

支持向量机 (SVM)

翟婷婷

扬州大学
信息工程（人工智能）学院
zh tt@yzu.edu.cn

2023年春

课程目标

- 掌握软间隔最大化思想和线性SVM算法。
- 理解线性SVM优化问题的等价优化问题。
- 掌握核技巧和非线性SVM算法。
- 了解SVM算法的实现。

问题的提出

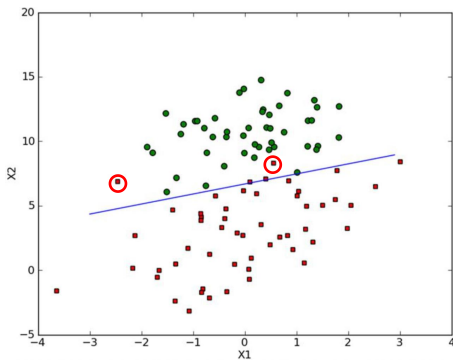
- 使用感知机和基于硬间隔最大化思想的支持向量机的前提是训练数据集是线性可分的。
- 在实际的模式识别问题中，线性可分的问题占少数，并且很难预先判断一个样本集是否是线性可分的。
- 因此需要更一般的算法，不仅能用在线性可分的数据集上，也能用在线性不可分的数据集上。
- 区分两种线性不可分的情况：
 - ① 数据集中存在少量的噪声/异常点导致线性不可分；
 - ② 问题本质上就是线性不可分的。
- 本节学习支持向量机如何处理两种线性不可分的情况。

第一种线性不可分情形—近似线性可分

➤ 考虑第一种线性不可分的情况：

数据集中存在少量的噪声/异常样本点导致线性不可分；

➤ 噪声/异常样本点存在的原因：模式采集过程中存在噪声干扰或采集出现误差。



“少量”意味着：去除这些噪声/异常样本点以后，剩下的大部分样本点组成的集合是线性可分的。

这类分类问题本质上仍是线性分类问题。

线性支持向量机

- 硬间隔最大化SVM的原始优化问题:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2$$

$$s. t. \ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1, \ \forall i = 1 \cdots N$$

- 少量的样本点导致线性不可分，意味着对于少量的样本点 (\mathbf{x}_i, y_i) ，约束 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$ 不能满足。因此，可以对每个样本 (\mathbf{x}_i, y_i) 引进一个松弛变量 $\xi_i \geq 0$ ，使得 $y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i$ ，同时要最小化松弛变量 ξ_i 的取值，这样优化问题变为：

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

$$s. t. \ y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \ \forall i = 1 \cdots N$$

$$\xi_i \geq 0, \ \forall i = 1 \cdots N$$

软间隔最大化

线性支持向量机

- $C > 0$ 是一个预定义的常数，称为惩罚系数， C 值越大表明对误分类的惩罚越大。 C 的合适取值一般由具体的应用问题决定。
- 想象一下：
 - ✓ 对于线性可分的数据集，上述优化问题求得的 $(\xi_1 \cdots \xi_N)$ 一定满足 $\xi_i = 0, \forall i = 1 \cdots N$ 。
 - ✓ 对于近似线性可分数据集，绝大部分样本点对应的松弛变量 $\xi_i = 0$ ，只有少数的噪声/异常点对应的 $\xi_i > 0$ 。

线性支持向量机器学习算法

- 输入: 线性可分或近似线性可分的训练数据集 $\{(\mathbf{x}_1, y_1) \cdots (\mathbf{x}_N, y_N)\}$, 其中 $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, C .
- 输出: 软间隔最大的分离超平面和分类决策函数。

1. 求解如下优化问题:

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

$$\text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1 \cdots N$$

得到最优解 \mathbf{w}^* , b^* ;

