SVM

Mathematical Modeling

Prof. Dr. Jingzhi Li

Department of Mathematics,
Southern University of Science and Technology

2025 Spring

1. SVM
2. SVM 的实现
3. 课堂练习

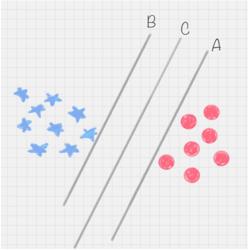
概述

SVM（Support Vector Machine）中文名为支持向量机，是一种有监督的学习模型。

- SVM 常用于分类，模式识别和回归

SVM 的工作原理

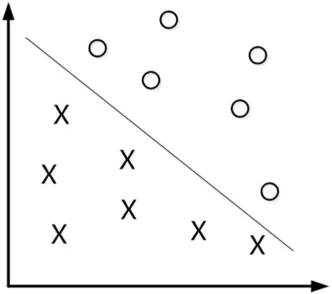
SVM 的主要思想是找到一个超平面，使得它能够尽可能多地将两类数据点正确分开，同时使分开的两类数据点距离超平面最远。



- 平面 A 更靠近蓝色球
- 平面 B 更靠近红色球
- 相比于 A 和 B，平面 C 的鲁棒性更强

线性可分

- 定义：D_0 和 D_1 是 n 维欧氏空间中的两个点集。如果存在 n 维向量 ω 和实数 b，使得所有属于 D_0 的点 x_i 都有 $\omega^T x_i + b > 0$，而对于所有属于 D_1 的点 x_j 则有 $\omega^T x_j + b < 0$，则我们称 D_0 和 D_1 线性可分。
- 直观理解：在 n 维空间上，两类点被一个超平面完全分开叫做线性可分。



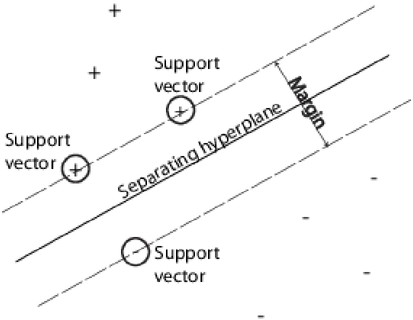
最大间隔超平面

为了使找到的超平面更具有鲁棒性，我们需要寻找最佳超平面，以最大间隔把两类样本分开的超平面，称为最大间隔超平面。

- 两类样本分别分割在该超平面的两侧
- 两侧距离超平面最近的样本点到超平面的距离被最大化

支持向量

样本中距离超平面最近的一些点称为支持向量（Support Vector）



SVM 的数学形式

在 n 维空间中，超平面可以用下面的线性方程来描述：

$$\omega^T x + b = 0$$

其中 $\omega, x \in R^n$。
任意点 $x = (x_1, x_2 \dots x_n)$ 到直线 $\omega^T x + b = 0$ 的距离为：

$$\frac{|\omega^T x + b|}{||\omega||}$$

其中 $||\omega|| = \sqrt{\omega_1^2 + \dots \omega_n^2}$ 。

假设支持向量到超平面的距离为 d，则其他点到超平面的距离大于 d，于是我们有

$$\begin{cases} \frac{\omega^T x+b}{\|\omega\|} \geq d & y=1 \\ \frac{\omega^T x+b}{\|\omega\|} \leq -d & y=-1 \end{cases}$$

即

$$\begin{cases} \frac{\omega^T x+b}{\|\omega\| d} \geq 1 & y=1 \\ \frac{\omega^T x+b}{\|\omega\| d} \leq -1 & y=-1 \end{cases}$$

$\|\omega\| d$ 恒为正数，为了方便优化，我们可以将其省略：

$$\begin{cases} \omega^T x+b \geq 1 & y=1 \\ \omega^T x+b \leq -1 & y=-1 \end{cases}$$

将两个方程合并后，对于数据集 $\mathcal{D}=\{(x_1,y_1),\ldots,(x_m,y_m)\}$，其中 $x_i\in\mathbb{R}^n$ 是输入向量，$y_i\in\{-1,1\}$ 是输入数据的分类，我们有

$$y_i\left(\omega^T x_i+b\right)\geq 1$$

接着我们要找到最大间隔超平面，简化后得到 SVM 的最优化问题

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

$$\text{s.t.}\quad y_i\left(\omega^T x_i+b\right)\geq 1$$

引入 Lagrange 乘子

构造拉格朗日函数

$$\min_{\omega,b}\max_{\lambda} L(\omega, b, \lambda) = \frac{1}{2}\|\omega\|^2 + \sum_{i=1}^{m}\lambda_i\left[1 - y_i\left(\omega^T x_i + b\right)\right]$$

