Author：Liu Jian

Time：2020-04-25

机器学习3-SVM

平面： $\boldsymbol{w}^T\boldsymbol{x}+b=0$

平面的法向量:

思路一:
点 $\boldsymbol{x}_1$ 和点 $\boldsymbol{x}_2$ 在平面上，则有:

$$\begin{cases}\boldsymbol{w}^T\boldsymbol{x}_1+b=0\\ \boldsymbol{w}^T\boldsymbol{x}_2+b=0\end{cases}\Rightarrow \boldsymbol{w}^T(\boldsymbol{x}_1-\boldsymbol{x}_2)=0$$

可知， $\boldsymbol{w}$ 与平面内任意向量 $\boldsymbol{x}_1-\boldsymbol{x}_2$ 垂直，则 $\boldsymbol{w}$ 就是平面的法向量。

思路二：上升一个维度，构造函数 $y=f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b$，平面 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 为函数的一个等值面。易知，函数的梯度向量为 $\boldsymbol{w}$，梯度向量是使函数值上升最快的方向，与等值面垂直，则 $\boldsymbol{w}$ 是平面 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 的法向量，且指向使 $\boldsymbol{w}^T\boldsymbol{x}+b$ 值变大即 $\boldsymbol{w}^T\boldsymbol{x}+b>0$ 的那一侧。

点到平面距离:
$\boldsymbol{x}_1$ 为平面内一点， $\boldsymbol{x}^*$ 为平面外一点，点 $\boldsymbol{x}^*$ 到平面的距离:

$$\frac{\left|\boldsymbol{w}^T(\boldsymbol{x}^*-\boldsymbol{x}_1)\right|}{||\boldsymbol{w}||}=\frac{\left|\boldsymbol{w}^T\boldsymbol{x}^*+b\right|}{||\boldsymbol{w}||}$$

1 硬间隔最大化 SVM

1.1 模型的构建

前提条件：数据可分

注意，标记 $y=-1\text{ or }+1$，不要认为 $y=\boldsymbol{w}^T\boldsymbol{x}+b$。 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 可以看做一个平面，也可以看作一个函数，要是看成函数，则是一个 $(n-1)$ 维的函数(假设 $\boldsymbol{x}$ 是一个 n 维向量)。实际上，将平面上升一个维度令 $f(\boldsymbol{x})=\boldsymbol{w}^T\boldsymbol{x}+b$，标记 $y=\text{sign}(f(\boldsymbol{x}))$：

函数间隔:

超平面 $(\boldsymbol{w},b)$ 关于**样本点** $(\boldsymbol{x}_i,y_i)$ **的函数间隔**:

$$\hat{\gamma}_i = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$$

超平面 $(\boldsymbol{w}, b)$ 关于**数据集的函数间隔**：

$$\hat{\gamma} = \min_{i=1,\cdots,N} \hat{\gamma}_i$$

其中， N 为数据集中样本点的个数。因为前提条件假设线性可分，在超平面上方的点的标记为 $+1$ ，在超平面下方的点的标记为 -1 ，可知乘以标记 y_i 的作用是使函数间隔为正。

若 $(\boldsymbol{w}, b)$ 成比例变化，超平面没有变化，但函数间隔会成比例地变化，下面通过归一化，引入保持不变的几何间隔。

几何间隔：

超平面 $(\boldsymbol{w}, b)$ 关于**样本点 $(\boldsymbol{x}_i, y_i)$ 的几何间隔**：

$$\gamma_i = \frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} = \frac{\hat{\gamma}_i}{\|\boldsymbol{w}\|}$$

超平面 $(\boldsymbol{w}, b)$ 关于**数据集的几何间隔**：

$$\gamma = \min_{i=1,\cdots,N} \gamma_i = \frac{\gamma_i}{\|\boldsymbol{w}\|}$$

支持向量机(SVM)：

SVM的思想就是将数据集的几何间隔最大化： $\max_{\boldsymbol{w},b}\gamma$ ，又 $\gamma = \min_{i=1,\cdots,N}\gamma_i$ ，即：

$$\max_{\boldsymbol{w},b} \gamma$$
$$\text{s.t. } \gamma_i = \frac{y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)}{\|\boldsymbol{w}\|} \geqslant \gamma, \text{ for all } i$$

上式改写成函数间隔的形式：

$$\max_{\boldsymbol{w},b} \gamma = \frac{\hat{\gamma}}{\|\boldsymbol{w}\|}$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant \hat{\gamma}, \text{ for all } i$$

由 $(\boldsymbol{w}, b)$ 可成比例地变化，则不妨令函数间隔 $\hat{\gamma} = 1$ ，则**SVM可最终表示为求解一个如下的凸二次规划问题**：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant 1, \text{ for all } i$$

凸二次规划问题：凸优化问题中，目标函数是二次函数，不等式约束函数是仿射函数。

SVM的性质：

- 若数据集线性可分，则SVM存在且唯一；
- 存在两条与分离超平面平行的超平面，分别位于分离超平面两侧，给出了正负两类样本距离分离超平面最近的边界(即， $\boldsymbol{w}^T\boldsymbol{x}_i + b = \pm 1$)。那些少量的在间隔边界上的样本点被称为**支持向量**，分离超平面只与这些支持向量有关，因此，这种算法叫支持向量机。两条间隔边界到分离超平面的距离相等(由反证法易知)，大小为 $\frac{1}{\|\boldsymbol{w}\|}$ 。

1.2 模型的求解

线性可分，凸二次优化，则强对偶性成立，KKT条件为充要条件。

选择求解凸二次优化的对偶问题。拉格朗日函数：

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

原变量是 $(\boldsymbol{w}, b)$，对偶变量为拉格朗日乘子 $\boldsymbol{\alpha} \geqslant \mathbf{0}$。

