机器学习笔记–支撑向量机（SVM）

空修菜

1 支撑向量机（Support Vector Machines）

对于向量 ω 来说，我们可以通过内积得到一个支撑超平面。

$$\omega \cdot x = t$$

其中，$t \in \mathbb{R}$。

这个方程就刻画了一个超平面，该平面以 ω 为外法向量。该方程描述的是：超平面上任何一点在 ω 方向上投影到长度均为 t。

在本节中，我们令 $y \in \{-1, 1\}$,

$$h_{\omega,b}(x) = g(\omega^T x + b) = g(z).$$

若 $z \geq 0$, 则 $g(z) = 1$; 否则,$g(z) = -1$。

1.1 Functional and Geometric margins

1. 超平面 H. 以 α 为单位外法向量且到原点距离为 $-b$ 的支撑超平面 H 定义为:
$$H := \{x \in \mathbb{R}^n : \alpha^T x + b = 0\}.$$
   - 由 H 的定义, 对任意的 $x_0 \in H$, 均有 $\alpha^T x_0 = -b$;
   - 对于 $x^{(i)}$，$\alpha^T x^{(i)} = \ell$ 表示 $x^{(i)}$ 在单位方向 α 上投影的长度是 ℓ;
   - 若 $\ell = -b$，则该点在支撑超平面 H 上;
   - 若该长度 $\ell \neq -b$, 则 $\ell - (-b) = \ell + b = \alpha^T x^{(i)} + b$ 就是点 $x^{(i)}$ 到超平面 H 的距离.

2. 函数间隔（Functional margins）的定义。给定一个训练集 $(x^{(i)}, y^{(i)})$，其函数间隔 $\hat{\gamma}_i$ 定义为
$$\hat{\gamma}_i = y^{(i)}(\omega^T x^{(i)} + b).$$

- 由第一点的超平面定义可知, $\omega^T x^{(i)} + b$ 计算的就是 $x^{(i)}$ 到某个超平面的距离 (正负暂时不确定).
- 若数据集是线性可分的, 则超平面将数据集分为两部分，一部分是正类，另一部分是负类；发方向指向的一侧是正类，另一侧是负类。
- 由第二点, 因为 $y^{(i)} \in \{-1, 1\}$, 所以当 $x^{(i)}$ 被正确分类时, $\hat{\gamma}_i \geq 0$, 即 $\hat{\gamma}_i$ 就是点 $x^{(i)}$ 到某个超平面的距离 (符号确定).
- 给定一个训练集 $S = \{(x^{(i)}, y^{(i)}) \mid i = 1, 2, ..., m\}$. (ω, b) 关于 S 的函数间隔 $\hat{\gamma}$ 是所有函数边界中最小的一个，定义为:

$$\hat{\gamma} = \min_{i=1,2,...,m} \hat{\gamma}_i.$$

3. **函数间隔的问题**. 当成比例增加 ω 和 b 时, 比如, 增大 k 倍, $k\omega, kb$, 此时, 由函数间隔的计算公式可知, 函数间隔 $\hat{\gamma}_i$ 也增加了 k 倍, 但是超平面却没有移动, 因为 $k\omega^T x + kb = 0$, 所以 $\omega^T x + b = 0$.

4. 几何间隔 (Geometric margins). 将 $x^{(i)}$ 的函数间隔除以外法方向向量 ω 的长度 $\|\omega\|$ 就得到了 $x^{(i)}$ 的几何边界 γ_i (外法向单位化):

$$\gamma_i = \frac{\hat{\gamma}_i}{\|\omega\|} = \frac{y^{(i)}(\omega^T x^{(i)} + b)}{\|\omega\|} = y^{(i)}\Big(\Big(\frac{\omega}{\|\omega\|}\Big)^T x^{(i)} + \frac{b}{\|\omega\|}\Big).$$