其中 $\lambda_i \geq 0$ 为 Lagrange 乘子。利用强对偶性转化得

$$\min_{\omega,b}\max_{\lambda} L(\omega, b, \lambda) = \max_{\lambda}\min_{\omega,b} L(\omega, b, \lambda)$$

对参数 ω 和 b 的偏导置零，得

$$\sum_{i=1}^{m} \lambda_i x_i y_i = \omega$$

$$\sum_{i=1}^{m} \lambda_i y_i = 0$$

回代得到原问题的 Lagrange 对偶问题

$$\max_{\lambda} \sum_{j=1}^{m} \lambda_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \lambda_i \lambda_j y_i y_j \left(x_i \cdot x_j \right)$$

$$\text{s.t.} \quad \sum_{i=1}^{m} \lambda_i y_i = 0 \quad \lambda_i \geq 0$$

KKT 条件

求解上述最优化问题后得到最优解 λ_i^*，计算得到最终模型

$$f(x)=\omega^T x+b=\sum_{i=1}^{m}\lambda_i^* y_i {x_i}^T x+b$$

注意到该优化问题为凸优化问题，且满足 Slater 条件，即存在严格可行解。那么上述求解过程需满足 KKT 条件

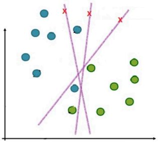
$$\begin{cases}\lambda_i \geqslant 0\\ y_i f\left(x_i\right)-1 \geqslant 0\\ \lambda_i\left(y_i f\left(\boldsymbol{x}_i\right)-1\right)=0\end{cases}$$

KKT 条件是一个非线性优化问题能有最优解法的充分必要条件。

于是，对于任意样本 (x_i, y_i)，总有 $\lambda_i^* = 0$ 或 $y_i f(x_i) = 1$，若 $\lambda_i^* = 0$，则对应样本点不会在最终模型的求和中出现。若 $\lambda_i^* > 0$，则有 $y_i f(x_i) = 1$，即该样本点是一个支持向量。我们会得到 SVM 的一个重要的性质：训练完成后，大部分样本点不需要保留，最终模型只与支持向量有关。

软间隔

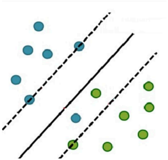
如果数据不是完全线性可分的，例如



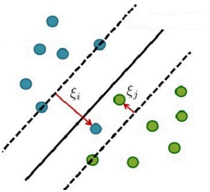
此时我们可以适当放宽约束条件，允许个别点出现在间隔层里面，即不满足

$$1 - y_i\left(\omega^T x_i + b\right) \leq 0$$

我们称其称为软间隔。



为了度量软间隔软的程度，我们为每个数据点引入松弛变量 ξ_i ，令 $\xi_i \geq 0$ ，且 $1 - y_i\left(\omega^T x_i + b\right) - \xi_i \leq 0$ 。对应如下图所示：



增加软间隔后我们的优化目标变为：

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

$$\text{s.t.} \quad g_i(\omega, b) = 1 - y_i\left(\omega^T x_i + b\right) - \xi_i \leq 0$$

$$\xi_i \geq 0$$

其中 C 是惩罚系数，代表对分类器中存在异常点的容忍程度。

构造拉格朗日函数

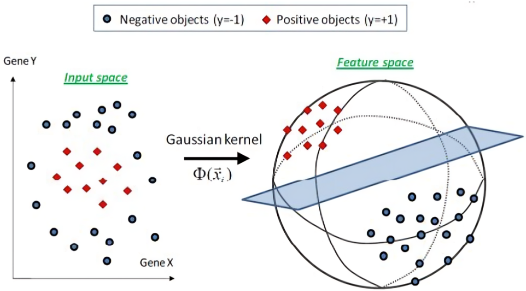
$$\min_{\omega,b,\xi}\max_{\lambda,\mu} L(\omega,b,\xi,\lambda,\mu)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{m}\xi_i$$
$$+\sum_{i=1}^{n}\lambda_i\left[1-\xi_i-y_i\left(\omega^T x_i+b\right)\right]-\sum_{i=1}^{n}\mu_i\xi_i$$
$$\text{s.t.}\quad \lambda_i\geq 0\quad \mu_i\geq 0$$

经过转化后得到

$$\max_{\lambda}\left[\sum_{j=1}^{n}\lambda_i-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\lambda_i\lambda_j y_i y_j\left(x_i\cdot x_j\right)\right]$$
$$\text{s.t.}\quad \sum_{i=1}^{n}\lambda_i y_i=0,\quad \lambda_i\geq 0,\quad C-\lambda_i-\mu_i=0$$