对偶问题：

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \ \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geqslant 0, \ \text{for all } i$$

记对偶变量和原变量的解分别为 $\boldsymbol{\alpha}^*, (\boldsymbol{w}^*, b^*)$，解得 $\boldsymbol{\alpha}^*$ 后(可采用SMO算法计算)，**由原问题的 KKT 条件**：

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

对于偏移量 b^*，**由原问题的 KKT 条件**，对任意支持向量 $(\boldsymbol{x}_j, y_j)$（$\alpha_j^* > 0$），均有 $y_j(\boldsymbol{w}^{*T}\boldsymbol{x}_j + b^*) = 1$，则：

$$b^* = \frac{1}{y_j} - \sum_{i=1}^{N} \alpha_i^* y_i \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle$$

可以看到，$\boldsymbol{w}^*$ **也只与那些拉格朗日乘子非零的样本点即支持向量有关。而偏移量的解不唯一，更鲁棒的做法是考虑所有支持向量的结果然后取平均。**

2 软间隔最大化 SVM

没有第一章中硬间隔最大化SVM的中线性可分的要求，函数间隔大于等于 1 的条件被放宽，也允许有误分类的样本。

2.1 模型的构建

思路：硬间隔最大化SVM模型中要求所有样本点都满足函数间隔大于等于 1 的约束，软间隔最大化SVM模型则放宽了这一要求，去掉了函数间隔均大于等于 1 的要求，但对那些小于 1 的样点进行了惩罚，得到模型如下：

$$\min_{\boldsymbol{w}, b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N} l_{0/1}(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)$$

其中，$C > 0$ 为给定的惩罚参数，$l_{0/1}(\cdot)$ 为如下的"0/1损失函数"：

$$l_{0/1}(z) = \begin{cases} 1, & z < 0 \\ 0, & z \geqslant 0 \end{cases}$$

一般地，对于分类问题，正确分类($z \geqslant 0$)时，损失为 0；错误分类($z < 0$)时，损失为 1。当函数间隔 $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 0$ 时，点被正确分类，但这里并不是令 $z = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$，而是令 $z = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1$，因为我们要求函数间隔大于等于 1 时才令损失为 0。

“0/1损失函数”是二分类问题真正的损失函数，但它非凸、非连续，数学性质不太好，不易于求解。因此，人们通常会选用一些凸的连续函数作为替代损失函数：

$$\text{hinge损失：} l_{hinge}(z) = \max(0, 1-z)$$
$$\text{指数损失：} l_{exp}(z) = \exp(-z)$$
$$\text{对率损失：} l_{log}(z) = log(1+\exp(-z))$$

hinge损失函数(也称合页损失函数)，由其函数图像可知，点被正确分类且间隔大于等于 1 ($z \geqslant 1$)时损失函数才为 0 ，即被正确分类的确信度足够高时才没有损失。此时，令 $z = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)$ ，模型变为：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))$$

可以看到， $\frac{1}{2}\|\boldsymbol{w}\|^2$ 为正则化项，用来描述划分超平面的“间隔”大小；
$C\sum_{i=1}^{N}\max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))$ 为经验损失/风险，用来表述训练集上的误差；两者相加并最小化就是结构风险最小化。

当然，也可以采用上述其他的损失函数构建模型，但因为它们没有损失完全为 0 的点，因此每个样本点都会对模型产生影响，因此都是“支持向量”；而合页损失函数存在损失为完全为 0 的区域，因此最终只有少量的点会对模型产生影响，这使得支持向量的解具有稀疏性。

虽然上面我们构建出了一个无约束的凸优化问题，但因为目标函数不可导，问题仍然不便于求解。为此，我们可以**引入松弛变量 $\boldsymbol{\xi}$ ，问题等价于**：

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b) \geqslant 1-\xi_i$$
$$\xi_i \geqslant 0, \text{ for all } i$$

其中， $C > 0$ 为给定的惩罚参数。上述问题仍然是一个凸二次规划问题，可证明解存在， $\boldsymbol{w}$ 的解唯一，但 b 的解存在于一个区间。**两个问题等价的证明**：实际上，令 $\xi_i' = \max(0, 1-y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b))$ ，以 ξ_i' 为纵坐标， $y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)$ 为横坐标，就得到了 l_{hinge} 函数曲线。而等价模型中的 ξ_i 由其不等式关系可取 l_{hinge} 函数曲线及上方所有区域，而通过取最小化 $\min_{\boldsymbol{w},b,\boldsymbol{\xi}}$ 的操作，使得等价模型中的 ξ_i 实际上也为 l_{hinge} 的函数曲线。因此，两个模型等价。

2.2 模型的求解

原问题：

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi}} \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{N}\xi_i$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b) \geqslant 1-\xi_i$$
$$\xi_i \geqslant 0, \text{ for all } i$$

仍然是一个凸二次规划问题。通过求解对偶问题来求解原问题。

构造拉格朗日函数：

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

原变量是 $(\boldsymbol{w}, b, \boldsymbol{\xi})$ ，对偶变量为拉格朗日乘子 $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 且非负。令拉格朗日函数对原变量的导数为 0 ，可得：

$$\begin{cases} \boldsymbol{w} = \sum_{i=1}^{N} \alpha_i y_i \boldsymbol{x}_i \\ \sum_{i=1}^{N} \alpha_i y_i = 0 \\ \alpha_i + \beta_i = C, \qquad i = 1, \cdots, N \end{cases}$$