2. 得到分割超平面: $\mathbf{w}^{*\top} \mathbf{x} + b^* = 0$

分类决策函数 $f(\mathbf{x}) = \text{sign}(\mathbf{w}^{*\top} \mathbf{x} + b^*)$

软间隔最大化的SVM学习算法

线性支持向量机的对偶问题

➤ 线性支持向量机的原始优化问题:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1 \cdots N \\ & \xi_i \geq 0, \quad i = 1 \cdots N \end{aligned}$$

➤ 原始优化问题满足改进的slater条件, 强对偶性成立。

➤ 推导对偶优化问题:

① 定义拉格朗日函数为:

$$\begin{aligned} L(\mathbf{w}, b, \xi, \alpha, \beta) \\ = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i (1 - \xi_i - y_i(\mathbf{w}^\top \mathbf{x}_i + b)) - \sum_{i=1}^N \beta_i \xi_i \end{aligned}$$

线性支持向量机的对偶问题

② 对偶优化目标函数为:

$$\theta_D(\alpha, \beta) = \min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \beta)$$

如何求 $\min_{\mathbf{w}, b, \xi} L(\mathbf{w}, b, \xi, \alpha, \beta)$?

令 $L(\mathbf{w}, b, \xi, \alpha, \beta)$ 关于 \mathbf{w}, b, ξ_i 的偏导数分别为零:

$$\nabla_{\mathbf{w}} L(\mathbf{w}, b, \alpha) = 0 \Rightarrow \mathbf{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

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

$$\nabla_{\xi_i} L(\mathbf{w}, b, \alpha) = 0 \Rightarrow C = \alpha_i + \beta_i, i = 1, \dots, N$$

将上述等式代入拉格朗日函数得到:

线性支持向量机的对偶问题

$$\begin{aligned}\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) &= \frac{1}{2} \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right)^\top \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right) + \textcolor{violet}{c} \sum_{i=1}^N \xi_i + \sum_{i=1}^N \alpha_i - \sum_{i=1}^N \textcolor{violet}{\alpha}_i \xi_i \\ &\quad - \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right)^\top \left(\sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \right) - \textcolor{red}{b} \sum_{i=1}^N \alpha_i y_i - \sum_{i=1}^N \textcolor{violet}{\beta}_i \xi_i\end{aligned}$$

➤ 整理得到:

$$\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta}) = -\frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j + \sum_{i=1}^N \alpha_i$$

➤ 因为 $\arg \max_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha \geq 0, \beta \geq 0} \theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta})$ 等价于 $\arg \min_{\boldsymbol{\alpha}, \boldsymbol{\beta}: \alpha \geq 0, \beta \geq 0} -\theta_D(\boldsymbol{\alpha}, \boldsymbol{\beta})$

线性支持向量机的对偶问题

➤ 线性支持向量机的对偶优化问题为:

$$\min_{\alpha, \beta} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^N \alpha_i$$

$$\text{subject to } \sum_{i=1}^N \alpha_i y_i = 0$$

$$\alpha_i + \beta_i = C, i = 1, \dots, N$$

$$\alpha_i \geq 0, \beta_i \geq 0, i = 1 \dots N$$

$$0 \leq \alpha_i \leq C$$

➤ 简化后, 线性支持向量机的对偶优化问题为:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^\top \mathbf{x}_j - \sum_{i=1}^N \alpha_i$$

$$\text{subject to } \sum_{i=1}^N \alpha_i y_i = 0$$

$$0 \leq \alpha_i \leq C, i = 1 \dots N$$

由对偶问题的解构造原始问题的解

- 假设对偶问题的最优解为 $\alpha^* = (\alpha_1^* \cdots \alpha_N^*)$, 则 $\beta^* = C - \alpha^*$, 如何构造出原始问题的最优解 w^*, b^* ?
- 思路: KKT条件
- 定理3告诉我们, 满足KKT条件的 w^*, b^*, ξ^* 和 α^*, β^* 一定分别是原始问题和对偶问题的最优解。
- 由KKT条件可得到:

