

# 机器学习-svm全手写推导

2020年5月19日 星期二

下午6:34

该文档包含一下4部分内容：

- 1.硬间隔SVM手写推导
- 2.软间隔SVM手写推导
- 3.核技巧和非线性SVM构造流程
- 4.svm面试问答

-浪矢73

## • 硬间隔SVM手写推导

MEMO NO. \_\_\_\_\_  
DATE \_\_\_\_\_

SVM (硬间隔)

① margin 计算

a. 点到超平面的距离  
(S:  $Wx+b=0$ )

$d = \frac{1}{\|W\|} |W \cdot x_0 + b|$

表示点  $x_0$  到超平面  $S: Wx+b=0$  的距离。

\* 证明:

设点  $x_0$  在平面  $S$  上的投影为  $x_1$ , 则  $Wx_1+b=0$

则  $\overline{x_0 x_1} \parallel W$  (平面的法向量)

故  $\|W \cdot \overline{x_0 x_1}\| = \|W\| \|\overline{x_0 x_1}\| = \|W\| d$

又  $W \cdot \overline{x_0 x_1} = W^1(x_1^1 - x_0^1) + W^2(x_1^2 - x_0^2) + \dots + W^n(x_1^n - x_0^n)$

$= W^1 x_1^1 + W^2 x_1^2 + \dots + W^n x_1^n - (W^1 x_0^1 + W^2 x_0^2 + \dots + W^n x_0^n)$

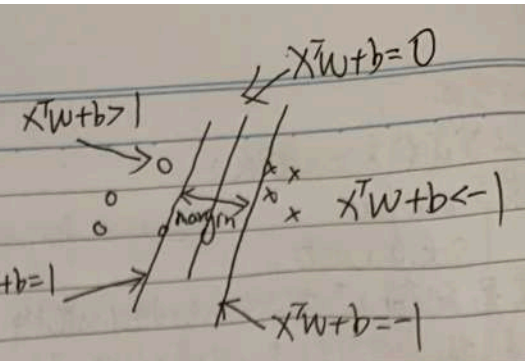
$= W^1 x_1^1 + W^2 x_1^2 + \dots + W^n x_1^n \xrightarrow{Wx_1+b=0} -b - Wx_0$

$\therefore \|W\| d = \|W \cdot \overline{x_0 x_1}\| = |-Wx_0 - b|$

$\Rightarrow d = \frac{|Wx_0 + b|}{\|W\|}$

b. 求 margin

设点  $a_1$  在  $x^T W + b = 1$  上的一点,  $a_2$  到  $x^T W + b = -1$  的距离为:



$$d = \frac{1 \times w + b}{\|w\|} = \frac{1}{\|w\|}$$

$$\text{by margin} = \frac{2}{\|w\|}$$

$$\text{求最大 margin} \Rightarrow \max (\text{margin})^2 \Rightarrow \max \frac{2}{\|w\|^2} \Rightarrow \min \frac{1}{\|w\|^2}$$

$$\text{故最大 margin 等价于 } \min_{w,b} \frac{1}{\|w\|^2}$$

$$\text{s.t. } y_i (x_i^T w + b) \geq 1 \quad (y_i = \pm 1)$$

② KKT条件:

定义: 不等式约束优化问题:

$$\begin{cases} \min f(x) \\ \text{s.t. } g(x) \leq 0, \end{cases} \quad (L(x) = f(x) + \lambda g(x))$$

设其最优解  $x^*$  满足如下性质 (KKT条件):

互补松弛性, 原始可行性, 对偶可行性, 互补松弛性:

$$\begin{cases} \nabla L = \nabla f + \lambda \nabla g = 0 \\ g(x) \leq 0 \\ \lambda \geq 0 \\ \lambda g(x) = 0 \end{cases} \quad (\text{不等式约束处理类似})$$

推广到 SVM:

$$\begin{cases} \min_{w,b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } -[y_i (x_i^T w + b) - 1] \leq 0 \Leftrightarrow y_i (x_i^T w + b) \geq 1 \end{cases}$$