将上述关系代入拉格朗日函数中，消除原变量并化简，得**对偶问题**：

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle + \sum_{i=1}^{N} \alpha_i$$

$$\text{s.t.} \ \sum_{i=1}^{N} \alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C, \ \text{for all} \ i$$

对比硬间隔最大化SVM，区别只在于多了 $\alpha_i \leqslant C$ 的限制。上式中只含有拉格朗日乘子 α_i ，拉格朗日乘子 β_i 因为存在关系 $\alpha_i + \beta_i = C$ 而被消去。解得对偶问题后，由原变量与对偶变量之间的关系可得原变量。

硬间隔最大化SVM中，样本点的拉格朗日乘子非零时，不等式约束起作用，样本点在间隔边界上，样本为支持向量。**软间隔最大化SVM中的支持向量的情况更加复杂，解读如下**：

首先，我们明确，拉格朗日乘子： $0 \leqslant \alpha_i \leqslant C$ ，对应约束： $1 - \xi_i - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) \leqslant 0$ ；拉格朗日乘子： $0 \leqslant \beta_i \leqslant C$ ，对应约束： $-\xi_i \leqslant 0$ ，且 $\alpha_i + \beta_i = C$ 。先考虑两个极端情况：

1. $\alpha_i = 0$ ， $\beta_i = C$ 。由KKT条件， α_i 对应不等式约束不起作用： $1 - \xi_i - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) < 0$ ； β_i 对应不等式约束起作用: $-\xi_i = 0$ 。则，此时 $1 < y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b)$ ，样本点落在间隔外且被正确分类，对模型没有影响。
2. $\alpha_i = C$ ， $\beta_i = 0$ 。由KKT条件， α_i 对应不等式约束起作用： $1 - \xi_i - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 0$ ； β_i 对应不等式约束不起作用: $-\xi_i < 0$ 。则，此时 $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 - \xi_i$ ，需进一步分类讨论如下：
    1. $0 < \xi_i \leqslant 1$ ，则函数间隔 $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 - \xi_i \in [0, 1)$ ，样本点落在间隔边界内且被正确分类，对模型有影响，是支持向量。
    2. $1 < \xi_i$,则函数间隔 $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1 - \xi_i < 0$ ，样本点被 SVM 错误分类，间隔内外都有可能分布，对模型有影响，是支持向量。
3. $0 < \alpha_i < C$ ， $0 < \beta_i < C$ 。由KKT条件，此时对应约束都起作用： $1 - \xi_i - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 0, \xi_i = 0$ 。可知，函数间隔 $y_i(\boldsymbol{w}^T \boldsymbol{x}_i + b) = 1$ ，样本点落在间隔边界上且被正确分类，是支持向量。

可以看到，支持向量对应的拉格朗日乘子 $0 < \alpha_i \leqslant C$ 。

3 核技巧

硬/软间隔最大化SVM统称线性支持向量机，对非线性分类问题，可基于对偶问题通过使用核技巧来构建非线性支持向量机。

一个例子：
若训练数据由 $X_1^2 + X_2^2 = 1$ (对应标记 $y = -1$)和 $X_1^2 + X_2^2 = 9$ (对应标记 $y = +1$)生成，上述线性SVM就无法对此进行分类，但事实上分离面 $X_1^2 + X_2^2 = 4$ 是可以很好地对样本进行分类。若将原空间中的点 (X_1, X_2) 映射为新空间中的点 $(Z_1, Z_2) = (X_1^2, X_2^2)$ ，那么在新空间中就可以构建线性SVM对数据进行分类，这里的原空间叫做**输入空间**，记为 $\mathcal{X}$ ，新空间叫做**特征空间**，记为 $\mathcal{H}$ 。事实上，机器学习都是在特征空间中进行的，只不过有时候输入空间无需再做映射，其本身就是特征空间。

由上面的例子可以看到，通过合适的**非线性映射** $\phi(\cdot)$ ，将输入空间样本点 (上例中 $\boldsymbol{x}=(X_1,X_2)$) 映射到特征空间中 (上例中 $\boldsymbol{z}=(Z_1,Z_2)$)，就可在特征空间中使用线性SVM解决输入空间中的非线性分类问题 (输入空间中的分离超曲面模型对应特征空间中的分离超平面模型)，相应的公式如下：

对偶问题：

$$\begin{aligned}\min_{\boldsymbol{\alpha}} \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle-\sum_{i=1}^{N}\alpha_i\\ \text{s.t.}\ \sum_{i=1}^{N}\alpha_i y_i=0\\ 0\leqslant\alpha_i\leqslant C,\ \text{for all}\ i\end{aligned}$$

解上述方程，得对偶变量的解 $\boldsymbol{\alpha}^*$ 后，由原问题的 KKT 条件可得原变量 $\boldsymbol{w}$ 的解：

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

对于偏移量的解 b^* ，**由原问题的 KKT 条件**，对任意 $C>\alpha_j^*>0$ 的支持向量 $(\boldsymbol{x}_j,y_j)$ ，均有 $y_j(\boldsymbol{w}^{*T}\phi(\boldsymbol{x}_j)+b^*)=1$ ，则：

$$b^*=\frac{1}{y_j}-\boldsymbol{w}^{*T}\boldsymbol{x}_j=\frac{1}{y_j}-\sum_{i=1}^{N}\alpha_i^* y_i\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$$

可以看到，**$\boldsymbol{w}^*$ 的解只与支持向量 ($0<\alpha_i^*\leqslant C$ 的样本点)有关。而偏移量 b^* 不唯一，更鲁棒的做法是对所有 $0<\alpha_j^*<C$ 的支持向量的结果取平均。事实上，$\alpha_j^*=C$ 对应的样本点也是支持向量，但此时 $y_j(\boldsymbol{w}^{*T}\boldsymbol{x}_j+b^*)=1-\xi_j^*$ ，关系式中还存在待求松弛变量 ξ_j^* ，而我们要得到 ξ_j^* 会先求 b^* ，再代入到关系式中解出 ξ_j^* ，ξ_j^* 随 b^* 的解变化而变化，并不唯一。**