$$\nabla_w L(w^*, b^*, \xi^*, \alpha^*, \beta^*) = 0 \Rightarrow w^* = \sum_{i=1}^N \alpha_i^* y_i x_i$$

$$\nabla_b L(w^*, b^*, \xi^*, \alpha^*, \beta^*) = 0 \Rightarrow \sum_{i=1}^N \alpha_i^* y_i = 0$$

$$\nabla_{\xi_i} L(w^*, b^*, \xi^*, \alpha^*, \beta^*) = 0 \Rightarrow \alpha_i^* + \beta_i^* = C, i = 1, \cdots, N$$

由对偶问题的解构造原始问题的解

对偶互补条件:

$$\alpha_i^*(1 - \xi_i^* - y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*)) = 0, i = 1 \cdots N$$

$$\beta_i^* \xi_i^* = 0, i = 1 \cdots N$$

原问题约束: $y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) \geq 1 - \xi_i^*, i = 1 \cdots N$

$$\xi_i^* \geq 0, i = 1 \cdots N$$

对偶约束: $\sum_{i=1}^N \alpha_i^* y_i = 0$

$$0 \leq \alpha_i^* \leq C, i = 1 \cdots N$$

➤ 根据对偶互补条件可得:

若 $0 < \alpha_j^* < C$, 则 $y_j(\mathbf{w}^{*\top} \mathbf{x}_j + b^*) = 1 - \xi_j^*$

若 $0 < \alpha_j^* < C$, 则 $\beta_j^* > 0$, 所以 $\xi_j^* = 0$

综上, 若 $0 < \alpha_j^* < C$, 则 $y_j(\mathbf{w}^{*\top} \mathbf{x}_j + b^*) = 1$

由对偶问题的解构造原始问题的解

结合 $y_j^2 = 1$ 得:

$$b^* = y_j - \mathbf{w}^{*\top} \mathbf{x}_j = y_j - \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i^\top \mathbf{x}_j$$

✓ **定理:** 若线性SVM对偶问题的最优解为 $\boldsymbol{\alpha}^* = (\alpha_1^* \cdots \alpha_N^*)$, 则若存在下标 j , 使得 $0 < \alpha_j^* < C$, 可按如下公式得到原问题的最优解 (\mathbf{w}^*, b^*) :

$$\mathbf{w}^* = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i,$$
$$b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i^\top \mathbf{x}_j$$

线性SVM的对偶学习算法

➤ 输入: 线性可分或线性近似可分的训练数据集

$\{(\mathbf{x}_1, y_1) \cdots (\mathbf{x}_N, y_N)\}$, 其中 $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{+1, -1\}$, C .

➤ 输出: 最大间隔的分离超平面和分类决策函数。

1. 求解如下优化问题:

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^N \alpha_i \\ \text{s.t.} \quad & \sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1 \cdots N \end{aligned}$$

得到最优解 $\boldsymbol{\alpha}^* = (\alpha_1^* \cdots \alpha_N^*)$;

2. 计算原问题最优解: $\mathbf{w}^* = \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i$, 选择 $\boldsymbol{\alpha}^*$ 的一个分量满足 $0 < \alpha_j^* < C$, 计算 $b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i \mathbf{x}_i^T \mathbf{x}_j$

3. 得到分割超平面: $\mathbf{w}^{*T} \mathbf{x} + b^* = 0$

分类决策函数 $f(\mathbf{x}) = \text{sign}(\mathbf{w}^{*T} \mathbf{x} + b^*)$

KKT条件推出的结论

► 由KKT条件可以得到以下结论:

① 若 $\alpha_i^* = 0 \Rightarrow \beta_i^* = C \Rightarrow \xi_i^* = 0 \Rightarrow y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) \geq 1$
 $\Rightarrow \alpha_i^* = 0$ 对应的 (\mathbf{x}_i, y_i) 能以充分的置信度被正确分类。

