

Machine Learning Note

目录

1 模型评估与选择	1
1.1 经验误差与过拟合	1
1.2 评估方法	1
1.2.1 留出法	2
1.2.2 交叉验证法	2
2 支持向量机	2
2.1 间隔与支持向量	2
2.2 对偶问题	4
2.3 核函数	5
2.4 软间隔与正则化	6
2.5 支持向量回归	8

1. 模型评估与选择

1.1 经验误差与过拟合

学习器在训练集上的误差称为“训练误差”或“经验误差”(empirical error), 在新样本上的误差称为“泛化误差”(generalization error).

“过拟合”: 训练样本学的太好, 泛化性能下降 /quad 是机器学习的关键障碍, 无法彻底避免

“欠拟合”: 是指对训练样本的一般性质尚未学好

1.2 评估方法

现实中要考虑时间开销, 存储开销 (这俩也算是降维兴起的原因), 可解释性等方面因素, 这里只考虑泛化误差.

1.2.1 留出法

直接将数据集 D 划分为两个互斥的集合, 训练集 S 和测试集 T . 训练/测试集的划分要尽量保持数据分布的一致性.

单次使用留出法得到的估计结果不够稳定可靠, 一般采用多次随机划分, 重复试验取平均值作为评估结果.

1.2.2 交叉验证法

先将数据集 D 划分成 k 个 (k 折) 大小相似的互斥子集, 每个子集都尽可能保持数据分布的一致性, 即从 D 中通过分层采样得到. 然后每次用 $k-1$ 个子集的并集作为训练集, 余下子集作为测试集, 进行 k 次训练和测试, 最终返回的是 k 个测试结果的均值.

为减小因样本划分不同而引入的差别, k 折要随机用不同的划分重复 p 次, 最终评估结果是 p 次 k 折交叉验证结果的均值.

假定 D 包含 m 个样本, 若令 $k=m$, 得到特例: 留一法 (Leave-One-Out, LOO)

2. 支持向量机

2.1 间隔与支持向量

样本空间中划分超平面:

$$\omega^T \mathbf{x} + b = 0 \quad (2.1)$$

$\omega = \{\omega_1; \omega_2; \dots; \omega_d\}$ 为法向量, 决定超平面的方向; b 为位移项, 决定超平面与原点之间的距离.

样本空间任一点 \mathbf{x} 到超平面 (ω, b) 的距离:

$$r = \frac{|\omega^T \mathbf{x} + b|}{\|\omega\|} \quad (2.2)$$

设超平面 (ω, b) 能将训练样本正确分类, 即对于超平面 $(x_i, y_i) \in D$, 若 $y_i = +1$, 则有 $\omega^T \mathbf{x} + b > 0$; 若 $y_i = -1$, $\omega^T \mathbf{x} + b < 0$. 另

$$\begin{cases} \omega^T \mathbf{x}_i + b \geqslant +1 & y_i = +1 \\ \omega^T \mathbf{x}_i + b \leqslant -1 & y_i = -1 \end{cases} \quad (2.3)$$