- 当 ω 是单位向量时, $\hat{\gamma}_i = \gamma_i$;
- 当 ω, b 成比例变化时, 超平面位置不变、点 $x^{(i)}$ 的函数间隔不变.
- 集合 S 的几何边界 γ 定义为:

$$\gamma = \min_{i=1,2,...,m} \gamma_i.$$

1.2 几何间隔最大化问题

1. 第一个优化问题:

$$\begin{aligned} &\max_{\gamma,\omega,b} \gamma \\ &\quad s.t. \quad y^{(i)}(\omega^T x^{(i)} + b) \geq \gamma, i = 1, 2, ..., m \\ &\quad \|\omega\| = 1 \end{aligned}$$

- 对于数据集 S 中的每一个元素, 该元素的函数边界和几何边界最少为 γ;

- $\|\omega\| = 1$ 为的是保证函数边界与几何边界相等

2. 因为 $\gamma = \frac{\hat{\gamma}}{\|\omega\|}$, 可以将优化问题进一步简化为优化问题二:

$$\max_{\hat{\gamma},\omega,b} \frac{\hat{\gamma}}{\|\omega\|}$$
$$s.t. \quad \frac{y^{(i)}(\omega^T x^{(i)} + b)}{\|\omega\|} \geq \frac{\hat{\gamma}}{\|\omega\|}, i = 1, 2, ..., m$$

- 由几何边界关于 (ω, b) 的膨胀不变性，我们将 (ω, b) 膨胀为原来的 $\hat{\gamma}$ 倍，即 $\hat{\gamma}(\omega, b)$。那么上面的优化问题就可以进一步整理为:

$$\max_{\hat{\gamma},\omega,b} \frac{1}{\|\omega\|}$$
$$s.t. \quad y^{(i)}(\omega^T x^{(i)} + b) \geq 1 \quad , i = 1, 2, ..., m$$

3. 因为求 $1/\|\omega\|$ 的最大值等价于求 $\frac{1}{2}\|\omega\|^2$ 的最小值，所以我们就得到了最终的优化问题三:

$$\min_{\gamma,\omega,b} \frac{1}{2}\|\omega\|^2$$
$$s.t. \quad y^{(i)}(\omega^T x^{(i)} + b) \geq 1 \quad , i = 1, 2, ..., m$$

1.3 拉格朗日对偶（Lagrange Dual）

一般上面所得到的最后一个优化问题是通过对偶 Lagrange 问题来解决的.

一、原始问题 (Primal Preblem)

假设现在要解决的优化问题是

$$\begin{aligned} \min_x \quad & f(x) \\ s.t. \quad & g_i(x) \leq 0, \quad i = 1, \cdots k, \\ & h_i(x) = 0, \quad i = 1, \cdots l \end{aligned}$$

为此, 先构造一般的 Lagrange 函数:

$$\mathcal{L}(x, \alpha, \beta) = f(x) + \sum_{i=1}^{k} \alpha_i g_i(x) + \sum_{i=1}^{l} \beta_i h_i(x).$$

固定 x, 令

$$\theta_{\mathcal{P}}(x) = \max_{\alpha,\beta} \mathcal{L}(x,\alpha,\beta),$$

在上式中, 若 g_i 或 h_i 中某个不符合约束条件, 即 $g_i > 0$ 或 $h_i \neq 0$, 设 $\alpha_i \geq 0$, 那么立即可得

$$\theta_{\mathcal{P}}(x) = \max_{\alpha,\beta} \mathcal{L}(x,\alpha,\beta) = \infty.$$

因此, 在上述假设下就有

$$\sum_{i=1}^{k} \alpha_i g_i(x) + \sum_{i=1}^{l} \beta_i h_i(x) \leq 0,$$

所以,

$$\theta_{\mathcal{P}}(x) = f(x),$$

由此可知, 在符合约束条件时, 有

$$\min_{x}\ f(x) = \min_{x}\ \theta_{\mathcal{P}}(x) = \min_{x} \max_{\alpha,\beta} \mathcal{L}(x,\alpha,\beta).$$

设 $p^* = \min_x\ f(x)$, 则 $p^* = \min_x\ \theta_{\mathcal{P}}(x)$, p^* 称为原始问题 (primal problem) 的值.

