SVM 的数学推导和 Python 实现

赵新锋

2021 年 7 月 19 日

摘要

支持向量机（support vector machines, SVM）是一种分类模型，该模型在特征空间中求解间隔最大的分类超平面。当训练数据近似线性可分时，可以通过增加软间隔学习一个线性分类器。当线性不可分时，利用核技巧，隐式的将特征空间映射到高维特征空间，从而达到线性可分。使用序列最小最优化算法（SMO），可以快速求解模型的参数。

关键词：支持向量机; SVM; SMO; 矩阵运算; 矩阵求导;numpy;sklearn.

目录

1 数学推导与 python 实现

1.1 分类超平面

当训练数据线性可分，可以得到一个线性超平面 $x \cdot w + b = 0$, 将在超平面上方的归为正类，将在超平面下方的归为负类。当数据点与超平面距离越远时，表示分类的确定性越高，这样虽然线性分类的超平面可能有无数多个，但是我们可以找到一个所有点距离超平面最大的一个超平面。相应的决策函数为:

$$f(x) = sign(x \cdot w^* + b^*)$$

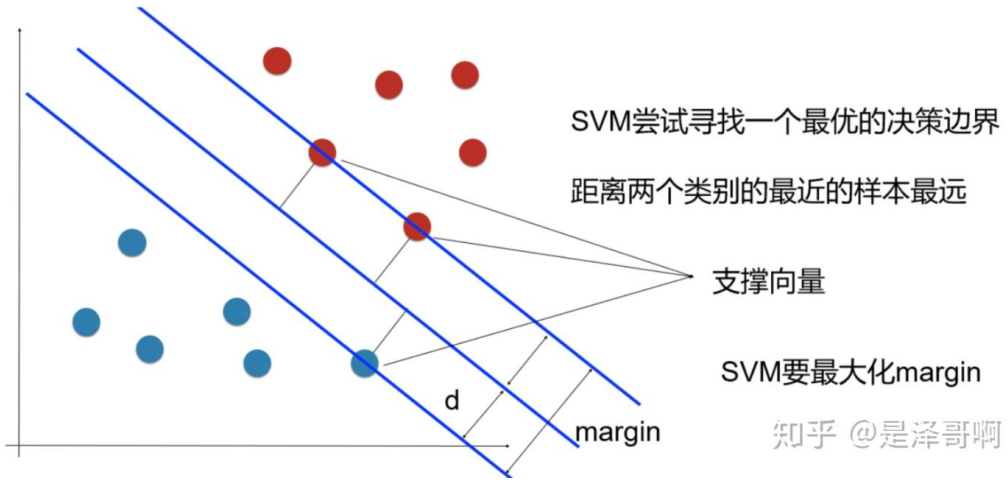

图 1: SVM 线性可分

1.2 公式推导

$x_i^T \cdot w_0 + b_0$ 为正时，y_i 为正，当 $x_i^T \cdot w_0 + b_0$ 为负时，y_i 为负，则可以定义 $\frac{y_i \cdot (x_i^T \cdot w_0 + b_0)}{\|w_0\|}$ 为几何间隔，表示数据点距离超平面的距离。模型最终会找到一个参数为 w_0 和 b_0 的分离超平面，所有点距离超平面的距离都大于等于 d，将距离正好等于 d 的数据点称之为支持向量。

$$\begin{aligned} \arg\max_{w_0, b_0} \ & d = \frac{y_0 \cdot (x_0^T \cdot w_0 + b_0)}{\|w_0\|} \\ \text{s.t.} \ & \frac{y_i \cdot (x_i^T \cdot w_0 + b_0)}{\|w_0\|} \geq d \end{aligned} \tag{1}$$

将 w_0 和 b_0 进行一定比例的缩放

$$w = \frac{w_0}{y_0 \cdot (x_0^T \cdot w_0 + b_0)}$$
$$b = \frac{b_0}{y_0 \cdot (x_0^T \cdot w_0 + b_0)}$$