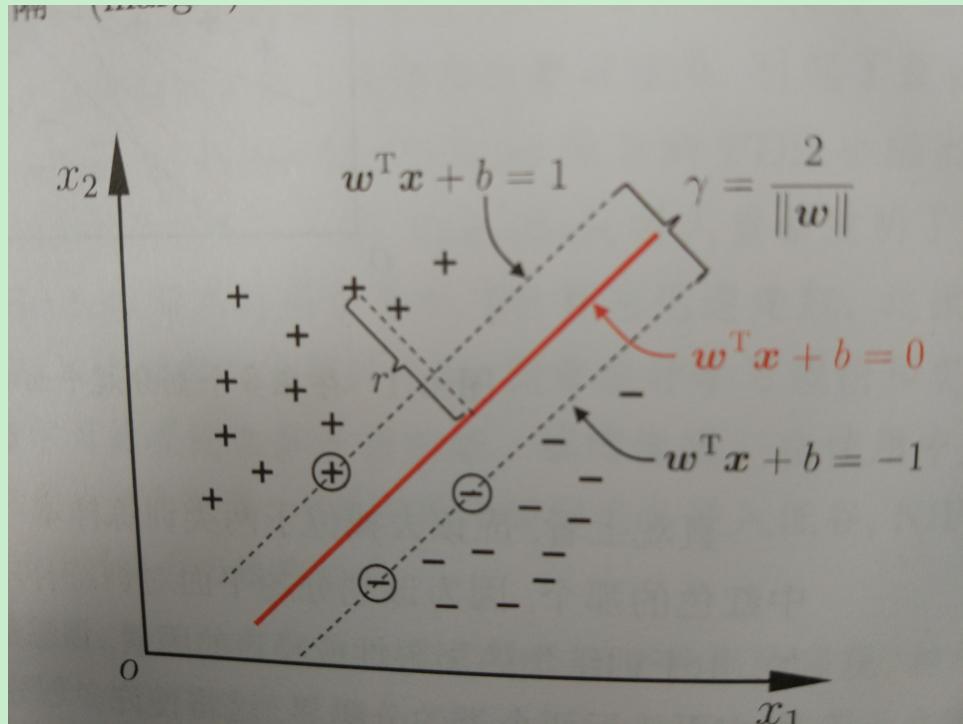


图 1: 支持向量与间隔

如图1, 距离超平面最近的几个样本点使式??的等号成立, 每个样本点对应一个特征向量, 他们被称为“支持向量”, 两个异类支持向量到超平面的距离和:

$$\gamma = \frac{2}{\|\omega\|} \quad (2.4)$$

称为“间隔”(margin).

欲找到具有“最大间隔”(maximum margin) 的划分超平面, 即

$$\begin{aligned} & \max_{\omega, b} \frac{2}{\|\omega\|} \\ \text{s.t. } & y_i(\omega^T x_i + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned} \quad (2.5)$$

b 通过约束隐式地影响 ω , 所以间隔与 b 和 ω 都有关.

式2.5可以重写为:

$$\begin{aligned} & \min_{\omega, b} \frac{\|\omega\|^2}{2} \\ \text{s.t. } & y_i(\omega^T x_i + b) \geq 1, \quad i = 1, 2, \dots, m. \end{aligned} \quad (2.6)$$

即支持向量机 (SVM) 的基本型.

2.2 对偶问题

希望通过解式2.6得到大间隔划分超平面对应的模型

$$f(\mathbf{x}) = \boldsymbol{\omega}^T \mathbf{x} + b \quad (2.7)$$

对式2.6使用拉格朗日乘子法可得到其“对偶问题”(线性规划有一个有趣的特性，就是任何一个求极大的问题都有一个与其匹配的求极小的线性规划问题). 对2.6每条约束添加拉格朗日乘子 $\alpha_i \geq 0$, 拉格朗日函数可写为

$$L(\boldsymbol{\omega}, b, \boldsymbol{\alpha}) = \frac{\|\boldsymbol{\omega}\|^2}{2} + \sum_{i=1}^m \alpha_i (1 - y_i (\boldsymbol{\omega}^T \mathbf{x}_i + b)) \quad (2.8)$$

$\boldsymbol{\alpha} = (\alpha_1; \alpha_2; \dots; \alpha_m)$, 令 $L(\boldsymbol{\omega}, b, \boldsymbol{\alpha})$ 对 $\boldsymbol{\omega}$ 和 b 的偏导为 0 得

$$\boldsymbol{\omega} = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i \quad (2.9)$$

$$0 = \sum_{i=1}^m \alpha_i y_i \quad (2.10)$$

2.9代入2.8, 考虑2.10的约束, 就可得到2.6的对偶问题

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned} \quad (2.11)$$

解出 $\boldsymbol{\alpha}$, 求出 $\boldsymbol{\omega}$ 与 b 即可得到模型

$$f(\mathbf{x}) = \boldsymbol{\omega}^T \mathbf{x} + b = \sum_{i=1}^m \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b \quad (2.12)$$