二、Lagrange Dual Problem

上面的原始问题先求解最大值，然后再求解最小值得到最后的结果. 对偶 Lagrange 问题在顺序上则刚好相反. 定义函数 $\theta_{\mathcal{D}}$ 为:

$$\theta_{\mathcal{D}}(\alpha,\beta) = \min_{x} \mathcal{L}(x,\alpha,\beta),$$

其中, $\alpha_i \geq$). 对偶 (dual) 优化问题就是:

$$\max_{\alpha,\beta} \theta_{\mathcal{D}}(\alpha,\beta) = \max_{\alpha,\beta} \min_{x} \mathcal{L}(x,\alpha,\beta),$$

同时, 定义对偶优化问题的优化值为 d^*:

$$d^* = \max_{\alpha,\beta} \theta_{\mathcal{D}}(\alpha,\beta).$$

三、p^* 和 d^* 的关系

实际上, 就是通过 p^* 和 d^* 的关系来解决问题. 两者具有若对偶性 (weak duality) 以及强对偶性 (strong duality) 两种关系.

1. 弱对偶性 (weak duality)

   Proposition 1.1.

   $$d^* = \max_{\alpha,\beta} \min_x \mathcal{L}(x,\alpha,\beta) \le \min_x \max_{\alpha,\beta} \mathcal{L}(x,\alpha,\beta) = p^*.$$

   Proof. 由定义

   $$\begin{aligned} \mathcal{L}(x,\alpha,\beta) &\le \max_{\alpha,\beta} \mathcal{L}(x,\alpha,\beta) = \mathcal{L}_1(x) \\ \mathcal{L}_2(\alpha,\beta) &= \min_x \mathcal{L}(x,\alpha,\beta) \le \mathcal{L}(x,\alpha,\beta), \end{aligned}$$

   所以,

   $$\mathcal{L}_2(\alpha,\beta) \le \mathcal{L}_1(x),$$

   由 (x,α,β) 的任意性, 有

   $$\max_{\alpha,\beta} \mathcal{L}_2(\alpha,\beta) \le \min_x \mathcal{L}_1(x)$$

   □

2. 强对偶性 (strong duality)

   强对偶性即,$p^* = d^*$. 对于对偶优化问题来说, 强对偶性总是成立的.

   强对偶性的证明依赖的假设有:1,$f(x), g_i(x)$ 是凸的, h_i 是线性的以及至少存在一个点 x_0, 使得 $g_i(x_0) < 0$, $h_i(x_0) = 0$. 上面这些假设, 统称为 Slater's condition.

在上面的假设下, 可以求得 (x^*, α^*, β^*), 其中, x^* 是原始问题的解, α^*, β^* 是对偶优化问题的解. 此外, x^*, α^*, β^* 还满足 **KKT** 条件:

$$\begin{aligned} \frac{\partial \mathcal{L}(x^*,\alpha^*,\beta^*)}{\partial x_i} &= 0, \quad i = 1,2,\cdots,n \\ \frac{\partial \mathcal{L}(x^*,\alpha^*,\beta^*)}{\partial \beta_i} &= 0, \quad i = 1,2,\cdots,l \\ \alpha_i g_i(x_i) &= 0, \quad i = 1,2,\cdots,k \\ g_i(x_i) &\le 0, \quad i = 1,2,\cdots,k \\ \alpha_i &\ge 0, \quad i = 1,2,\cdots,k \end{aligned}$$

若任意的 x, α, β 满足 KKT 条件, 那么它们也是原始问题和对偶优化问题的解.