可得**分离超平面**：

$$\boldsymbol{w}^{*T}\boldsymbol{x}+b^*=\sum_{i=1}^{N}\alpha_i^* y_i\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x})\rangle+b^*=0$$

决策函数：

$$y=\operatorname{sign}(\boldsymbol{w}^{*T}\boldsymbol{x}+b^*)=\operatorname{sign}\left(\sum_{i=1}^{N}\alpha_i^* y_i\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x})\rangle+b^*\right)$$

一般从输入空间映射到的是高维空间(但上面的例子只是二维到二维)，特征空间维数可能很高，或者是无穷维。因此，先将输入空间中的点映射到特征空间中，再对此构建线性SVM是不可行的，除此之外，根据样本数据的特征找到合适的映射 $\phi(\cdot)$ 难度非常大，也不可行。从上述公式中可以看到，在求解对偶变量和计算分离超平面的过程中，只涉及到数据点内积的运算：$\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$ 、$\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x})\rangle$ 。为此，我们引入**核函数 (Kernel Function)** $K(\cdot,\cdot)$ 以避免上述问题。我们把计算两个向量在映射后空间中的内积的函数叫做核函数：

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\langle\phi(\boldsymbol{x}_i),\phi(\boldsymbol{x}_j)\rangle$$

即 $\boldsymbol{x}_i$ 与 $\boldsymbol{x}_j$ 在特征空间的内积等于它们在原始空间中通过函数 $K(\cdot,\cdot)$ 计算的结果，显然 $K(\cdot,\cdot)$ 是一个**对称函数**，$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=K(\boldsymbol{x}_j,\boldsymbol{x}_i)$ 。注意上式中的“等号”不是“定义为”的意思，等式的左右两边分别表示计算内积的两种方式：一种 (等式右边) 是先映射到高维空间中，然后再根据内积的公式进行计算；而另一种 (等式左边) 则直接在原来的低维空间中进行计算，而不需要显式地写出映射后的结果。可

以看到，当遇到映射后的维度很高或是无限维时，采用等式左边的方法已经很难或无法计算的情况下，使用核函数依旧能从容处理。

映射 $\phi(\cdot)$ 和核函数 $K(\cdot,\cdot)$ 的关系：对于给定的核 $K(\cdot,\cdot)$，特征空间 $\mathcal{H}$ 和映射函数 $\phi(\cdot)$ 的取法并不唯一，可以取不同的特征空间，即便在同一特征空间里也可以取不同的映射。举例见李航《统计学习方法》P117。

在实际操作中，我们只定义核函数 $K(\cdot,\cdot)$，学习与预测是隐式地在特征空间中进行而不需要显示地定义特征空间和映射函数 $\phi(\cdot)$，而我们选择的核函数是否有效则需要实验来验证。几种常用的核函数如下：

1. 线性核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\boldsymbol{x}_i^T\boldsymbol{x}_j$
2. 多项式核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=(\boldsymbol{x}_i^T\boldsymbol{x}_j+1)^d, d\in\mathbb{N}^+$
3. 高斯核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right)$，$\sigma>0$ 为高斯核的带宽
4. 拉普拉斯核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_j\|}{\sigma}\right)$，$\sigma>0$
5. Sigmoid核函数：$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\tanh\left(\beta\boldsymbol{x}_i^T\boldsymbol{x}_j+\theta\right)$，$\tanh(\cdot)$ 为双曲正切函数，$\beta>0$，$\theta<0$

线性核核存在的主要目的是使得“映射后空间中的问题”和“映射前空间中的问题”两者在形式上统一起来。前面提到过的输入空间映射为无穷维空间对应核函数的一种情况就是高斯核函数。

另外，举一个映射为无穷维空间的例子：$\phi:\boldsymbol{x}\longrightarrow K(\cdot,\boldsymbol{x})$，$\phi$ 将向量 $\boldsymbol{x}$ 映射为一个函数 $K(\cdot,\boldsymbol{x})$。函数 $K(\cdot,\boldsymbol{x})$ 为无穷维的理解如下：一个向量和一个函数其实是很类似的，一个 N 维向量可以看成是一个定义域为 $\{1,2,\ldots,N\}$ 的函数，反过来，我们平时看到的函数可以看成是（不可数）无穷维的向量。广义的向量的定义并不要求是有限维的，只要满足向量空间的性质即可。

核函数还有如下性质：

1. 若 K_1 和 K_2 为核函数，则对于任意正数 γ_1、γ_2，其线性组合 $\gamma_1K_1+\gamma_2K_2$ 也是核函数；
2. 若 K_1 和 K_2 为核函数，则核函数的直积 $K_1\otimes K_2(\boldsymbol{x}_i,\boldsymbol{x}_j)=K_1(\boldsymbol{x}_i,\boldsymbol{x}_j)K_2(\boldsymbol{x}_i,\boldsymbol{x}_j)$ 也是核函数
3. 若 K_1 为核函数，则对于任意函数 $g(\cdot)$，$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=g(\boldsymbol{x}_i)K_1(\boldsymbol{x}_i,\boldsymbol{x}_j)g(\boldsymbol{x}_j)$ 也是核函数。