② 若 $0 < \alpha_i^* < C \Rightarrow y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) = 1$
 $\Rightarrow 0 < \alpha_i^* < C$ 对应的 (\mathbf{x}_i, y_i) 在间隔边界 h_1 和 h_2 上:

$$h_1: \mathbf{w}^{*\top} \mathbf{x} + b^* = 1, \quad h_2: \mathbf{w}^{*\top} \mathbf{x} + b^* = -1$$

③ 若 $\alpha_i^* = C \Rightarrow \beta_i^* = 0 \Rightarrow \xi_i^* \geq 0$

若 $0 < \xi_i^* \leq 1 \Rightarrow y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) \geq 1 - \xi_i^* \geq 0$

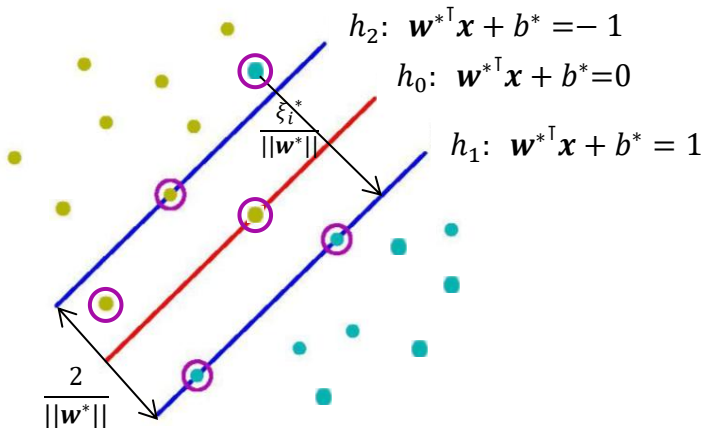
对应的 (\mathbf{x}_i, y_i) 能被正确分类, 但没有充分的置信度。

若 $\xi_i^* > 1 \Rightarrow y_i(\mathbf{w}^{*\top} \mathbf{x}_i + b^*) \geq 1 - \xi_i^*, 1 - \xi_i^* < 0$

对应的 (\mathbf{x}_i, y_i) 会被错误分类。

基于软间隔最大化学习到的分类器允许少量的样本被错分类! !

支持向量

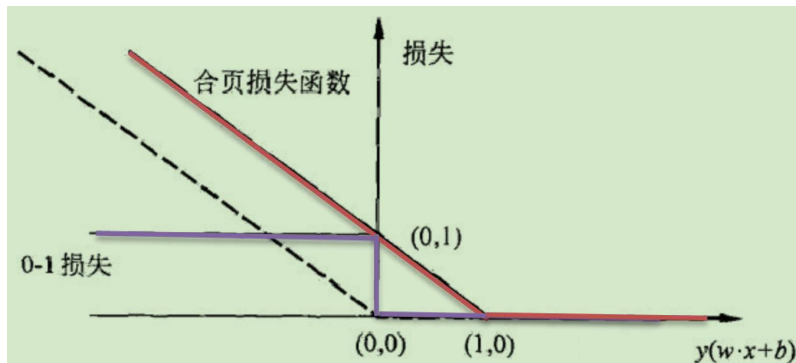


- 对应于 $\alpha_i^* > 0$ 的样本点 (\mathbf{x}_i, y_i) 称为软间隔的支持向量。
- 圈起来的都是支持向量。

线性支持向量机的等价问题

➤ 定义合页损失函数(hinge loss function):

$$l(y(\mathbf{w}^T \mathbf{x} + b)) = \max\{0, 1 - y(\mathbf{w}^T \mathbf{x} + b)\}$$
$$= [1 - y(\mathbf{w}^T \mathbf{x} + b)]_+$$



线性支持向量机的等价问题

➤ 线性SVM的原始优化问题:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1 \cdots N \\ & \xi_i \geq 0, \quad i = 1 \cdots N \end{aligned}$$