\* SVM 是一个凸优化问题, 解一定全局最优:

KKT条件:

$$\begin{cases} \nabla f + \lambda \nabla g = 0 \\ \alpha_i \geq 0 \\ -[y_i (x_i^T w + b) - 1] \leq 0 \\ -\alpha_i [y_i (x_i^T w + b) - 1] = 0 \end{cases} \quad (L(x) = \frac{1}{2} \|w\|^2 - \sum \alpha_i [y_i (x_i^T w + b) - 1])$$

③ 拉格朗日函数及问题转化:

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum \alpha_i [y_i (x_i^T w + b) - 1]$$

$\therefore$  不满足  $y_i (x_i^T w + b) \geq 1$ , 则  $\max_{\alpha} L(w, b, \alpha) = +\infty$  (或者  $\geq \frac{1}{2} \|w\|^2$ )



若满足  $y_i(x_i^T w + b) \geq 1$ , 则  $\max_{\alpha} L(w, b, \alpha) = \sum \|w\|$

故  $\begin{cases} \min_{w, b} \frac{1}{2} \|w\|^2 \\ \text{s.t. } -1 \leq y_i(x_i^T w + b) - 1 \leq 0 \end{cases} \Leftrightarrow \min_{w, b} \max_{\alpha} L(w, b, \alpha)$

MEMO NO. \_\_\_\_\_  
DATE \_\_\_\_\_

④ 对偶问题:

根据拉格朗日对偶性:

$$\min_{w, b} \max_{\alpha} L(w, b, \alpha) \Leftrightarrow \max_{\alpha} \min_{w, b} L(w, b, \alpha)$$

⑤ 求解:

1) 求  $\min_{w, b} L(w, b, \alpha)$

$$\nabla_w L(w, b, \alpha) = w - \sum_{i=1}^n \alpha_i y_i x_i = 0$$

故:  $w = \sum_{i=1}^n \alpha_i y_i x_i$

$$\nabla_b L(w, b, \alpha) = -\sum_{i=1}^n \alpha_i y_i = 0$$

将  $w = \sum_{i=1}^n \alpha_i y_i x_i$  代入  $L(w, b, \alpha)$ :

$$\begin{cases} \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

得:  $\min_{w, b} L(w, b, \alpha) = \max_{\alpha} \frac{1}{2} \left\| \sum_{i=1}^n \alpha_i y_i x_i \right\|^2 - \sum_{i=1}^n \left[ \alpha_i y_i x_i^T \cdot \sum_{j=1}^n \alpha_j y_j x_j \right] + \sum_{i=1}^n \alpha_i$

$\|x\|^2 = x^T x$

$$\Rightarrow \frac{1}{2} \sum_{i=1}^n \alpha_i y_i x_i^T \cdot \sum_{j=1}^n \alpha_j y_j x_j - \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i$$

$$\Rightarrow \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

则:

$$\max_{\alpha} L(w, b, \alpha) \Rightarrow \max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j$$

$$\Downarrow$$

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^n \alpha_i$$

$$s.t. \sum_{i=1}^n \alpha_i y_i = 0, \alpha_i \geq 0.$$

smo/二次规划解出最优解  $\hat{\alpha}_i (i=1, 2, \dots, n)$ .

① 硬间隔 SVM:

若  $\alpha_i = 0$ , 约束不起作用, 对应样本点的支持向量.

若  $\alpha_i > 0$ , 则  $y_i (x_i^T w + b) = 1$  (KKT条件), 则该样本点位于最大间隔的边界上, 为支持向量.

由  $\hat{\alpha}$  可得:

$$\hat{w} = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\hat{b} = y_i - \sum_{j=1}^n \alpha_j y_j x_i^T x_j \quad (\text{下标不同}).$$

② 判别函数:

$$f(x) = \text{sign}(\sum_{i \in SV} \hat{\alpha}_i y_i x^T x_i + \hat{b})$$

\* SVM 的解只与支持向量 SV 有关, 与非支持向量无关.

