# 支撑向量机笔记

**咎志浩**
1801210709
软件与微电子学院
zanzhihao@pku.edu.cn

## Abstract

支撑向量机是一种判别式的线性分类方法。它是个二元线性分类器，核心思想在于找到一个超平面，对样本进行分类。要求将训练集中的所有样本都能正确地分类。模型的优化条件则是这个超平面与各类样本的距离达到最远，这样做的目的是：在保证模型经验误差等于0的同时最小化模型泛化误差。但这个模型仍存在着一些问题。首先，当样本不是线性可分时，SVM无法进行有效判别。为了解决这个问题，首先引入了软间隔概念。允许一定数量的样本被错误分类，同时在优化目标中加入最小化被错误分类的样本数量条件，增强模型的泛化能力。但软间隔无法解决“亦或”这类非线性问题，由此引入了核映射这一概念对SVM进行改进，通过将样本特征空间映射到高维空间中，实现非线性分类。最后，简单的支撑向量机只能解决二分类问题。通过使用One vs All 方法或同时训练多分类器实现多分类的支撑向量机

## 1 基础模型

### 1.1 基本思想

在数据是线性可分时，支撑向量机的理念是希望找到一个超平面对样本进行划分。样本按照类别分布在超平面两侧。但满足条件的超平面可能有无数个。因此，我们定义一个最优的条件：该超平面与最近的点的距离最大化，从而达到泛化误差最小化的目的。对于点$i$, 其与该超平面的距离为$\gamma_i$，令$\hat{\gamma} = \min \gamma_i \ for \ each \ i$。那么我们优化目标就是$\max \hat{\gamma}$ 那么如何将这样一个理念转换为数学表达呢？

### 1.2 数学表达

SVM的基本思想简而言之就是找到一个能够将所有训练样本正确分类且间隔最大化的超平面。那么我们可将其转化为如下的数学表达：

首先我们需要考虑如何计算点到平面的距离：对于点$X_i$，与平面方程$W*X+b=0$的距离为$\frac{W*X_i+b}{\|W\|_2}$。不妨令$a=W*X_i+b$,那么上文提到的优化目标$\hat{\gamma}=\frac{a}{\|W\|_2}$

SVM的限制条件是每一个样本都被正确地分类，同时每个样本点与超平面的距离都要大于$\hat{\gamma} \Rightarrow \gamma_i * Y_i > \hat{\gamma}$。在乘式中加入$Y_i$是为了将正例和反例都用相同的式子表达。将$\gamma$中的分母$\|W\|_2$去除可得$(W*X_i)Y_i >= a$ 优化目标和限制条件中的$a$是相同的，因而可以被约掉。

那么我们的模型条件可以被归结为：

$$\max \frac{1}{\|W\|_2}$$

$$s.t.(W*X_i+b)*Y_i >= 1 \ \forall j$$

由于我们习惯性求解最小化问题，因此上述条件可改写为：

$$\min \|W\|_2$$

$$s.t.(W*X_i+b)*Y_i >= 1 \ \forall j$$

从而将寻找最大间隔超平面问题抽象为了一个二次规划求解的问题。

## 2 约束优化

直接求解这个二次规划问题的时间复杂度与样本特征的维度有关。如果在进行了后文中的核映射之后样本的特征空间会非常大。从而也会带来大量的计算。而该问题的拉格朗日对偶问题的时间复杂度和样本数量相关却与特征维度无关。因此常常通过求解拉格朗日对偶问题来简化计算。

### 2.1 一般约束的拉格朗日对偶问题

在这里，我们首先探讨对于一般的约束问题，如何找到它的拉格朗日对偶问题。

$$\min f(x)$$

$$s.t. g(x) <= 0$$

$$h(x) = 0$$

其中，$h(x)=0$可以看做是$h(x)<=0$和$h(x)>=0$两个条件同时成立

那么这个约束问题的拉格朗日对偶问题则是：

$$\min_x \max_{\alpha,\beta} L(x,\alpha,\beta) = f(x) + \alpha g(x) + \beta h(x) \ , \alpha >= 0$$

从约束优化变化到了无约束优化。

那么$p^* = \min_x \max_{\alpha,\beta} L(x,\alpha,\beta)$是否与$q^* = \max_{\alpha,\beta} \min_x L(x,\alpha,\beta)$等价？而根据弱对偶性，一般有

$$d^* <= p^*$$

而当满足KKT条件时，强对偶性成立。 KKT条件

$$\nabla L(x^*,\alpha^*,\beta^*) = 0$$

$$\alpha^* >= 0$$

$$g(x^*) <= 0$$

$$h(x^*) = 0$$

$$\alpha^* g(x^*) = 0$$

上述条件可以理解为

- 拉格朗日函数梯度为0
- 非负约束的参数大于零
- 满足所有约束
- 互补松弛

对于凸优化问题，KKT条件成立等价于强对偶成立。需要注意的是，对于$h(x) = 0$,必须有$h(x)$和$-h(x)$同时为凸函数。只有仿射函数$wx + b$形式满足这样的条件。

### 2.2 SVM的拉格朗日对偶问题

因此，SVM模型中的约束优化是满足这个条件的。那么将SVM中的条件带入到约束优化问题中，它的拉格朗日对偶问题是：

$$\max_{\alpha} \min_{w,b} L(w,b,a\alpha) = \frac{1}{2} W * W - \sum_i \alpha_i (W * X_i + b) Y_i - 1)$$