等价于如下的优化问题:

$$\min_{\mathbf{w}, b} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+$$

$\lambda > 0$ 是一个预定义的常数, 称为正则化项系数。

有约束



无约束

线性支持向量机的等价问题

➤ 证明:

令 $\xi_i = [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+ = \max\{0, 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)\}$

则 $\xi_i \geq 0$, 且 $\xi_i \geq 1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)$ 。

也就是说, ξ_i 的取值里就蕴含了原问题的约束条件,
所以原问题就等价于:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+$$

进一步等价于:

$$\min_{\mathbf{w}, b} \frac{1}{2C} \|\mathbf{w}\|^2 + \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+$$

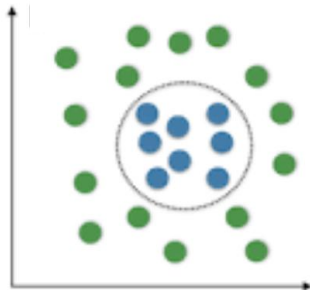
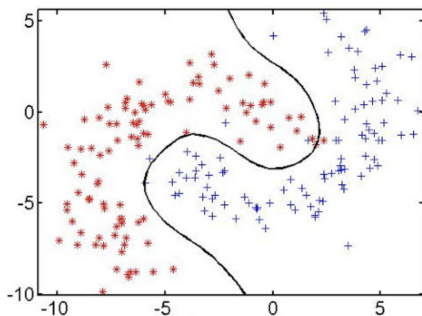
➤ 取 $\lambda = \frac{1}{2C}$, 就得到等价的优化问题。

第二种线性不可分情形—决策边界非线性

➤ 考虑第二种线性不可分的情况：

问题本质上就是线性不可分的，即决策边界是非线性的。

这种非线性问题，用一个超曲面才能将两类样本分开。



非线性SVM—求解思路

- SVM是如何解决线性不可分问题的呢？
- 广义线性化：
 - ① 将原始的线性不可分的数据集，通过一个非线性映射 Φ ，映射到一个新的特征空间中变为线性可分或近似线性可分的数据集；
 - ② 然后利用线性SVM算法在新的特征空间中找到一个分类决策超平面。
 - ③ 对未知类别的实例进行分类时，也需要首先将其映射到新的特征空间中，再利用找到的决策超平面对其分类。
- 用数学语言描述上述过程：

非线性SVM—求解思路

- 原始空间的样本 \mathbf{x}_i 经过映射后变为 $\Phi(\mathbf{x}_i)$;
- 求解线性SVM的对偶优化问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \underline{\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)} - \sum_{i=1}^N \alpha_i$$

$$\text{subject to } \sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, \quad i = 1 \dots N$$

得到最优解 $\boldsymbol{\alpha}^* = (\alpha_1^* \dots \alpha_N^*)$, 计算得到 $\mathbf{w}^* = \sum_{i=1}^N \alpha_i^* y_i \Phi(\mathbf{x}_i)$

选择 $\boldsymbol{\alpha}^*$ 的一个分量 $0 < \alpha_j^* < C$, 计算 $b^* = y_j -$

$$\sum_{i=1}^N \alpha_i^* y_i \underline{\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)}$$

- 分类决策函数可以表示为:

$$f(\mathbf{x}) = \text{sign}(\mathbf{w}^{*T} \Phi(\mathbf{x}) + b^*) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \underline{\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x})} + b^*\right)$$

非线性SVM—核技巧(kernel trick)

- 然而，实现广义线性化的难点有：
 - ① 如何找到一个合适的映射 Φ ？
 - ② 新的特征空间，即 $\Phi(\mathbf{x}_i)$ 往往维度很高，会造成巨大计算量问题，如何解决这个问题？
- 观察发现：在映射后的特征空间中求解线性SVM的对偶优化问题时，要优化的目标函数和最终得到的分类决策函数中，都只涉及到样本在新空间中的内积 $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ 。
- 如果样本的内积 $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ 在即使不知道映射 Φ 的情况下也能高效计算出，那么就能解决上述两个难点。
- 问题变为：高效 $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ 可能吗？