1.4 优化边界分类（Optimal margin classifiers）

对于 $i=1,2,...,m$, 原始优化问题是:

$$\min_{\gamma,w,b} \frac{1}{2}\|\omega\|^2$$
$$s.t.\ y^{(i)}(\omega^T x^{(i)}+b)\geq 1$$

令约束条件为 g_i:

$$g_i(\omega)=1-y^{(i)}(\omega^T x^{(i)}+b)\leq 0.$$

g_i 度量的是每一个 sample x_i 到支撑平面的距离. 当 x_i 到支撑平面的距离大于 1 时, 由 KKT 对偶补充条件 $\alpha_i g_i(\omega)=0$ 可知, 此时对应的 $\alpha_i=0$. 当 x_i 到支撑平面的距离刚好为 1 时, $g_i=0$, 此时, $\alpha_i>0$. 当 $g_i(\omega)=0$ 时, $\alpha_i>0$ 所对应的向量称为支撑向量。(α_i 并不会恒为 0)

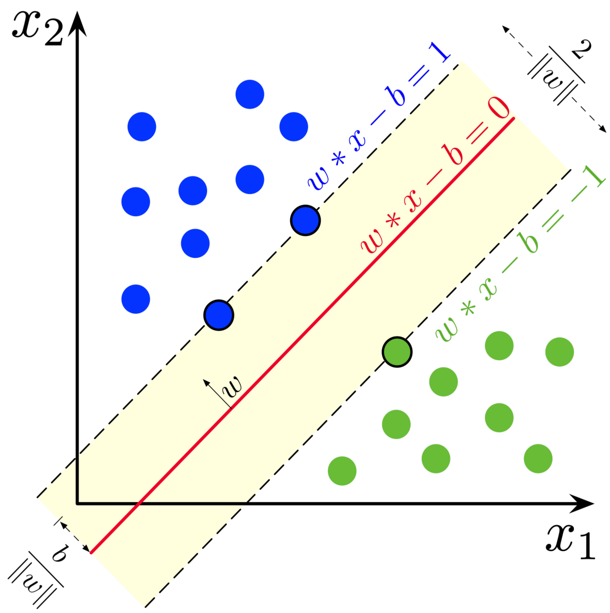

Figure 1.1: 虚线上的三个向量即为支撑向量

由前一节的 Lagrange Dual 理论知, 要求解上述原始优化问题, 可以求解与之等价的对偶优化问题, 所以, 解决步骤分为两步. 1, 先对 $\mathcal{L}$ 关于 w,b 求极小值; 2, 再对 $\mathcal{L}$ 关于 α 求极大值. 所得解就是原始问题的解.

$$\mathcal{L}(\omega,b,\alpha)=\frac{1}{2}\|\omega\|^2-\sum_{i=1}^{m}\alpha_i\Big(y^{(i)}(\omega^T x^{(i)}+b)-1\Big). \tag{1.1}$$

一、Lagrange 函数的关于 w, b 的最小化

对于固定的 α, 考虑 $\mathcal{L}$ 的最小化问题. 我们先将 $\mathcal{L}$ 写为:

$$\mathcal{L}(\omega, b, \alpha) = \frac{1}{2}\sum_{i=1}^{n}\omega_i^2 - \sum_{i=1}^{m}\alpha_i y^{(i)}\Big(\sum_{j=1}^{n}\omega_i x_j^{(i)}\Big) - b\sum_{i=1}^{m}\alpha_i y^{(i)} + \sum_{i=1}^{m}\alpha_i. \quad (1.2)$$

事实上, 容易证明 $\mathcal{L}(w, \alpha, b)$ 关于 w, b 是凸的, 所以可以通过分别对 w, b 求偏导的方式来求取 $\mathcal{L}$ 的最小值. 已知

$$\frac{\partial \mathcal{L}}{\partial \omega} = \Big(\frac{\partial \mathcal{L}}{\partial \omega_1}, \frac{\partial \mathcal{L}}{\partial \omega_2}, ..., \frac{\partial \mathcal{L}}{\partial \omega_n}\Big).$$

由求导法则，我们有

$$\frac{\partial \mathcal{L}}{\partial \omega_k} = \omega_k - \sum_{i=1}^{m}\alpha_i y^{(i)} x_k^{(i)}.$$