流程:

margin 计算

↓

KKT 条件

↓

拉格朗日函数及问题转化.

↓

求解  $\Rightarrow$  判别函数.



## 2. 间隔问题转化

### • 软间隔SVM手写推导

软间隔SVM:

① 优化目标:

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

正则项      hinge loss.

\*  $C \rightarrow +\infty$  变成硬间隔

$C$  越大越容易过拟合, 越大越容易欠拟合.

引入松弛变量  $\xi_i \geq 0$ :

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

$$s.t. \ y_i(x_i^T w + b) \geq 1 - \xi_i \Rightarrow \xi_i \geq [1 - y_i(x_i^T w + b)]$$

$\xi_i \geq 0, i=1, 2, \dots, n$ .

$$(\text{软间隔约束}) \quad 0 \geq 1 - \xi_i - y_i(x_i^T w + b)$$

$0 \geq -\xi_i$ .

② 拉格朗日函数(约束条件):

$$L(w) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i [ \xi_i - 1 + y_i(x_i^T w + b) ] - \sum_{i=1}^n \beta_i \xi_i$$

若满足:  $\xi_i \geq [1 - y_i(x_i^T w + b)], \xi_i \geq 0$ .

$$\text{by } \max_{\alpha_i, \beta_i} L(w) \Rightarrow \frac{1}{2} \|w\|^2 + C \sum_{i=1}^n \xi_i$$

若不满足:  $\xi_i \geq [1 - y_i(x_i^T w + b)], \xi_i \geq 0$ .

$$\text{by } \max_{\alpha_i, \beta_i} L(w) \Rightarrow +\infty$$

故原问题转化为:  $\min_{w, b, \xi_i} \max_{\alpha_i, \beta_i} L(w)$ .

③ 转为对偶问题:

$$\min_{w, b, \xi_i; \alpha_i, \beta_i} \max L(\alpha) \Leftrightarrow \max_{\alpha_i, \beta_i} \min_{w, b, \xi_i} L(\alpha).$$

④ 计算:

对  $\min_{w, b, \xi_i} L(\alpha)$ :

$$\nabla_w L(\alpha) = w - \sum_{i=1}^n \alpha_i y_i x_i \Rightarrow w = \sum_{i=1}^n \alpha_i y_i x_i$$

$$\nabla_b L(\alpha) = \sum_{i=1}^n \alpha_i y_i = 0$$

$$\nabla_{\xi_i} L(\alpha) = \sum_{i=1}^n C - \sum_{i=1}^n \alpha_i - \sum_{i=1}^n \beta_i = 0 \Rightarrow C = \alpha_i + \beta_i$$

代入式 (1)  $L(\alpha)$ :

$$\|X\|^2 = X^T X$$

$$L(\alpha) = \frac{1}{2} \sum_{j=1}^n \alpha_j y_j x_j^T \sum_{i=1}^n \alpha_i y_i x_i + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \xi_i - \frac{1}{2} y_i x_i^T \sum_{j=1}^n \alpha_j y_j x_j + y_i b - \sum_{i=1}^n \beta_i \xi_i$$

$$\Rightarrow \frac{1}{2} \sum_{j=1}^n \alpha_j y_j x_j^T \sum_{i=1}^n \alpha_i y_i x_i + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \xi_i + \sum_{i=1}^n \alpha_i$$

$$- \frac{1}{2} \sum_{i=1}^n \alpha_i y_i x_i^T \sum_{j=1}^n \alpha_j y_j x_j - \sum_{i=1}^n \alpha_i y_i b - \sum_{i=1}^n \beta_i \xi_i$$

$$\Rightarrow -\frac{1}{2} \sum_{i=1}^n \alpha_i y_i x_i^T \sum_{j=1}^n \alpha_j y_j x_j$$

$$\Rightarrow -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i + C \sum_{i=1}^n \xi_i - \sum_{i=1}^n \alpha_i \xi_i - \sum_{i=1}^n \beta_i \xi_i$$

$$\Rightarrow -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i$$

$$\text{to } \max_{\alpha_i, \beta_i} \min_{w, b, \xi_i} L(\alpha) \Leftrightarrow \max_{\alpha_i, \beta_i} \left( -\frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j x_i^T x_j + \sum_{i=1}^n \alpha_i \right)$$