可以将 (1) 式化简为:

$$\begin{aligned}
&\arg\max_{w}\ d = \frac{1}{\|w\|} \\
\Leftrightarrow\ &\arg\min_{w}\ d = \|w\| \\
\Leftrightarrow\ &\arg\min_{w}\ d = \frac{1}{2}w^T \cdot w \qquad (2) \\
&\text{s.t.}\ \ y_i \cdot (x_i^T \cdot w + b) \geq 1 \qquad (3)
\end{aligned}$$

将(2) 和 (3) 利用拉格朗日乘数法，获得拉格朗日原始问题形式:

$$\begin{aligned}
\arg\min_{w,b}\max_{\alpha}\ L(w,b,\alpha) &= \frac{1}{2}w^T \cdot w - \alpha^T \cdot (y \odot (X \cdot w + b) - 1) \\
&= \frac{1}{2}w^T \cdot w - (\alpha \odot y)^T \cdot (X \cdot w + b) + \alpha^T \cdot \mathbf{1}^m \\
&= \frac{1}{2}w^T \cdot w - (\alpha \odot y)^T \cdot (X \cdot w + b) + \mathbf{1}^T \cdot \alpha \qquad (4)
\end{aligned}$$

当满足 KTT 条件时，拉格朗日对偶问题的解等价于(4) 的解:

$$\arg\max_{\alpha}\min_{w,b}\ L(w,b,\alpha) = \frac{1}{2}w^T \cdot w - (\alpha \odot y)^T \cdot (X \cdot w + b) + \mathbf{1}^T \cdot \alpha \qquad (5)$$

首先求 L 以 w、b 为参数的极小值，通过求微分得到偏导数形式。

$$\begin{aligned}
\mathrm{d}L = &\frac{1}{2}tr[(\mathrm{d}w)^T \cdot w + w^T \cdot \mathrm{d}w) - (\alpha \odot y)^T \cdot (X\mathrm{d}w) \\
&- (1^T \cdot (\alpha \odot y))^T \cdot \mathrm{d}b \\
&- (\mathrm{d}\alpha)^T \cdot (y \odot (X \cdot w + b) - 1)] \\
= &\, tr[w^T\mathrm{d}w - (X^T \cdot (\alpha \odot y))^T\mathrm{d}w - (\alpha \odot y)^T \cdot \mathrm{d}b - (y \odot (X \cdot w + b) - 1)^T\mathrm{d}\alpha]
\end{aligned}$$

从而得到 w、b 偏导数，并令偏导数为 0。

$$\begin{aligned}
\frac{\partial L}{\partial w} &= w - X^T \cdot (\alpha \odot y) = \mathbf{0} \\
\frac{\partial L}{\partial b} &= -1^T \cdot (\alpha \odot y) = -y^T \cdot \alpha = 0
\end{aligned}$$

推导出如下关系:

$$w = X^T \cdot (\alpha \odot y) \qquad (6)$$

$$y^T \cdot \alpha = 0 \qquad (7)$$

将(6)式和 (7)式代入(5)式中，

$$\begin{aligned}
\arg\max_{\alpha} L(w,b,\alpha) = &\frac{1}{2}(\alpha \odot y)^T \cdot X \cdot X^T \cdot (\alpha \odot y) \\
&- (\alpha \odot y)^T \cdot X \cdot X^T \cdot (\alpha \odot y) \\
&- (\alpha \odot y)^T \cdot b^m \\
&+ \mathbf{1}^T \cdot \alpha \\
= &-\frac{1}{2}(\alpha \odot y)^T \cdot X \cdot X^T \cdot (\alpha \odot y) + (\alpha \odot y)^T \cdot b + \mathbf{1}^T \cdot \alpha \\
= &-\frac{1}{2}(\alpha \odot y)^T \cdot X \cdot X^T \cdot (\alpha \odot y) + \mathbf{1}^T \cdot \alpha
\end{aligned}$$