接下来是拓展的比较理论的一些内容，只给出结论，了解其意义就行了：

通常所说的核函数就是正定核函数(positive definite kernel function)。

核函数的充要条件：表明只要一个对称函数所对应的核矩阵 (Kernel Gram Matrix) 半正定，它就能作为核函数使用。

表示定理：表明对一般的损失函数和单调递增的正则化项，优化问题的最优解都可以表示为核函数的线性组合，这条定理显示出了核函数的巨大威力。

输入空间 $\mathcal{X}$：可以是欧式空间 $\mathbb{R}^n$ 或离散集合

特征空间 $\mathcal{H}$：是一个希尔伯特空间 (欧氏空间的一个推广，简单地说就是完备的内积空间)，更具体的是“再生核希尔伯特空间 (Reproducing Kernel Hilbert Space, RKHS)”。事实上，任意给定一个核函数 $K(\cdot,\cdot)$，都能找到一个与之对应的映射 ϕ，都可以生成一个再生核希尔伯特空间，生成方式见李航《统计学习方法》P118-P121。

实际上，若映射后的特征空间是无限维，那么SVM的推广并不是像特征空间是有限维时那么直接，博文 支持向量机：Kernel II 中给出了映射后的空间是无限维时SVM的推导过程。

符号解释：
$f(\cdot)$ 表示一个函数，点号是占位符，表示函数的参数，而 $f(x)$ 有时候表示函数，就是 $f(\cdot)$ 的意思，有时候表示一个"数"，此时 x 是一个固定的值而不是未知量。 $K(\cdot, x)$ 表示 $K(\cdot,\cdot)$ 这个二元函数固定第二个参数为 x （定值）之后得到的一元函数，比如 $f = g(x,\cdot)$ 就是 $f(y) = g(x,y)$ 的意思。另一个重要的应用场景是卷积运算中。比如， $f(x,\cdot) * g(\cdot,x)$ means "convolve f and g in their second arguments, with the first arguments fixed as x and y respectively". It might be written as $f(x,z) *_z g(z,x)$ 。

4 序列最小最优化 (SMO) 算法

待优化的对偶模型：

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

$$\text{s.t.} \quad \sum_{i=1}^{N}\alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C, \quad \text{for all } i$$

虽然可以使用通用的二次规划算法来求解上述问题，但该问题的规模正比于训练样本数，但训练样本容量很大时，会存在低效甚至无法使用的问题。因此，我们根据问题的特点，提出了序列最小最优化(SMO)算法。

算法思路：类似于坐标下降法，在 $\boldsymbol{\alpha}$ 给定初始值后，因为存在 $\sum_{i=1}^{N}\alpha_i y_i = 0$ 的限制，对 $\boldsymbol{\alpha}$ 每次只优化其中的两个分量 α_i 和 α_j ，而其他分量保持不变，即 $y_i\alpha_i + y_j\alpha_j = const$ 。优化问题就变为只关于变量 α_i 的二次优化问题，并可推得解析解，十分高效。有了上述思路后还需要解决的问题是：

1. 每次更新如何选取 α_i 和 α_j ?
2. 更新计算何时停止?

我们借助KKT条件解决这两个问题。前面提到，此时，KKT条件是原变量和对偶变量最优点的充要条件。那么，对于第二个问题，当结果满足KKT条件时，停止更新。对于第一个问题，我们选择 α_j 为此时违反如下KKT条件最严重的的点：

$$\text{样本点}(\boldsymbol{x}_j, y_j)\text{，对应的拉格朗日乘子}\alpha_j$$

$$\text{由KKT条件可知}\begin{cases}\alpha_j = 0\text{时，} y_j(\boldsymbol{w}^T\boldsymbol{x}_j + b)\text{应该大于}1\\ \alpha_j = C\text{时，} y_j(\boldsymbol{w}^T\boldsymbol{x}_j + b)\text{应该小于}1\\ \alpha_j = 0\text{时，} y_j(\boldsymbol{w}^T\boldsymbol{x}_j + b)\text{应该等于}1\end{cases}$$

那么我们选取不满足上述条件，且违反程度(可采用 $|y_j(\boldsymbol{w}^T\boldsymbol{x}_j + b) - 1|$ 来衡量)最大的点作为 α_j 。选定 α_j 后， α_i 的选择标准是希望优化前后 α_i 有足够大的变化，这可由解析公式作相应的计算来判断，不再赘述，可参见李航《统计学习方法》。直观上看，采用上述 α_i 和 α_j 的选择策略，能较快地使目标函数变小，从而快速地收敛到最优点。

5 支持向量回归 SVR

5.1 SVR 的构建

我们首先从一般的回归视角推导 SVR，再探讨 SVR 与 SVM 之间的关系。一般地，回归模型 $f(\boldsymbol{x};\boldsymbol{\theta}) : \mathbb{R}^n \to \mathbb{R}$ ，其中 $\boldsymbol{x}$ 为 n 维输入， $\boldsymbol{\theta}$ 为待定参数。给定样本集 $\{(\boldsymbol{x}_i, y_i)_{i=1}^N\}$ ，学习就是要解决如下优化问题：

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} l(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i)$$

其中 $l(\cdot,\star)$ 为损失函数，上式即为经验风险最小化。一般地，我们还会加入正则项，即结构风险最小化，如 L_2 正则化：

$$\min_{\boldsymbol{\theta}} \sum_{i=1}^{N} l(f(\boldsymbol{x}_i;\boldsymbol{\theta}), y_i) + C\|\boldsymbol{\theta}\|_2^2$$