答案：可以，通过核技巧

非线性SVM—核技巧(kernel trick)

- 核技巧依赖于核函数。
- 核函数：是一类函数，它的输入是低维空间中的两个向量，输出是这两个向量经过同一个映射到另一个空间中的内积。
- 换句话说，使用核函数，能在低维空间中直接计算某些高维空间中的向量内积，而无需进行向量从低维空间到高维空间的映射变换。
- 什么是核函数，满足什么条件的函数才能成为核函数？

非线性SVM—核技巧(kernel trick)

➤ 核函数定义:

设 \mathcal{X} 是输入空间(欧氏空间 \mathbb{R}^n 的子集或离散集合), \mathcal{H} 为特征空间(希尔伯特空间), $\mathcal{K}(\mathbf{x}, \mathbf{z}): \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 是一个函数。如果存在一个从 \mathcal{X} 到 \mathcal{H} 的映射 $\Phi(\mathbf{x}): \mathcal{X} \rightarrow \mathcal{H}$, 使得对于任意的 $\mathbf{x}, \mathbf{z} \in \mathcal{X}$, 函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 都满足

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^\top \Phi(\mathbf{z})$$

则称 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 为核函数。

- 上述定义告诉我们, 核函数是定义在输入空间上的, 但它蕴含了一个从输入空间 \mathcal{X} 到特征空间 \mathcal{H} 的映射 Φ 。给定核函数, 特征空间 \mathcal{H} 和映射 Φ 的取法不唯一。

非线性SVM—核技巧(kernel trick)

- 通常我们使用的核函数都是正定核函数。
- 关于正定核函数的知识，简单了解。
- 正定核函数定义：设 $\mathcal{X} \subset \mathbb{R}^n$ 是输入空间， $\mathcal{K}(\mathbf{x}, \mathbf{z}): \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ 是一个对称函数，即对任意 $\mathbf{x}, \mathbf{z} \in \mathcal{X}$ ，都有 $\mathcal{K}(\mathbf{x}, \mathbf{z}) = \mathcal{K}(\mathbf{z}, \mathbf{x})$ 。如果对于任意有限个元素 $\mathbf{x}_i \in \mathcal{X}, i = 1, 2, \dots, m$ ， $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 对应的Gram矩阵

$$[\mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)]_{m \times m} = \begin{bmatrix} \mathcal{K}(\mathbf{x}_1, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_1, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_1, \mathbf{x}_m) \\ \mathcal{K}(\mathbf{x}_2, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_2, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_2, \mathbf{x}_m) \\ \vdots & \vdots & \ddots & \vdots \\ \mathcal{K}(\mathbf{x}_m, \mathbf{x}_1) & \mathcal{K}(\mathbf{x}_m, \mathbf{x}_2) & \cdots & \mathcal{K}(\mathbf{x}_m, \mathbf{x}_m) \end{bmatrix}$$

是半正定的，则称 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 是正定核函数。

- 根据这个定义，要检验一个函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 是否是正定核函数并不容易，因为需要对输入空间中的任意有限个元素 $\{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$ 验证该函数对应的Gram矩阵是否是半正定的。

非线性SVM—核技巧(kernel trick)

➤ 实际问题中，常常使用已有的核函数。

➤ 常用的核函数：

✓ 多项式核函数：（在自然语言处理中很受欢迎）

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = (\mathbf{x}^\top \mathbf{z} + c)^d, \quad d = 1, 2, \dots$$

d 是多项式的度。当 $d = 2$ 时， $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 对应的特征映射为：

$$\Phi(\mathbf{x}) = (x_n^2, \dots, x_1^2, \sqrt{2}x_n x_{n-1}, \dots, \sqrt{2}x_n x_1, \\ \sqrt{2}x_{n-1} x_{n-2}, \dots, \sqrt{2}x_{n-1} x_1, \dots, \sqrt{2}x_2 x_1, \sqrt{2}c x_n, \dots, \sqrt{2}c x_1, c)$$