去除负号，将极大转换成极小形式：

$$\begin{aligned} \arg\min_{\alpha} L(w,b,\alpha) &= \frac{1}{2}(\alpha \odot y)^T \cdot X \cdot X^T \cdot (\alpha \odot y) - \mathbf{1}^T \cdot \alpha \\ \text{s.t.} \quad & y^T \cdot \alpha = 0 \end{aligned} \tag{8}$$

为了使拉格朗日对偶问题的解与原始问题解相同，需要同时满足 KTT 条件：

$$\begin{aligned} \frac{\partial L}{\partial w} &= 0 \\ \frac{\partial L}{\partial b} &= 0 \\ \alpha \odot (y \odot (X \cdot w + b) - 1) &= \mathbf{0}^m \\ y \odot (X \cdot w + b) - 1 &\geq \mathbf{0}^m \\ \alpha &\geq \mathbf{0}^m \end{aligned}$$

1.3 软间隔

当训练数据近似线性可分，有些异常点或噪声点导致无法找到分离超平面，可以对每个数据点加一个松弛变量 ξ_i，从而让所有数据点均满足约束。

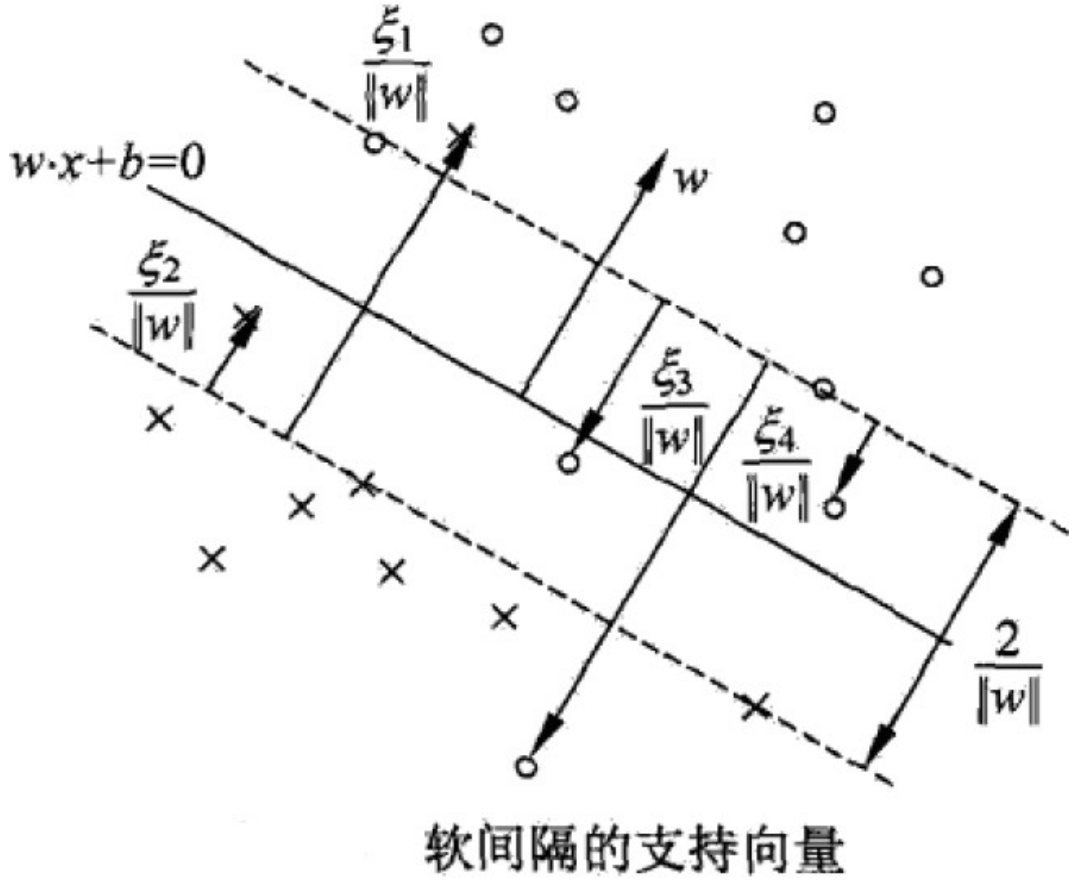

图 2: SVM 软间隔

加入软间隔参数, 每个向量点距离分类超平面距离增加 ξ_i，同时增加一个惩罚系数 C, 代入(5)新形式如下：

$$\begin{aligned}\arg\min_{w}\ & d=\frac{1}{2}w^T\cdot w+C\cdot\mathbf{1}^m\cdot\xi\\ \text{s.t.}\ & y_i\cdot(x_i^T\cdot w+b)\geq 1-\xi_i\\ & \xi_i\geq 0\end{aligned}$$

将其转换为拉格朗日对偶形式：

$$L(w,b,\xi,\alpha,\mu)=\frac{1}{2}w^T\cdot w+C\cdot\mathbf{1}^T\cdot\xi-\alpha^T\cdot(y\odot(X\cdot w+b)-1+\xi)-\mu^T\cdot\xi \tag{9}$$

求得对于 w、b、ξ 的偏导并令其为 0：

$$\begin{aligned}\frac{\partial L}{\partial w}&=w-X^T\cdot(\alpha\odot y)=\mathbf{0}\\ \frac{\partial L}{\partial b}&=-1^T\cdot(\alpha\odot y)=-y^T\cdot\alpha=0\\ \frac{\partial L}{\partial \xi}&=C-\alpha-u=\mathbf{0}\end{aligned}$$

代入 (9) 式中：