核函数

对于数据点完全线性不可分的情况，我们引入核函数 $k(x,y)=(\phi(x),\phi(y))$ 将低维线性不可分的数据点映射到高维空间中，使其变得线性可分。



常见的核函数

- 线性核函数

$$k\left(x_i, x_j\right) = x_i^T x_j$$

- 多项式核函数

$$k\left(x_i, x_j\right) = \left(x_i^T x_j\right)^d$$

- 高斯核函数

$$k\left(x_i, x_j\right) = \exp\left(-\frac{\|x_i - x_j\|}{2\delta^2}\right)$$

- sigmoid 核函数

$$k\left(x_i, x_j\right) = \tanh(\beta x_i^T x_j + c)$$

可以证明，带有 sigmoid 核的 SVM 等价于两层的多层感知机

用 SVM 解决多分类问题

- 一对多法：假设我们要将数据集分为 A、B、C、D 四类，那么我们可以先把其中的一类作为分类 1，其他类统一归为分类 2。这种方法，针对 K 个分类，需要训练 K 个分类器。
    - 分类速度较快，但训练速度较慢
    - 负样本数量远大于正样本数量，会造成样本不对称的情况
    - 如果增加新的分类，需要对分类器重新构造
- 一对一法：在任意两类样本之间构造一个 SVM，这样针对 K 类的样本，就会有 $C(k,2)$ 类分类器。当对一个未知样本进行分类时，每一个分类器都会有一个分类结果，即为 1 票，最终得票最多的类别就是整个未知样本的类别。
    - 如果新增一类，只需要训练和新增这一类样本的分类器
    - 分类器的个数与 K 的平方成正比，当 K 较大时，训练和测试的时间会比较慢

SVM 的优缺点

- 优点：
  - 有严格的数学理论支持，可解释性强
  - 采用核技巧之后，可以处理非线性分类或回归任务
  - 样本量不大时，分类准确率高，泛化能力强
  - 解决高维特征的分类问题和回归问题很有效，在特征维度大于样本数时依然有很好的效果
- 缺点：
  - S 在样本量非常大，核函数映射维度非常高时，计算量过大，不太适合使用
  - 非线性问题的核函数的选择没有通用标准，难以选择一个合适的核函数
  - SVM 对缺失数据敏感

1. SVM
2. SVM 的实现
3. 课堂练习

MATLAB 中关于 SVM 的函数和参数

常用函数：

- `fitcsvm`：训练用于二分类的 SVM 分类器
- `fitcecoc`：训练用于多分类的 SVM 分类器
- `predict`：预测

`fitcsvm` 的常用参数：

- KernelFunction：核函数类型
- BoxConstraint：惩罚参数，控制软间隔
- Standardize：是否对特征进行标准化

分类学习器-SVM

可以使用分类学习器 APP 实现 SVM

利用 sklearn 实现 SVM

- 导入 SVM 工具包：`from sklearn import svm`
- 直接导入 SVM 分类器：`from sklearn.svm import svc`
- 创建 SVM 模型：`model = svc()`
  - 参数 `kernel`：指定核函数类型
  - 参数 `C`：惩罚项系数
  - 参数 `gamma`：高斯核函数系数
- 线性 SVM 分类器：`LinearSVC`
  - 使用线性核函数，不能选择其他核函数
  - 在线性可分的情况下效率更高
- SVC 实现多分类任务
  - `decision function shape='ovo'`：一对一
  - `decision function shape='ovr'`：一对多

1. SVM
2. SVM 的实现
3. 课堂练习

用 SVM 进行乳腺癌检测

数据集 data.csv 来自美国威斯康星州的乳腺癌诊断数据集。医疗人员采集了患者乳腺肿块经过细针穿刺 (FNA) 后的数字化图像，并且对这些数字图像进行了特征提取，这些特征可以描述图像中的细胞核呈现。肿瘤可以分成良性和恶性。

数据表一共包含了 32 个字段，代表的含义保存在字段含义.jpeg 中。mean 代表平均值，se 代表标准差，worst 代表最大值（3 个最大值的平均值）。每张图像都计算了相应的特征，得出了这 30 个特征值（不包括 ID 字段和分类标识结果字段 diagnosis），实际上是 10 个特征值（radius、texture、perimeter、area、smoothness、compactness、concavity、concave points、symmetry 和 fractal dimension mean）的 3 个维度，平均、标准差和最大值。这些特征值都保留了 4 位数字。字段中没有缺失的值。在 569 个患者中，一共有 357 个是良性，212 个是恶性。我们的目标是生成一个乳腺癌诊断的 SVM 分类器，并计算这个分类器的准确率。