所以

$$\begin{aligned}\frac{\partial \mathcal{L}}{\partial \omega} &= (\omega_1, \omega_2, ..., \omega_n) - \Big(\sum_{i=1}^{m}\alpha_i y^{(i)} x_1^{(i)}, \sum_{i=1}^{m}\alpha_i y^{(i)} x_2^{(i)}, ..., \sum_{i=1}^{m}\alpha_i y^{(i)} x_n^{(i)}\Big)\\ &= \omega - \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}.\end{aligned}$$

同理,

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{i=1}^{m}\alpha_i y^{(i)}.$$

分别令 $\frac{\partial \mathcal{L}}{\partial \omega} = 0, \frac{\partial \mathcal{L}}{\partial b} = 0$, 则有

$$\omega = \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}, \qquad \sum_{i=1}^{m}\alpha_i y^{(i)} = 0.$$

下面是一个简单结论，很好证明。

Proposition 1.2. $y, x_i \in \mathbb{R}^n, i = 1, 2, ..., k,$

$$x = \sum_{i=1}^{k} x_i,$$

则

$$y \cdot x = \langle y, \sum_{i=1}^{k} x_i\rangle = \sum_{i=1}^{k} y \cdot x_i.$$

Proof.

$$\begin{aligned}
y\cdot x &= (y_1,...,y_n)\cdot\Big(\sum_{i=1}^{k}x_i^1,...,\sum_{i=1}^{k}x_i^n\Big)\\
&=\sum_{j=1}^{n}y_j\Big(\sum_{i=1}^{k}x_i^j\Big)=\sum_{j=1}^{n}\sum_{i=1}^{k}y_jx_i^j\\
&=\sum_{i=1}^{k}\sum_{j=1}^{n}y_jx_i^j=\sum_{i=1}^{k}y\cdot x_i
\end{aligned}$$

□

由 Proposition 1.2, 我们得到

$$\begin{aligned}
\|\omega\|^2=\omega^T\omega=\sum_{i=1}^{m}\alpha_iy^{(i)}\omega^Tx^{(i)}&=\sum_{i=1}^{m}\alpha_iy^{(i)}\Big(\sum_{j=1}^{m}\alpha_jy^{(j)}(x^{(j)})^Tx^{(i)}\Big)\\
&=\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy^{(i)}y^{(j)}(x^{(j)})^Tx^{(i)}.
\end{aligned}$$

再将所得结果代入到 (1.4) 式, 我们就得到下面的表达式:

$$\mathcal{L}(\omega,b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy^{(i)}y^{(j)}(x^{(j)})^Tx^{(i)}.$$

观察上式可知，其中只有 α 是未确定的, 也就是说, 在 α 确定以后，$\mathcal{L}$ 的值将由数据集 S 给出.

二、Lagrange 函数关于 α 最大化

在将上面得到的 Lagrange 函数 $\mathcal{L}$ 关于 α 最大化, 所得的解就是原始问题的解.

$$\begin{aligned}
\max_{\alpha}\quad & W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_jy^{(i)}y^{(j)}(x^{(j)})^Tx^{(i)}\\
s.t.\quad & \alpha_i\geq 0, i=1,2,...,m\\
& \sum_{i=1}^{m}\alpha_iy^{(i)}=0
\end{aligned}$$

关于预测，主要通过计算 $z = \omega^T x + b$ 来进行，若 z 远大于 0，则取 1；否则取-1。同时借助于前面的讨论，我们还可以经预测问题归结为（在 α 已知时）计算数据集 S 与 x 的内积：

$$\begin{aligned}\omega^T x + b &= \Big(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\Big)^T x + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} (x^{(i)})^T x + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b\end{aligned}$$

1.5 核（Kernels）

在前面关于 SVM 的讨论中，均假设数据集是线性可分离的，即认为数据集可以用一个长平面将它们分开。但在实际处理问题时，常常不是线性可分离的，这就涉及到一种新的方法: 核 (kernels)。它的作用就是将低维度的数据转换为高维的进行处理。例如:ϕ 是一个 feature map，$\phi : \mathbb{R} \mapsto \mathbb{R}^3$

$$\phi(x) = \begin{bmatrix} x \\ x^2 \\ x^3 \end{bmatrix}$$

这样，ϕ 就把一个一维的数映成了一个三维的向量。在给定一个特征映射 ϕ 以后，我们将与之对应的核 $K(x,z)$ 定义为：

$$K(x,z) = \phi(x)^T \phi(z).$$

实际上,如果使用内积的记号,我们就可以将上式写为 $K(x,z) = \langle \phi(x), \phi(z) \rangle$. 此外，核的一个优点是它比较容易计算。