解出KKT条件

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_i \alpha_i y_i X_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_i \alpha_i Y_i = 0$$

那么有

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_i , j \alpha_i \alpha_j Y_i Y_j X_i X_j$$

$$s.t. \sum_i \alpha_i Y_i = 0$$

$$\alpha_i >= 0$$

在上述过程中，我们可以看到：决定超平面位置的只有处于判定边界上的点。而远离判定边界的样本存在与否并不影响超平面的位置。因此，我们把处于判定边界上的点称为支撑向量。

## 3 非线性可分

当数据不是线性可分时，基本模型无法求解。主要有两个改进方向。

### 3.1 松弛变量

当数据不是线性可分时，一个直观的想法便是允许部分样本分类错误。减少错误分类样本的数量也被作为优化目标之一。令$C$为被错误分类的样本数量。那么有

$$\min W^T * W + C$$

$$(W * X + b)Y_i >= 1 \, \forall j$$

但这样的优化目标会带来两个问题，一是无法使用对偶问题求解简化计算，二是没有衡量样本被错误分类的程度。错误分类样本与超平面的距离越远错误应该更严重，但这个目标中只是简单的考虑了错误样本的数量。

改进：引入松弛变量$\xi$。$\xi_i >= 0$恒成立，当样本i被错误分类时，$\xi_i$>1。$C$在模型中作为超参数使用。那么有

$$\min W^T * W + C \sum_i \xi_i$$

$$s.t.(W * X_i + b)Y_i >= 1 - \xi_i$$

进行这样的改进之后，优化目标又变为了一个可进行二次优化的问题。同时，当把$C$设置为inf时，该问题又变为了硬间隔问题。

从另一个角度理解优化目标。我们还可以将松弛变量$\xi$理解为边缘损失。

$$\xi_i = loss(f(X_i) - Y_i)$$

$$f(X_i) = sgn(W * X_i + b)$$

$$\xi_i = (1 - W * X + b)Y_i$$

优化条件中的$W^T * W$则是$W$的L2范数，可以视为正则项。

这样，SVM的优化问题从一个二次规划问题变为了一个和逻辑斯蒂回归相似的：加入了正则项的损失函数优化问题。

### 3.2 非线性核映射

另一种想法则是对特征进行组合变换如$\Phi x_1^2, x_1 x_2 ...$。但这会导致特征空间迅速膨胀。对于M个输入特征，如果进行d项式组合，那么就有$C_d^{d+M-1}$个组合。维度的迅速膨胀则会带来维度灾难导致样本空间变得极度稀疏。因此，在这里我们引入核方法。

核方法的思想在于，在训练和预测时定义出核函数$K(x,y) = \Phi(x) * Phi(y)$，而不是显式地定义$x,y$的映射函数$\Phi$。这样做的原因时直接计算$K(x,y)$比较容易而通过$\Phi(x)$和$\Phi(y)$计算$K(x,y)$往往不容易。在SVM的对偶问题中，我们可以看到，优化目标中只涉及到样本之间的内积，因此我们考虑将对偶问题的优化目标中的$x_i * x_j$用核函数$K(x_i, x_j)$代替。此时的优化目标为：

$$\max_{\alpha} \sum_{i} \alpha_i - \frac{1}{2} \sum_{i}, j\alpha_i\alpha_j Y_i Y_j \Phi(X_i X_j)$$

而新样本的判别式为：

$$y = \arg\max_{W} \Phi(x) + b$$

一般的核方法有如下四种

- d阶多项式
$$K(x,y) = (x * y)^d$$
- 直到d阶
$$K(x,y) = (x * y + 1)^d$$
- 高斯核
$$K(x,y) = \exp\left(-\frac{\|x-y\|_2}{2\sigma^2}\right)$$
- Sigmoid核
$$K(x,y) = \tanh(\eta x * y + v)$$

## 4 多分类问题

关于将SVM应用于多分类问题主要有两种方案。

### 4.1 One Vs All

简单来说，就是对K分类问题，训练K个SVM模型。每个模型在训练过程中将某个类别的样本作为正例，其余样本作为反例。从而训练出K个分类器。在对样本进行分类预测时，每个分类器都对其进行判别。选择置信度最高的分类结果。

### 4.2 多组权重训练

多分类的SVM模型意味着对k分类问题训练k组权重。数学表达为：

$$\min \sum_{y} W^{(y)} W$$

$$s.t. W^{(y_i)} X_i + b^{(y_i)} >= W^{(y')} X_i + b^{(y')} + 1$$

判别式为：

$$y = \arg\max W^{(k)} X + b^k$$

而软间隔版本的多分类SVM则是

$$\min \sum_y W^{(y)} W + C \sum_i \xi_i$$

$$s.t. W^{(y_i)} X_i + b(y_i) >= W^{(y')} X_i + b^{(y')} + 1 - \xi_i$$

## 5 Reference


[1] 周志华. 机器学习[M]. 北京：清华大学出版社，2016，121-145.

[2] 李航. 统计学习方法[M]. 北京：清华大学出版社，2012，95-123.