对于 SVR，模型为线性回归模型： $f(\boldsymbol{x};\boldsymbol{w},b)=\boldsymbol{w}^T\boldsymbol{x}+b$ ，损失函数取 ϵ -不敏感误差函数 (ϵ -insensitive error function， ϵ 为给定正数)：

$$l_\epsilon(f(\boldsymbol{x};\boldsymbol{w},b),y)=\begin{cases}0, & |f(\boldsymbol{x};\boldsymbol{w},b)-y|\leqslant\epsilon\\ |f(\boldsymbol{x};\boldsymbol{w},b)-y|, & |f(\boldsymbol{x};\boldsymbol{w},b)-y|>\epsilon\end{cases}$$

加入正则化项 $\|\boldsymbol{w}\|_2^2$ ，优化问题为：

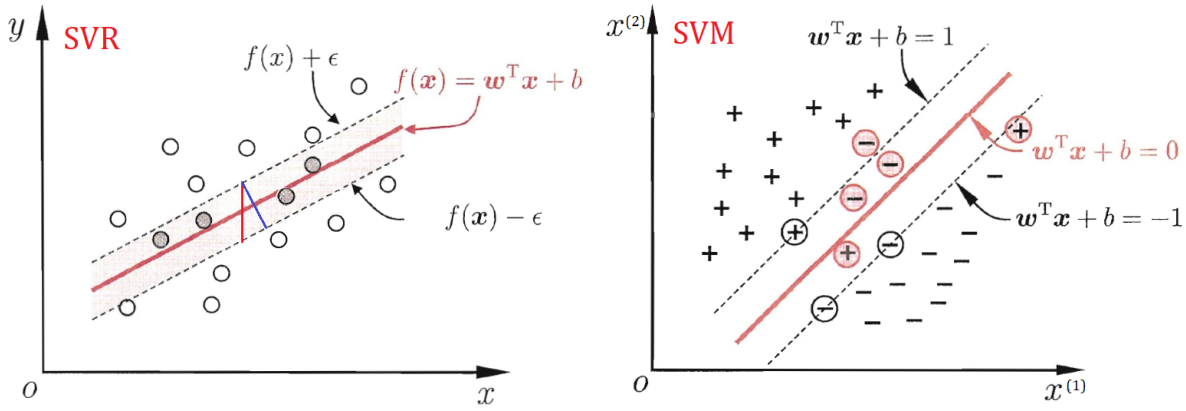
$$\min_{\boldsymbol{w},b} \sum_{i=1}^{N} l_\epsilon(f(\boldsymbol{x}_i;\boldsymbol{w},b),y_i) + C\|\boldsymbol{w}\|_2^2$$

$$\Updownarrow$$

$$\min_{\boldsymbol{w},b} C\sum_{i=1}^{N} l_\epsilon(f(\boldsymbol{x}_i;\boldsymbol{w},b),y_i) + \frac{1}{2}\|\boldsymbol{w}\|_2^2$$

可以看到，SVR 实际上就是损失函数取 ϵ -不敏感误差函数的结构风险最小化。

接下来，我们来探讨 SVR 与 SVM 的关系。SVR 与 SVM示意图如下：



1. SVR 寻找的是一个拟合函数 $f(\boldsymbol{x};\boldsymbol{w},b)=\boldsymbol{w}^T\boldsymbol{x}+b$ ，而 SVM 寻找的只是一个超平面 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ ，注意，二者的维度是不同的。SVR 拟合函数表示的超平面为 $y-\boldsymbol{w}^T\boldsymbol{x}-b=0$ ，相比 SVM 的 $\boldsymbol{w}^T\boldsymbol{x}+b=0$ 多出了一个维度。SVR 与 SVM 的示意图看起来相似，但实际上二者的坐标轴都不同，SVR 的坐标轴分别为 $\boldsymbol{x},y$ ，而 SVM 的坐标轴则为输入 $\boldsymbol{x}$ 的各分量 $x^{(1)},x^{(2)}$ 。
2. SVR 取损失函数为 ϵ -insensitive error function，这意味着当预测值 $f(\boldsymbol{x};\boldsymbol{w},b)$ 与样本观察值相差不超过 ϵ 时，我们认为损失为 0 ，也就是落在超平面 $y=\boldsymbol{w}^T\boldsymbol{x}+b+\epsilon$ 和超平面 $y=\boldsymbol{w}^T\boldsymbol{x}+b-\epsilon$ 所形成的通道中的样本点的损失为 0 。注意，通道的宽度，即这两个超平面间的距离 Δ (SVR 示意图中蓝色线段的长度) 并不是 2ϵ ，与 y 轴平行的红色线段长度才为 2ϵ ：

$$\Delta=\frac{2\epsilon}{\sqrt{1^2+\|\boldsymbol{w}\|_2^2}}$$

又 ϵ 为给定正数，则 $\boldsymbol{w}$ 越小，通道就越宽，最宽为 2ϵ ，此时超平面水平。

3. 从结构风险最小化的角度来看，SVM 和 SVR 中的 $\|\boldsymbol{w}\|_2^2$ 就是正则化项，而其在几何上则表征了间隔大小。我们希望经验损失尽量小的同时，正则化项也尽量小，也就是间隔尽量大。特别地，因为样本可分，硬间隔 SVM 的经验损失可为 0，因此，经验损失并没有出现在优化目标中，而是以约束的形式给出，即：

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$\text{s.t. } y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geqslant 1, \text{ for all } i$$

4. 软间隔 SVM 中损失函数取 l_{hinge}，而 SVR 中损失函数取 l_ϵ。可以看到，相比平方损失和绝对损失，二者均存在一段为 0 的区域，这就使得最终只有少量的样本点会对模型产生影响，而这些点就被称为支持向量。除了正则化，支持向量的存在也会有助于稀疏解的产生。事实上，$\boldsymbol{w}$ 最优解的形式为 $\sum_{i=1}^N \mu_i \boldsymbol{x}_i$，其中 $\mu_i \geqslant 0$，又因为只有少量支持向量 (对应 $\mu_i > 0$) 才会对最终结果产生影响，这意味着向量 $\boldsymbol{\mu}$ 是稀疏的。可以看到，**软间隔 SVM 和 SVR 从正则化项和损失函数两个方面保证了解的稀疏性。**