- 核矩阵

  假设 K 是与某个特征映射 ϕ 相对应的核。有 m 个点 $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$。那么我们定义一个 $m \times m$ 的矩阵 K, 其中 K 的第 i 行第 j 列的元素 K_{ij} 由核来给出:

  $$K_{ij} = K(x^{(i)}, x^{(j)}),$$

  这样得到的矩阵 K 就称为核矩阵。注意到，核矩阵 K 是对称阵，

  $$K_{ij} = K(x^{(i)}, x^{(j)}) = \phi(x^{(i)})^T \phi(x^{(j)}) = \phi(x^{(j)})^T \phi(x^{(i)}) = K(x^{(j)}, x^{(i)}) = K_{ji}.$$

  同时，可以证明核矩阵 K 半正定的，这样我们就得到下面的一个定理。

Theorem 1.1. *Let $K : \mathbb{R}^n \times \mathbb{R}^n \mapsto \mathbb{R}$ be given. Then for K to be a valid (Mercer) kernel, it is necessary and sufficient that for any $\{x^{(1)}, x^{(2)}, ..., x^{(m)}\}$, $(m < \infty)$, the corresponding kernel matrix is symmetric positive semi-definite.*

上面的都是在线性可分离的情形下讨论的,下面讨论不是线性可分离的情形。

常见的核函数

1. 多项式核函数. $x, z \in \mathbb{R}^n$, p 是正整数:

$$K(x, z) = (x \cdot z + 1)^p.$$

2. 高斯核函数:

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

1.6 规则化以及不可分离的情形（Regularization and the non-separable case）

数据集中存在一些点，这些点到超平面的距离小于 1，即 $\omega^T x + b < 1$。如果按照之前的优化问题的要求，那么这些点都属于特殊情况（outlier）。所以需要对已经得到的模型进行优化，这样：

$$\begin{aligned} \min_{\xi, \omega, b} \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{m} \xi_i \\ s.t. \quad & y^{(i)}(\omega^T x^{(i)} + b) \geq 1 - \xi_i \\ & \xi_i \geq 0, i = 1, 2, ..., m \end{aligned}$$

上式中, $C > 0$ 是惩罚参数. 如果 C 比较小, 这就表明, 可以接受相对多一些的点成为 outlier,(总体的最小值不会增加过快), 所以最后的决策“带”就会比较宽, 也就是会有正 sample 也会有负 sample 落在“带”中; 若 C 值较大, 则每产生一个 outlier, 最小值增加的幅度会很大, 这就使得决策“带”会比较宅, 只允许很少的 sample 落在“带”中.

将上面的优化问题写成 Lagrange 函数 $\mathcal{L}$:

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_i - \sum_{i=1}^{m} \alpha_i \Big(y^{(i)}(w^T x^{(i)} + b) + \xi_i - 1\Big) - \sum_{i=1}^{m} \mu_i \xi_i,$$

仿照前面求关于 w, b 的导数的过程, 有

$$\frac{\partial \mathcal{L}}{\partial w} = w - \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$$

$$\frac{\partial \mathcal{L}}{\partial b} = -\sum_{i=1}^{m} \alpha_i y^{(i)}$$

$$\frac{\partial \mathcal{L}}{\partial \xi_i} = C - \alpha_i - \mu_i$$

分别令其为 0, 再将所得关系带入 $\mathcal{L}(w, b, \xi, \alpha, \mu)$, 就得

$$\mathcal{L}(w, b, \xi, \alpha, \mu) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(i)}.$$