因式2.6中有不等式约束, 上述过程需满足 KKT(Karush-Kuhn-Tucker) 条件, 即要求

$$\begin{cases} \alpha_i \geq 0; \\ y_i f(\mathbf{x}_i) - 1 \geq 0; \\ \alpha_i(y_i f(\mathbf{x}_i) - 1) = 0. \end{cases} \quad (2.13)$$

于是, 对于任意训练样本 (\mathbf{x}_i, y_i) , 总有 $\alpha_i = 0$ 或 $y_i f(\mathbf{x}_i) = 1$. 若 $\alpha_i = 0$, 则该样本不会在式2.12中出现, 也就不会对 $f(\mathbf{x})$ 产生影响; 若 $\alpha_i > 0$, 则必有 $y_i f(\mathbf{x}_i) = 1$, 所对应的样本点位于最大间隔边界上, 是一个支持向量. 因此, 训练完成后, 大部分的训练样本都不需要保留, 最终模型仅与支持向量有关. 此外, 求解式2.11, 可以用 SMO 算法.

2.3 核函数

不是所有的原始样本空间都存在能正确划分两类样本的超平面. 可将样本从原始空间映射到一个更高维的特征空间. 如果原始空间是有限维, 那么一定存在一个高维特征空间使样本可分.

令 $\Phi(\mathbf{x})$ 表示将 \mathbf{x} 映射后的特征向量, 在特征空间中划分超平面所对应的模型为

$$f(\mathbf{x}) = \boldsymbol{\omega}^T \Phi(\mathbf{x}) + b \quad (2.14)$$

对偶问题什么的和之前都差不多

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \\ \text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\ & \alpha_i \geq 0, \quad i = 1, 2, \dots, m \end{aligned} \quad (2.15)$$

特征空间维数可能很高, 直接计算 $\Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j)$ 通常困难, 设想一个函数:

$$\kappa(\mathbf{x}_i, \mathbf{x}_j) = \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}_j) \rangle = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) \quad (2.16)$$

即 \mathbf{x}_i 和 \mathbf{x}_j 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” $\kappa(.,.)$ 计算的结果. 式2.15重写为

$$\begin{aligned}
\max_{\alpha} \quad & \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i=1}^m \sum_{j=1}^m \alpha_i \alpha_j y_i y_j \kappa(\mathbf{x}_i, \mathbf{x}_j) \\
\text{s.t.} \quad & \sum_{i=1}^m \alpha_i y_i = 0, \\
& \alpha_i \geq 0, \quad i = 1, 2, \dots, m
\end{aligned} \tag{2.17}$$

求解后得到

$$\begin{aligned}
f(\mathbf{x}) &= \boldsymbol{\omega}^T \Phi(\mathbf{x}) + b \\
&= \sum_{i=1}^m \alpha_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + b \\
&= \sum_{i=1}^m \alpha_i y_i \kappa(\mathbf{x}, \mathbf{x}_i) + b
\end{aligned} \tag{2.18}$$

式2.18显示出模型最优解就可通过训练样本的核函数展开, 这一展式也称“支持向量展示”.

定理 6.1 (核函数) 令 χ 为输入空间, $\kappa(\cdot, \cdot)$ 是定义在 $\chi \times \chi$ 上的对称函数, 则 κ 是核函数当且仅当对于任意数据 $D = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m\}$, “核矩阵” K 总是半正定的. 定理 6.1 表明, 只要一个对称函数所对应的核矩阵半正定, 它就能作为核函数使用.

常用核函数: 线性核, 多项式核, 高斯核, 拉普拉斯核, Sigmoid 核, 还有函数组合的核.