映射 Φ 将 \mathbf{x} 从 n 维空间中映射到 $\frac{(n+1)(n+2)}{2}$ 维空间！

非线性SVM—核技巧(kernel trick)

✓ 高斯径向基核函数:

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

SVM非线性分类中最常用的核函数，它蕴含的映射将原始空间中样本映射到一个**无穷维度**的特征空间中。 σ 是核函数的参数。

➤ 一旦选择一个核函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 后，不需要知道该核函数蕴含的映射 Φ 和映射后的高维特征空间，就可以按照如下方式计算在高维特征空间中的内积：

$$\Phi(\mathbf{x})^\top \Phi(\mathbf{z}) = \mathcal{K}(\mathbf{x}, \mathbf{z})$$

非线性SVM的学习算法

- 选择合适的核函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 和惩罚参数 C ;
- 求解线性SVM的对偶优化问题:

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

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

得到最优解 $\boldsymbol{\alpha}^* = (\alpha_1^* \cdots \alpha_N^*)$, 选择 $\boldsymbol{\alpha}^*$ 的一个分量 $0 < \alpha_j^* < C$, 计算 $b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$

- 分类决策函数可以表示为:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b^*\right)$$

非线性SVM的学习算法

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b^*\right)$$

- 当采用高斯径向基核函数时:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{2\sigma^2}\right) + b^*\right)$$

分类决策函数 $f(\mathbf{x})$ 是一个非线性的函数。

- 最后一个问题：如何选择合适的核函数及其参数？
一般来说，只能依靠经验，或者通过不断尝试。
- 一般情况下，核函数方法配合软间隔方法，能够解决绝大多数的非线性分类问题。

核技巧的特点

- 核技巧，也称核函数方法，应用非常广泛，不仅限于在非线性SVM中。它的特点：
 - ① 核函数的引入避免了“维数灾难”，大大减小了计算量。
 - ② 无需知道非线性变换函数 Φ 的形式和参数。
 - ③ 核函数的形式和参数的变化会隐式地改变从输入空间到特征空间的映射，进而对特征空间的性质产生影响，最终改变各种核方法的性能。
 - ④ 核方法可以和不同的算法相结合，形成多种不同的基于核函数技术的方法。

SVM算法的实现和拓展

➤ SVM算法的实现:

- ① SMO (sequential minimal optimization): by John Platt in 1998 at Microsoft Research

参考: http://en.wikipedia.org/wiki/Sequential_minimal_optimization

实现库: LIBSVM, SVMLight, scikit-learn

- ② Pegasos (Primal Estimated sub-GrAdient SOLver for SVM): by Shai Shalev-Shwartz et al. in 2007

参考: <http://www.cs.huji.ac.il/~shais/code/>

- ③ LIBLINEAR (large-scale linear classification, for data with millions of instances and features)

参考: <https://www.csie.ntu.edu.tw/~cjlin/liblinear/>

SVM算法的实现和拓展

➤ SVM算法的拓展(继续阅读):

① 多分类SVM算法

参考paper: On the Algorithmic Implementation of Multiclass Kernel-based Vector Machines

② 非线性SVM算法

参考paper: Breaking the curse of kernelization: budgeted stochastic gradient descent for large-scale SVM training

➤ SVM和统计学习

阅读: A Tutorial on Support Vector Machines for Pattern Recognition

链接: <http://research.microsoft.com/pubs/67119/svmtutorial.pdf>

最新会议论文集:ICML、NIPS、AISTATS、COLT、...