代入  $\alpha_i$  和  $\beta_i$



⑤ 大K-条件:

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

$$\text{s.t. } 1 - \xi_i - y_i (x_i^T w + b) \leq 0, \\ -\xi_i \leq 0$$

KKT条件:

$$\begin{cases} \alpha_i \geq 0, \beta_i \geq 0 \\ 1 - \xi_i - y_i (x_i^T w + b) \leq 0, -\xi_i \leq 0 \end{cases}$$

$$\alpha_i [1 - \xi_i - y_i (x_i^T w + b)] = 0 \quad -\beta_i \xi_i = 0$$

⑥ 结论:

当  $\alpha_i = 0$  时, 该点的相关条件没有用, 不是支持向量.

~~在~~  $\alpha_i =$

当  $\alpha_i > 0$  时, 此时称一个支持向量.

当  $0 < \alpha_i < C$  时,  $0 < \beta_i < C$ , 此时为支持向量.

$$\begin{cases} [1 - \xi_i - y_i (x_i^T w + b)] \cdot \beta_i = 0 \\ -\beta_i \xi_i = 0 \end{cases} \Rightarrow \begin{cases} \xi_i = 0 \\ 1 = y_i (x_i^T w + b) \end{cases}$$

此时为最大间隔上的支持向量.

若  $\alpha_i = C, \beta_i = 0, 0 < \xi_i \leq 1$

$y_i (x_i^T w + b) = 1 - \xi_i$ , 则该点为间隔内部的支持向量.

若  $\alpha_i = C, \xi_i > 1$

$$y_i (x_i^T w + b) = 1 - \xi_i < 0,$$

该点为错误分类的支持向量.

⑦ 判别函数:

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

- 核技巧和非线性SVM构造流程

MEMO NO. \_\_\_\_\_  
DATE \_\_\_\_\_

非线性SVM - 核技巧

① 映射  $\phi(x): x \rightarrow H$ .

② 核函数  $\Rightarrow k(x, z) = \phi(x) \cdot \phi(z)$

③ 计算  $k(x, z)$  比较容易, 而  $\phi(x) \cdot \phi(z)$  计算  $k(x, z)$  不容易

非线性SVM

① 选择适当的核函数  $k(x, z)$  和参数  $C$ .

核函数最优化问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j k(x_i, x_j) - \sum_{i=1}^n \alpha_i$$

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

$$0 \leq \alpha_i \leq C, i=1, 2, \dots, n$$

② 二次规划问题求解  $\hat{\alpha}$

③ 选择满足  $0 < \hat{\alpha}_j < C$  的  $\hat{\alpha}_j$ .

$$\text{计算: } \hat{b} = y_j - \sum_{i \in SV} \hat{\alpha}_i y_i k(x_j, x_i)$$

④ 构造决策函数:

$$f(x) = \text{sign} \left( \sum_{i \in SV} \hat{\alpha}_i y_i k(x, x_i) + \hat{b} \right)$$





## • svm面试问答

1. 回顾：请参考 svm 系列视频以及课本 3.1 节在白纸上完成硬间隔 svm 公式的推导。其中推导中可略过 SMO 算法。

参考答案：参考讲解视频以及讲义进行推导。

2. 补充：为什么要将求解 SVM 的原始问题转换为其对偶问题？

参考答案：首先明确一点，不转化为对偶问题也能求解。所以该问题可以从转化为对偶问题后带来的优点进行阐述。

引入对偶问题所带来的优势是：

- (1) 对偶问题有时候更易求解， $w$ 、 $b$  是维度相关的，而  $\lambda$  是维度无关的，与样本数量相关，所以对高维数据且样本数量一定，适用于对偶问题。
- (2) 对偶问题产生内积，方便核函数的引入，进而推广到非线性分类问题。