2.4 软间隔与正则化

现实中很难确定合适的核函数使得训练样本在特征空间中线性可分, 即使找到了, 可分的结果也许是过拟合造成的.

缓解的一个办法是允许支持向量机在一些样本上出错. 为此, 引入图2“软间隔”(soft margin) 概念.

支持向量机形式是要求所有样本均满足约束2.3, 即所有样本都必须划分正确, 称“硬间隔”(hard margin). 软间隔允许某些样本不满足约束 (不满足约束的样本应尽可能少)

$$y_i(\boldsymbol{\omega}^T \mathbf{x}_i + b) \geq 1 \tag{2.19}$$

优化目标为

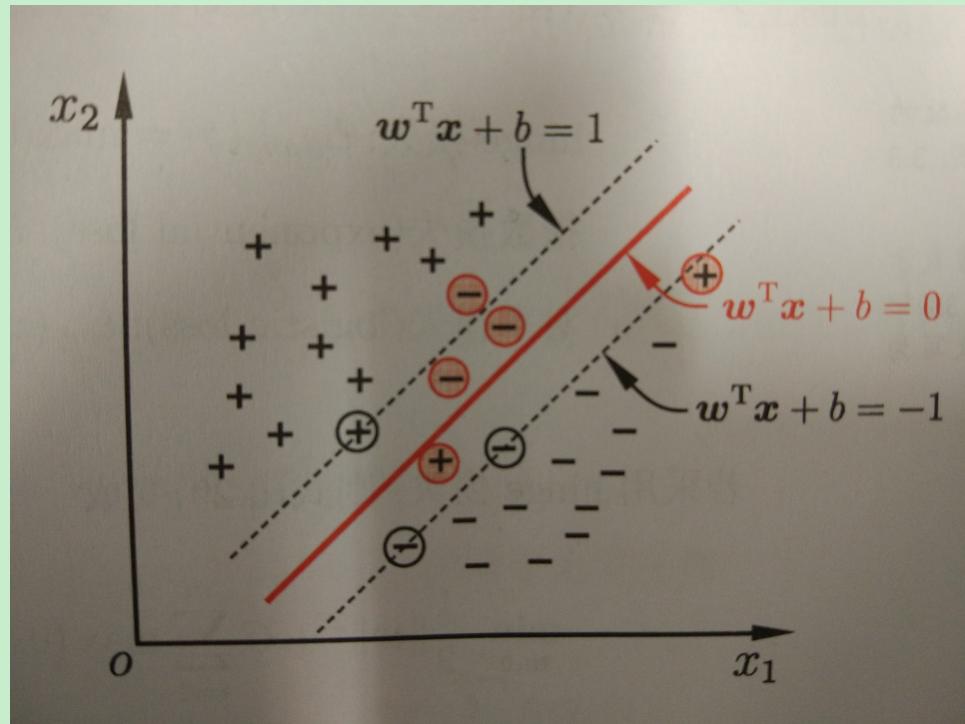


图 2: 软间隔

$$\min_{\omega, b} \quad \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m l_{0/1}(y_i(\omega^T \mathbf{x}_i + b) - 1) \quad (2.20)$$

$C > 0$ 是常数, $l_{0/1}$ 是“0/1 损失函数”

$$l_{0/1}(z) = \begin{cases} 1 & \text{if } z < 0; \\ 0 & \text{otherwise} \end{cases} \quad (2.21)$$

当 C 为无穷大时, 式 2.20 迫使所有样本均满足约束 2.19(不满足的话, 2.20 会无穷大), 2.20 等价于 2.6; C 取有限值时, 2.20 允许一些样本不满足约束.

$l_{0/1}$ 使式 2.20 不易直接求解, 常用其他函数替代之, 称“替代损失”. 之后对 2.20 用替代损失函数, 引入松弛变量, 构造拉格朗日函数.....

