面是直线还是曲线，即分类面一次拟合还是多项式拟合。

可分和不可分取决于数据是否能直接通过一条直线划分全部，即数据是硬间隔还是软间隔。

数据只存在线性可分和线性不可分(非线性)的区别，对于不可分情况，要么构造升维非线性 SVM，要么通过松弛变量和惩罚因子构造线性软间隔。

非线性 SVM 就可以解决数据不可分问题，但是带惩罚因子的线性 SVM 也可以解决数据不可分问题。

数据的分布以及复杂程度，决定了不同的优化方式。

对于线性可分的 SVM，通过一阶拉格朗日乘子法，QP 求解器，合页损失($Hinge\ Loss$)梯度下降，可以实现。

什么时候用对偶 SVM？当原问题不好直接求解时，可以先将原问题转为对偶问题，并构造拉格朗日乘子函数，进行求解。不管是线性还是非线性 SVM，都可以使用拉格朗日对偶法来解决。

对于线性不可分的 SVM，可以在拉格朗日对偶法的基础上，构造软间隔 -> 即添加松弛变量 ξ_i 和惩罚因子 C 来解决。

对于非线性的 SVM，QP 求解器存在一定局限，通常在拉格朗日对偶法的基础上，先构造核函数 $z_n = \Phi(x_n)$ 的映射(核 SVM)，然后一样构造拉格朗日函数求解对偶问题。

对于非线性的 SVM ，可以通过同时构造拉格朗日对偶，核函数，惩罚因子来进行求解。使用 SMO 分治策略能让这个优化问题的求解变得更简单。