$$\begin{aligned}\arg\max_{\alpha} L(\alpha)=&-\frac{1}{2}(\alpha\odot y)^T\cdot X\cdot X^T\cdot(\alpha\odot y)+\mathbf{1}^m\cdot\alpha\\ \text{s.t.}\ & y^T\cdot\alpha=0\\ & C-\alpha-\mu=\mathbf{0}\\ & \alpha_i\geq 0\\ & \mu_i\geq 0\end{aligned}$$

$C-\alpha-u=\mathbf{0}$、$\alpha_i\geq 0$ 、$u_i\geq 0$ 约束可以化简为：

$$\begin{aligned}\arg\min_{\alpha} L(\alpha)=&\frac{1}{2}(\alpha\odot y)^T\cdot X\cdot X^T\cdot(\alpha\odot y)-\mathbf{1}^m\cdot\alpha\\ \text{s.t.}\ & y^T\cdot\alpha=0\\ & 0\leq\alpha_i\leq C\Leftrightarrow\mathbf{0}^m\leq\alpha\leq\boldsymbol{C}^m\end{aligned} \tag{10}$$

为了使拉格朗日对偶问题的解与原始问题解相同，需要同时满足 KTT 条件：

$$\begin{aligned}
\frac{\partial L}{\partial w} &= 0 \\
\frac{\partial L}{\partial b} &= 0 \\
\frac{\partial L}{\partial \xi} &= 0 \\
\alpha \odot (y \odot (X \cdot w + b) - 1 + \xi) &= \mathbf{0}^m \\
y \odot (X \cdot w + b) - 1 + \xi &\geq \mathbf{0}^m \\
\alpha &\geq \mathbf{0}^m \\
\mu \odot \xi &= \mathbf{0}^m \\
\xi &\geq \mathbf{0}^m \\
\mu &\geq \mathbf{0}^m
\end{aligned}$$

1.4 核函数

近似线性可分用软间隔方式解决，然而当训练数据是非线性数据，会出现无法在原特征空间找到分离超平面。可以使用一个非线性变换，将数据从原特征空间映射到更高维的新空间，然后在新空间中寻找线性分类超平面，这种方法被称为核技巧。观察 (10) 式中计算 $X \cdot X^t$，即需要计算 $x_i \cdot x_i$ 内积。将核技巧应用到 SVM，定义核函数为：

$$K(x, z) = \phi(x) \cdot \phi(z)$$

即将原来的向量内积，改成先让向量映射到新空间，然后再求内积。核技巧的另外一个优点是，不需要显示的定义 ϕ 而是直接计算出 $\phi(x) \cdot \phi(z)$ 的结果，以高斯核函数为例：