不管用什么损失函数, 模型具有一个共性: 目标的第一项用来描述划分超平面的“间隔”大小, 另一项用来表述训练集上的误差 (参照 2.20), 更一般的形式: (“正则化”问题)

$$\min_f \quad \Omega(f) + C \sum_{i=1}^m l(f(\mathbf{x}_i), y_i) \quad (2.22)$$

以后可能还会补充.....

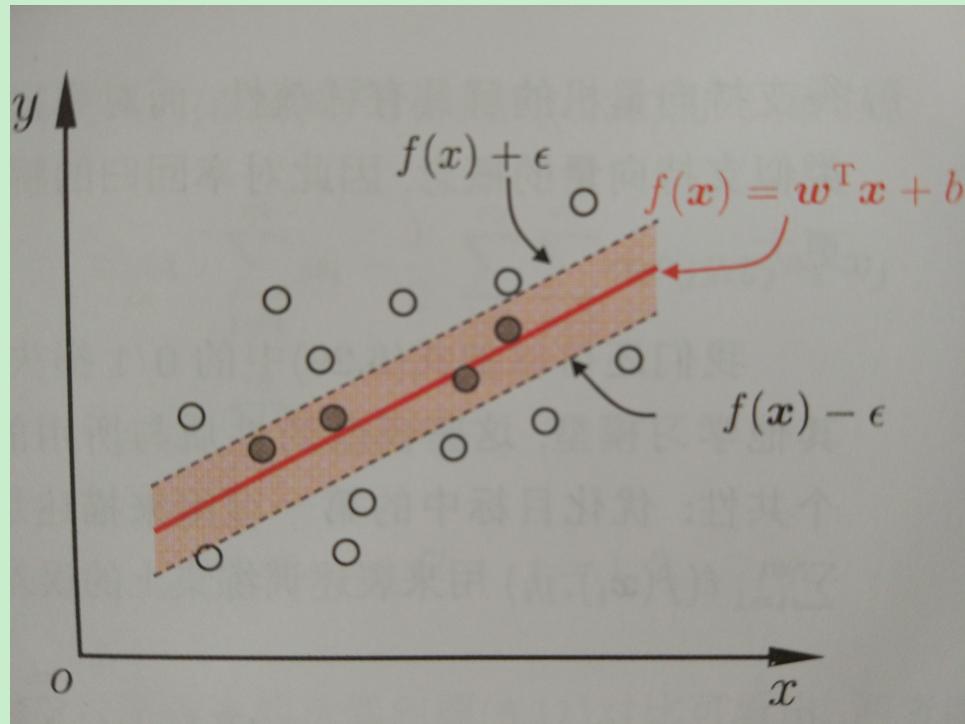


图 3: 支持向量回归, 落入橙色区域的样本不计算损失

2.5 支持向量回归

给定训练样本 $D = (\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_m, y_m), y_i \in R$, 希望学得一个形如式2.7的回归模型, 使得 $f(\mathbf{x})$ 与 y 尽可能接近, ω 和 b 是待确定的模型参数.

支持向量回归 (Support Vector Regression, SVR) 仅当 $f(\mathbf{x})$ 与 y 之间的差别绝对值大于 ϵ 时才计算损失. 如图3.

SVR 问题可形式化为

$$\min_{\omega, b} \frac{1}{2} \|\omega\|^2 + C \sum_{i=1}^m l_\epsilon(f(\mathbf{x}_i) - y_i) \quad (2.23)$$

C 为正则化常数, l_ϵ 是 ϵ -不敏感损失

$$l_\epsilon(z) = \begin{cases} 0, & \text{if } |z| \leq \epsilon; \\ |z| - \epsilon, & \text{otherwise} \end{cases} \quad (2.24)$$

引入松弛变量, 拉格朗日函数, 对偶问题, KKT, 核函数.....

SVR 可表示为

$$f(\boldsymbol{x}) = \sum_{i=1}^m (\hat{\alpha}_i - \alpha_i) \kappa(\boldsymbol{x}, \boldsymbol{x}_i) + b \quad (2.25)$$

2.6 核方法