小结

- 当训练数据集中存在少量的噪声或异常的样本点导致线性不可分时，可以通过引入松弛变量 ξ_i ，使噪声或异常样本点“可分”，从而得到线性SVM的原始优化问题：

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

$$\text{s.t. } y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \forall i = 1 \dots N$$



求得最优解记为 \mathbf{w}^*, b^* ;

得到的分类决策边界为 $\mathbf{w}^{*\top} \mathbf{x} + b^* = 0$ ，分类决策函数为 $f(\mathbf{x}) = \text{sign}(\mathbf{w}^{*\top} \mathbf{x} + b^*)$ 。

- 线性SVM优化问题的解 \mathbf{w}^* 是唯一的，但 b^* 不唯一。

小结

- 线性SVM的对偶优化问题:

$$\min_{\alpha} \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j - \sum_{i=1}^N \alpha_i$$

重点

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

通过求解对偶问题的最优解 α^* ，可以求出原始问题的最优解 \mathbf{w}^*, b^* 。

- 线性SVM对偶问题的解 α^* 中满足 $\alpha_i^* > 0$ 的 \mathbf{x}_i 称为支持向量。支持向量可以在间隔边界上，也可在间隔边界与决策边界之间，或在决策边界误分的那一侧。分类决策超平面完全由支持向量决定。
- 线性SVM算法比线性可分的SVM算法更通用。

小结


➤ 线性SVM的原始优化问题:

$$\begin{aligned} \min_{\mathbf{w}, b, \xi} \quad & \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i \\ \text{s.t.} \quad & y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1 - \xi_i, \xi_i \geq 0, \quad \forall i = 1 \dots N \end{aligned}$$

等价于: 正则化的合叶损失最小化问题:

$$\min_{\mathbf{w}, b} \lambda \|\mathbf{w}\|^2 + \sum_{i=1}^N [1 - y_i(\mathbf{w}^\top \mathbf{x}_i + b)]_+$$

小结

- 决策边界非线性的分类问题是真正的非线性分类问题。
- 对于输入空间中的非线性分类问题，可以通过非线性变换将它转化为某个高维特征空间中的线性分类问题，然后在高维特征空间中学习线性SVM。
- 由于在SVM对偶优化问题中，目标函数和分类决策函数中都只涉及到样本间的内积运算，所以不需要显式地指定非线性变换，而是用核函数替代内积。
- 核函数是一类函数，意味着存在一个从输入空间到特征空间的映射 Φ ，使得对输入空间中的任意两个点 \mathbf{x}, \mathbf{z} ，都有

$$\mathcal{K}(\mathbf{x}, \mathbf{z}) = \Phi(\mathbf{x})^\top \Phi(\mathbf{z})$$

小结

- 一个对称函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 是正定核的充要条件是：对于输入空间中的任意有限个点 $\mathbf{x}_1, \mathbf{x}_2 \cdots \mathbf{x}_m$ 和任意正整数 m , $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 对应的Gram矩阵是半正定的。
- 在线性SVM的对偶问题中，用核函数 $\mathcal{K}(\mathbf{x}, \mathbf{z})$ 替代内积，得到的就是非线性SVM的优化问题：

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

重点

subject to $\sum_{i=1}^N \alpha_i y_i = 0, \quad 0 \leq \alpha_i \leq C, i = 1 \cdots N$

得到最优解 $\boldsymbol{\alpha}^* = (\alpha_1^* \cdots \alpha_N^*)$;

选择 $\boldsymbol{\alpha}^*$ 的一个分量 $0 < \alpha_j^* < C$,

计算 $b^* = y_j - \sum_{i=1}^N \alpha_i^* y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}_j)$

小结

得到分类决策函数:

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \alpha_i^* y_i \mathcal{K}(\mathbf{x}_i, \mathbf{x}) + b^*\right)$$

- SVM的实现方法: SMO, Pegasos
- SVM的拓展: 多分类