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

则 (10) 式利用核技巧转化为：

$$\begin{aligned}
\arg\min_{\alpha} L(\alpha) = \frac{1}{2}(\alpha \odot y)^T \cdot K(X, X) \cdot (\alpha \odot y) - \mathbf{1}^m \cdot \alpha \qquad (11) \\
\text{s.t.} \quad y^T \cdot \alpha = 0 \\
\mathbf{0}^m \leq \alpha \leq \boldsymbol{C}^m \\
\alpha \odot (y \odot (X \cdot w + b) - 1 + \xi) = \mathbf{0}^m \\
y \odot (X \cdot w + b) - 1 + \xi \geq \mathbf{0}^m \\
\alpha \geq \mathbf{0}^m \\
\mu \odot \xi = \mathbf{0}^m \\
\xi \geq \mathbf{0}^m \\
\mu \geq \mathbf{0}^m
\end{aligned}$$

1.5 序列最小最优化算法 SMO

支持向量机的拉格朗日对偶问题是一个凸二次规划问题，具有全局最优解，序列最小最优化算法即 SMO 算法，是高效求解支持向量机解的一种算法。其基本思路是，如果所有变量都满足了 KTT 条件，那么就求得了问题的解。SMO 算法过程如下：

- 选择两个变量，如 α_1, α_2，固定其他变量，那么原问题就变成了两个变量的二次优化问题。
- 由于有约束 $y^T \cdot \alpha = 0$ 的存在，选择了两个变量，实际上自由变量只有一个。
- 求解两个变量的最优解，迭代直到所有变量满足 KTT 条件。
- 迭代求解使用的解析方法，效率高

当选择两个变量，如 α_1, α_2 时，将其他变量看做常数：

$$\begin{aligned}\arg\min_{\alpha} L(\alpha) &= \frac{1}{2}(\alpha \odot y)^T \cdot K(X,X) \cdot (\alpha \odot y) - \mathbf{1}^m \cdot \alpha \\ &= \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{m}\alpha_i\end{aligned}$$

则上式子去除不包含 α_1, α_2 的项后化简为:

$$\begin{aligned}\arg\min_{\alpha_1,\alpha_2} W(\alpha_1,\alpha_2) &= \frac{1}{2}K_{11}a_1^2 + y_1y_2K_{12}a_1a_2 + \frac{1}{2}K_{22}a_2^2 \\ &\quad + y_1a_1\sum_{i=3}^{m}y_ia_iK_{i1} \\ &\quad + y_2a_2\sum_{i=3}^{m}y_ia_iK_{i2} \\ &\quad - (a_1+a_2) \\ &= \frac{1}{2}K_{11}a_1^2 + y_1y_2K_{12}a_1a_2 + \frac{1}{2}K_{22}a_2^2 \\ &\quad + y_1a_1(\sum_{i=1}^{m}y_ia_i^{old}K_{i1} - y_1a_1^{old}K_{11} - y_2a_2^{old}K_{12}) \\ &\quad + y_2a_2(\sum_{i=1}^{m}y_ia_i^{old}K_{i2} - y_1a_1^{old}K_{12} - y_2a_2^{old}K_{22}) \\ &\quad - (a_1+a_2) \\ &= \frac{1}{2}K_{11}a_1^2 + y_1y_2K_{12}a_1a_2 + \frac{1}{2}K_{22}a_2^2 \\ &\quad + y_1a_1((y \odot a^{old})^T \cdot K(X,x_1) - y_1a_1^{old}K_{11} - y_2a_2^{old}K_{12}) \\ &\quad + y_2a_2((y \odot a^{old})^T \cdot K(X,x_2) - y_1a_1^{old}K_{12} - y_2a_2^{old}K_{22}) \\ &\quad - (a_1+a_2)\end{aligned}$$