5. 总而言之；
    - 硬间隔 SVM 就是在保证经验损失为 0 的情况下最大化间隔，落在边界上的样本点为支持向量；
    - 软间隔 SVM 就是希望经验损失尽量小的同时，间隔尽量大，其支持向量为：落在通道内 (包括边界) 被正确分类的样本点以及所有被错误分类的样本点 (通道内外都有可能分布)，也等价于，落在通道内 (包括边界) 的所有样本点以及通道外被错误分类的样本点。
    - 和软间隔 SVM 类似，SVR 也是希望经验损失尽量小的同时，间隔尽量大，但落在通道外 (包括边界) 的样本点才是支持向量，因此，通道外的样本点越少越好，也就是通道内的样本点越多越好。

5.2 SVR 的求解

和软间隔 SVM 一样，目前我们得到的 SVR 模型虽然是一个无约束的凸优化问题，但目标函数不可导。我们依旧引入松弛变量对原问题进行恒等变形，为了简便起见我们记 $f(\boldsymbol{x};\boldsymbol{w},b)$ 为 $f(\boldsymbol{x})$：

$$\begin{aligned}\min_{\boldsymbol{w},b,\boldsymbol{\xi},\hat{\boldsymbol{\xi}}} \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^N(\xi_i + \hat{\xi}_i)\\ \text{s.t. } f(\boldsymbol{x}_i) \leqslant y_i + \epsilon + \xi_i\\ f(\boldsymbol{x}_i) \geqslant y_i - \epsilon - \hat{\xi}_i\\ \xi_i \geqslant 0, \hat{\xi}_i \geqslant 0,\\ i = 1,2,\cdots,N\end{aligned}$$

和软间隔 SVM 类似，我们可以证明上述约束优化问题与 $\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}|_2^2 + C\sum_{i=1}^N l_\epsilon(f(\boldsymbol{x}_i), y_i)$ 等价。给定某个样本点 $(\boldsymbol{x}_i, y_i)$，模型预测值为 $f(\boldsymbol{x}_i)$，则若 $y_i - \epsilon \leqslant f(\boldsymbol{x}_i) \leqslant y_i + \epsilon$，则损失为 0，否则损失为 $|y_i - f(\boldsymbol{x}_i)| - \epsilon$。但不管 $f(\boldsymbol{x}_i)$ 是否落在 $[y_i - \epsilon, y_i + \epsilon]$ 中，我们总可以引入变量 $\xi \geqslant 0, \hat{\xi}_i \geqslant 0$，使得 $f(\boldsymbol{x}_i) \in [y_i - \epsilon - \hat{\xi}_i, y_i + \epsilon + \xi]$，而 $\xi, \hat{\xi}_i$ 也被称为松弛变量。我们可以看到，若对 $\xi, \hat{\xi}_i$ 不加限制，则损失函数 $l_\epsilon(f(\boldsymbol{x}_i), y_i) \leqslant \xi + \hat{\xi}_i$，但若限制 $\xi + \hat{\xi}_i$ 尽可能小即 $\min_{\xi,\hat{\xi}_i} \xi + \hat{\xi}_i$，则：

- 若 $f(\boldsymbol{x}_i) \in [y_i - \epsilon, y_i + \epsilon]$，则 $\xi = \hat{\xi}_i = 0$，该点损失为 $l_\epsilon(f(\boldsymbol{x}_i), y_i) = 0 = \xi + \hat{\xi}_i$；
- 若 $f(\boldsymbol{x}_i) < y_i - \epsilon$，则 $\xi = 0, \hat{\xi}_i = y_i - \epsilon - f(\boldsymbol{x}_i)$，该点损失为 $l_\epsilon(f(\boldsymbol{x}_i), y_i) = |y_i - f(\boldsymbol{x}_i)| - \epsilon = y_i - f(\boldsymbol{x}_i) - \epsilon = \xi + \hat{\xi}_i$；
- 若 $f(\boldsymbol{x}_i) > y_i + \epsilon$，则 $\xi = f(\boldsymbol{x}_i) - y_i - \epsilon, \hat{\xi}_i = 0$，该点损失为 $l_\epsilon(f(\boldsymbol{x}_i), y_i) = |y_i - f(\boldsymbol{x}_i)| - \epsilon = f(\boldsymbol{x}_i) - y_i - \epsilon = \xi + \hat{\xi}_i$；

因此，我们有 $\min_{\xi,\hat{\xi}_i} \xi + \hat{\xi}_i = l_\epsilon(f(\boldsymbol{x}_i), y_i)$，代入 $\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}|_2^2 + C\sum_{i=1}^N l_\epsilon(f(\boldsymbol{x}_i), y_i)$ 中，并考虑所有约束条件即证。

为求解上述约束优化问题，我们引入拉格朗日乘子 $\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}$，则拉格朗日函数为：

$$\begin{aligned}
&\mathcal{L}(\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}}, \boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}}) \\
&= \frac{1}{2}\|\boldsymbol{w}\|_2^2 + C\sum_{i=1}^N(\xi_i + \hat{\xi}_i) - \sum_{i=1}^N \beta_i\xi_i - \sum_{i=1}^N \hat{\beta}_i\hat{\xi}_i \\
&+ \sum_{i=1}^N \alpha_i(f(\boldsymbol{x}_i) - y_i - \epsilon - \xi_i) + \sum_{i=1}^N \hat{\alpha}_i\left(y_i - \epsilon - \hat{\xi}_i - f(\boldsymbol{x}_i)\right)
\end{aligned}$$

