Machine Learning Note

目录

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:

$$\boldsymbol{\omega}^T\boldsymbol{x}+b=0 \tag{2.1}$$

$\boldsymbol{\omega}=\{\omega_1;\omega_2;\dots\omega_d\}$ 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$\boldsymbol{x}$ 到超平面 $(\boldsymbol{\omega}, b)$ 的距离:

$$r=\frac{|\,\boldsymbol{\omega}^T\boldsymbol{x}+b\,|}{\|\,\boldsymbol{\omega}\,\|} \tag{2.2}$$

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

$$\begin{cases}\boldsymbol{\omega}^T\boldsymbol{x}_i+b\geqslant+1 & y_i=+1\\ \boldsymbol{\omega}^T\boldsymbol{x}_i+b\leqslant-1s & y_i=-1\end{cases} \tag{2.3}$$

如图1, 距离超平面最近的几个样本点使式2.3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:

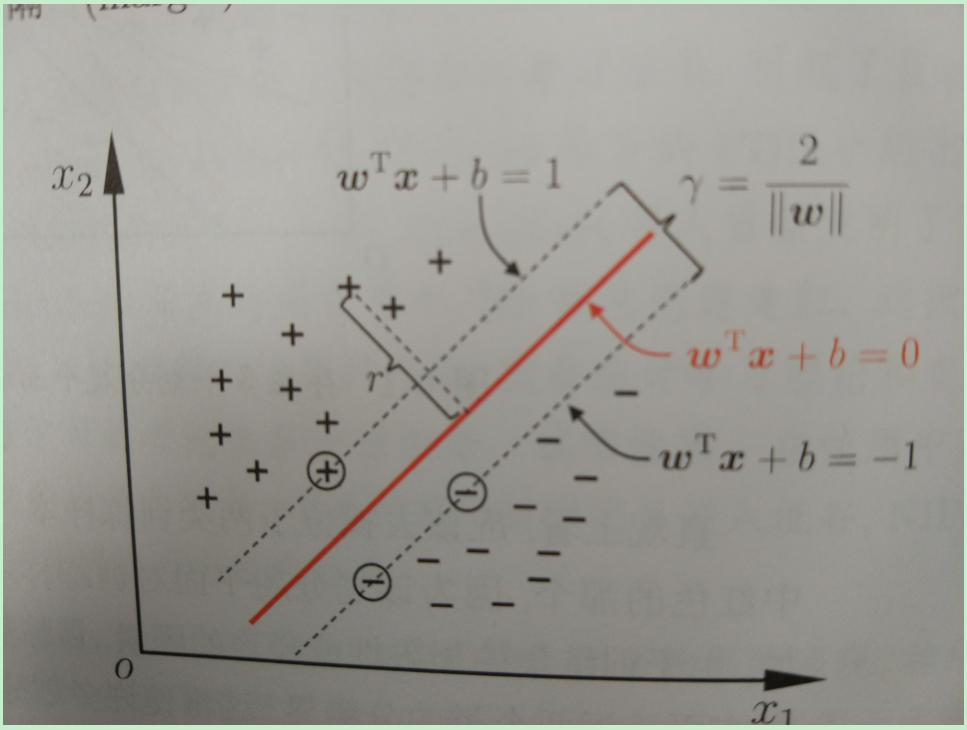

图 1: 支持向量与间隔

$$\gamma=\frac{2}{\|\boldsymbol{\omega}\|} \tag{2.4}$$

称为 “间隔”(margin).

欲找到具有 “最大间隔”(maximum margin) 的划分超平面, 即

$$\begin{aligned} &\max_{\boldsymbol{\omega},b} \quad \frac{2}{\|\boldsymbol{\omega}\|} \\ &s.t. \quad y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)\geqslant 1, \quad i=1,2,\ldots,m. \end{aligned} \tag{2.5}$$

b 通过约束隐式地影响 $\boldsymbol{\omega}$, 所以间隔与 b 和 $\boldsymbol{\omega}$ 都有关.

式2.5可以重写为:

$$\begin{aligned} &\max_{\boldsymbol{\omega},b} \quad \frac{\|\boldsymbol{\omega}\|^2}{2} \\ &s.t. \quad y_i(\boldsymbol{\omega}^T\boldsymbol{x}_i+b)\geqslant 1, \quad i=1,2,\ldots,m. \end{aligned} \tag{2.6}$$

2.2 对偶问题