利用约束 s.t. $y^T \cdot \alpha = 0$ 得到:

$$a_1 y_1 + a_2 y_2 = a_1^{old} y_1 + a_2^{old} y_2 = \zeta$$
$$a_1 = y_1(\zeta - a_2 y_2) = a_1^{old} + a_2^{old} y_1 y_2 - a_2 y_1 y_2$$

将 a_1 代入:

$$\begin{aligned}\arg\min_{\alpha_2} W(\alpha_2) = &\frac{1}{2}K_{11}(y_1(\zeta - a_2 y_2))^2 + y_1 y_2 K_{12} y_1(\zeta - a_2 y_2)a_2 + \frac{1}{2}K_{22}a_2^2 \\ &+ y_1 y_1(\zeta - a_2 y_2)((y \odot a^{old})^T \cdot K(X, x_1) - y_1 a_1^{old} K_{11} - y_2 a_2^{old} K_{12}) \\ &+ y_2 a_2((y \odot a^{old})^T \cdot K(X, x_2) - y_1 a_1^{old} K_{12} - y_2 a_2^{old} K_{22}) \\ &- (y_1(\zeta - a_2 y_2) + a_2)\end{aligned}$$

针对 a_2 求导，并令导数为 0:

$$\begin{aligned}\frac{\partial W}{\partial a_2} = &K_{11} y_2(a_2 y_2 - \zeta) + K_{12}(y_2\zeta - 2a_2) + K_{22}a_2 \\ &- y_2((y \odot a^{old})^T \cdot K(X, x_1) - y_1 a_1^{old} K_{11} - y_2 a_2^{old} K_{12}) \\ &+ y_2((y \odot a^{old})^T \cdot K(X, x_2) - y_1 a_1^{old} K_{12} - y_2 a_2^{old} K_{22}) \\ &+ y_1 y_2 - 1 \\ = &a_2(K_{11} - 2K_{12} + K_{22}) \\ &- K_{11} y_2(y_1 a_1^{old} + y_2 a_2^{old}) + K_{12} y_2(y_1 a_1^{old} + y_2 a_2^{old}) \\ &+ y_2((y \odot a^{old})^T \cdot K(X, x_2) - (y \odot a^{old})^T \cdot K(X, x_1)) \\ &+ y_1 y_2 a_1^{old} K_{11} + a_2^{old} K_{12} - y_1 y_2 a_1^{old} K_{12} - a_2^{old} K_{22} \\ &+ y_1 y_2 - 1 \\ = &a_2(K_{11} - 2K_{12} + K_{22}) \\ &+ y_2((y \odot a^{old})^T \cdot K(X, x_2) - (y \odot a^{old})^T \cdot K(X, x_1)) \\ &- a_2^{old}(K_{11} - 2K_{12} + K_{22}) \\ &+ y_1 y_2 - 1 \\ = &a_2(K_{11} - 2K_{12} + K_{22}) \\ &+ y_2((y \odot a^{old})^T \cdot K(X, x_2) - y2 - ((y \odot a^{old})^T \cdot K(X, x_1) - y_1)) \\ &- a_2^{old}(K_{11} - 2K_{12} + K_{22}) \\ = &0\end{aligned}$$

则可以推得:

$$a_2^{new,unclip} = a_2^{old} + \frac{y2(((\alpha \odot y)^T \cdot K(X, x_1) - y_1) - ((\alpha \odot y)^T \cdot K(X, x_2) - y_2))}{(K_{11} + K_{22} - 2K_{12})} \tag{12}$$

考虑到 $w = X^T \cdot (a \odot y)$ 令 $g(x) = w^T \cdot x + b = (a \odot y)^T \cdot X \cdot x + b$，将将内积转换成核函数形式，并且定义函数 E_i 为函数 $g(x)$ 与 y_i 的误差值:

$$E_i = g(x_i) - y_i = (a \odot y)^T \cdot K(X, x_i) + b$$
$$\eta = K_{11} - 2K_{12} + K_{22}$$

则 (12) 式可以转换为:

$$a_2^{new,unclip} = a_2^{old} + \frac{y2(E_1 - E_2)}{\eta} \tag{13}$$

End!!