原变量为 $(\boldsymbol{w}, b, \boldsymbol{\xi}, \hat{\boldsymbol{\xi}})$，对偶变量为拉格朗日乘子 $(\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}, \hat{\boldsymbol{\beta}})$ 且非负。令拉格朗日函数对原变量的导数为 0，可得：

$$\begin{cases}
\boldsymbol{w} = \sum_{i=1}^N(\hat{\alpha}_i - \alpha_i)\boldsymbol{x}_i & \\
\sum_{i=1}^N(\hat{\alpha}_i - \alpha_i) = 0 & \\
\alpha_i + \beta_i = C & i = 1, \cdots, N \\
\hat{\alpha}_i + \hat{\beta}_i = C & i = 1, \cdots, N
\end{cases}$$

将上述关系代入拉格朗日函数中，消除原变量并化简，得**对偶问题**：

$$\begin{aligned}
\max_{\boldsymbol{\alpha}, \hat{\boldsymbol{\alpha}}} &-\frac{1}{2}\sum_{i=1}^N\sum_{j=1}^N(\hat{\alpha}_i - \alpha_i)(\hat{\alpha}_j - \alpha_j)\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle + \sum_{i=1}^N\left(y_i(\hat{\alpha}_i - \alpha_i) - \epsilon(\hat{\alpha}_i + \alpha_i)\right) \\
&\text{s.t.} \quad \sum_{i=1}^N(\hat{\alpha}_i - \alpha_i) = 0 \\
&\qquad 0 \leqslant \alpha_i, \hat{\alpha}_i \leqslant C, \quad \text{for all } i
\end{aligned}$$

上式中只含有拉格朗日乘子 $\alpha_i, \hat{\alpha}_i$，拉格朗日乘子 $\beta_i, \hat{\beta}_i$ 因为存在关系 $\alpha_i + \beta_i = C, \hat{\alpha}_i + \hat{\beta}_i = C$ 而被消去。

和 SVM 类似，求解对偶问题后，**由原问题的 KKT 条件**，我们可得原变量解与对偶变量解之间的关系：

$$\boldsymbol{w}^* = \sum_{i=1}^N(\hat{\alpha}_i^* - \alpha_i^*)\boldsymbol{x}_i,$$

$$b^* = y_j + \epsilon - \sum_{i=1}^N(\hat{\alpha}_i^* - \alpha_i^*)\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \quad 0 < \alpha_j^* < C$$

$$\text{or}$$

$$b^* = y_j - \epsilon - \sum_{i=1}^N(\hat{\alpha}_i^* - \alpha_i^*)\langle\boldsymbol{x}_i, \boldsymbol{x}_j\rangle, \quad 0 < \hat{\alpha}_j^* < C$$

其中 $\boldsymbol{\alpha}^*, \hat{\boldsymbol{\alpha}}^*$ 为对偶变量的解，$\alpha_i^*, \hat{\alpha}_i^*$ 至少有一个为 0，即 $\alpha_i^*\hat{\alpha}_i^* = 0$；$\boldsymbol{w}^*, b^*$ 为原变量的解。b^* **的解不唯一，$0 < \alpha_j^* < C$ 和 $0 < \hat{\alpha}_j^* < C$ 时的计算公式不同，为了得到更鲁棒的结果，通常会对所有 $0 < \alpha_j^* < C$ 或 $0 < \hat{\alpha}_j^* < C$ 的结果取平均。**

和软间隔 SVM 的情况类似，$0 < \alpha_j^* < C$ 或 $0 < \hat{\alpha}_j^* < C$ 的点是落在间隔边界上的支持向量，$0 < \alpha_j^* < C$ 的点落在间隔下边界上，$0 < \hat{\alpha}_j^* < C$ 的点落在间隔上边界上。此外，还有一部分支持向量落在间隔外，对应的拉格朗日乘子 $\alpha_j^* = C$ (低于间隔下边界) 或 $\hat{\alpha}_j^* = C$ (高于间隔上边界)。以 $\alpha_j^* = C$ 为例，由原问题的 KKT 条件，我们有关系式 $\boldsymbol{w}^{*T}\boldsymbol{x}_j + b^* - y_j - \epsilon - \xi_j^* = 0$，$\xi_j^* > 0$ 为待求的松弛变量，我们无法由此关系式解得 b^*，事实上，我们一般是先解出 b^* 再代入关系式中求得 ξ_j^*，ξ_j^* 随 b^* 变换而变化。

模型最终为：

$$f(\boldsymbol{x}) = \sum_{i=1}^{N}(\hat{\alpha}_i^* - \alpha_i^*)\langle \boldsymbol{x}_i, \boldsymbol{x}\rangle + b^*$$

和前面一样，基于对偶问题，我们可以引入核方法。**关于对偶和核方法，《白板推导》给出了很精炼的总结可作参考：非线性带来高维转换 (从模型角度，$\boldsymbol{x} \to \phi(\boldsymbol{x})$)；对偶表示带来内积 (从优化角度，$\boldsymbol{x}_i^T\boldsymbol{x}_j$，因此可以使用核方法)。**

更多资料：

[损失函数总结]

[支持向量机: Kernel]

《Learning with Kernels》

SVM发明者 Vapnik 老爷爷的书《Statistical Learning Theory》和《The Nature of Statistical Learning Theory》，中文译本叫做《统计学习理论》和《统计学习理论的本质》。粗略地讲，后一本可以看成是前一本的几乎删掉了所有证明细节的精简版。