3. 补充：为什么 SVM 要引入核函数？

参考答案：首先明确一点，核函数并非 SVM 独有，它是一种解决问题的方法。

当样本在原始空间线性不可分时，可将样本从原始空间映射到一个更高维的特征空间，使得样本在这个特征空间内线性可分。而引入这样的映射后，所要求解的对偶问题的求解中，无需求解真正的映射函数，而只需要知道其核函数。核函数的定义： $K(x,y)=\langle \phi(x),\phi(y) \rangle$ ，即在特征空间的内积等于它们在原始样本空间中通过核函数  $K$  计算的结果。一方面数据变成了高维空间中线性可分的数据，另一方面不需要求解具体的映射函数，只需要给定具体的核函数即可，这样使得求解的难度大大降低。

4. 补充：为什么 SVM 对缺失数据敏感？哪些模型对缺失数据不那么敏感？（这里说的缺失数据是指缺失某些特征数据，向量数据不完整）

参考答案：

SVM 没有处理缺失值的策略。而 SVM 希望样本在特征空间中线性可分，所以特征空间的好坏对 SVM 的性能很重要。缺失特征数据将影响训练结果的好坏。

对于缺失数据敏感和不敏感的面试题可以参考下面这么答：

树模型一般对于缺失值的敏感度较低，大部分时候可以在数据有缺失时使用。

一般涉及到距离度量时，如计算两个点之间的距离，缺失数据就变得比较重要。因为涉及到“距离”这个概念，那么缺失值处理不当就会导致效果很差，如 K 近邻算法(KNN)和支持向量机(SVM)。

线性模型的损失函数往往涉及到距离的计算，计算预测值和真实值之间的差别，这容易导致对缺失值敏感。

神经网络的鲁棒性强，对于缺失数据不是非常敏感，但一般没有那么多数据可供使用。

贝叶斯模型对于缺失数据也比较稳定，数据量很小的时候首推贝叶斯模型。

总结来看，对于有缺失值的数据在经过缺失值处理后：

数据量很小，用朴素贝叶斯

数据量适中或者较大，用树模型，优先 xgboost

数据量较大，也可以用神经网络

避免使用距离度量相关的模型，如 KNN 和 SVM

5. 补充：谈谈你是怎么使用 SVM 中的核函数的。

参考答案：

一般选择线性核和高斯核，也就是线性核与 RBF 核。线性核：主要用于线性可分的情形，参数少，速度快，对于一般数据，分类效果已经很理想了。RBF 核：主要用于线性不可分的情形，参数多，分类结果非常依赖于参数。有很多人是通过训练数据的交叉验证来寻找合适的参数，不过这个过程比较耗时。如果 Feature 的数量很大，跟样本数量差不多，这时候选用线性核的 SVM。如果 Feature 的数量比较小，样本数量一般，不算大也不算小，选用高斯核的 SVM。

6. 补充: SVM 的优缺点：





**参考答案：**

**优点：**

- (1) 由于 SVM 是一个凸优化问题，所以求得的解一定是全局最优而不是局部最优。
- (2) 不仅适用于线性问题还适用于非线性问题(用核技巧)。
- (3) 拥有高维样本空间的数据也能用 SVM，这是因为数据集的复杂度只取决于支持向量而不是数据集的维度，这在某种意义上避免了“维数灾难”。
- (4) 理论基础比较完善(例如神经网络就更像一个黑盒子)。

**缺点：**

- (5) 二次规划问题求解将涉及  $m$  阶矩阵的计算( $m$  为样本的个数)，因此 SVM 不适用于超大数据集。(SMO 算法可以缓解这个问题)
- (6) 只适用于二分类问题。(SVM 的推广 SVR 也适用于回归问题；可以通过多个 SVM 的组合来解决多分类问题)