观察式子, 所含参数只剩下 α, 由对偶 Lagrange 优化问题, 有

$$\begin{aligned}
&\max_{\alpha} W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(i)} \\
&s.t.\ \sum_{i=1}^{m} \alpha_i y^{(i)} = 0, \quad i = 1, 2, \cdots, m \\
&\quad\quad C = \alpha_i + \mu_i \\
&\quad\quad \alpha_i \geq 0 \\
&\quad\quad \mu_i \geq 0
\end{aligned}$$

由于上面的 objective function 已经没有参数 μ 了, 考虑到它们之间的关系, 有

$$0 \leq \alpha_i = C - \mu_i \leq C,$$

最后得到的优化问题是:

$$\begin{aligned}
\max_{\alpha} \quad & W(\alpha) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(j)})^T x^{(i)} \\
s.t. \quad & 0 \leq \alpha_i \leq C, i = 1, 2, ..., m \\
& \sum_{i=1}^{m} \alpha_i y^{(i)} = 0
\end{aligned}$$

最后是实现上述优化问题的算法：序列最小优化（sequential minimal optimization）。

1.7 SMO 算法（The SMO algorithm）

SMO 算法就是选取两个 α_i, α_j 为变量，把其他 $\alpha_k, (k \neq i, j)$ 视作常量来进行求解。我们选取 α_1, α_2 作为变量来对问题进行求解。为了方便刻画，我们将 $\langle x^{(i)}, x^{(j)} \rangle$ 记为 A_{ij}, 当我们把所有包含 α_1, α_2 的项写出来之后，剩余不包含 α_1, α_2 的项就成了一个常数项 k:

$$k = -\frac{1}{2} \sum_{i=3}^{m} \sum_{j=3}^{m} \alpha_i \alpha_j y^{(i)} y^{(j)} A_{ij}.$$

而常数项对于 $W(\alpha_1, \alpha_2)$ 的优化是没有影响的，所以将 k 略去。那么上面的优化问题就可以写为:

$$\begin{aligned}
\max_{\alpha_1, \alpha_2} \quad W(\alpha_1, \alpha_2) =& (\alpha_1 + \alpha_2) - \frac{1}{2}\alpha_1^2 A_{11} - \frac{1}{2}\alpha_2^2 A_{22} - \alpha_1 \alpha_2 y^{(1)} y^{(2)} A_{12} \\
& - \alpha_1 y^{(1)} \sum_{i=3}^{m} \alpha_j y^{(j)} A_{1j} - \alpha_2 y^{(2)} \sum_{i=3}^{m} \alpha_j y^{(j)} A_{2j} \\
s.t. \quad & 0 \leq \alpha_i \leq C, \quad i = 1, 2 \\
& \alpha_1 y^{(1)} + \alpha_2 y^{(2)} = -\sum_{i=3}^{m} \alpha_i y^{(i)} = \xi
\end{aligned}$$

注意到

$$\alpha_1 y^{(1)} + \alpha_2 y^{(2)} = \xi \tag{1.3}$$

所以借助这个约束关系，我们可以将 α_2 作为变量，实际上，这就变成了一个单变量问题。

假设我们关于 W 得到初始解为 $\alpha_1^{old}, \alpha_2^{old}$, 最优解为 $\alpha_1^{new}, \alpha_2^{new}$。最优解正是我们要求的。由于是初始解，所以必定满足优化问题的约束条件，即

$$\alpha_1^{old} y^{(1)} + \alpha_2^{old} y^{(2)} = \xi \tag{1.4}$$

由于 $y^{(i)} \in \{-1, 1\}$, 所以从 (1.3) 式出发，我们就可以得到以下的四种情况:

(a). 对于 $y^{(1)} \neq y^{(2)}$.

- $y^{(1)} = 1, y^{(2)} = -1$
- $y^{(1)} = -1, y^{(2)} = 1$

(b). 对于 $y^{(1)} = y^{(2)}$.

- $y^{(1)} = y^{(2)} = 1$

- $y^{(1)} = y^{(2)} = -1$

- 情形一. 由 $y^{(i)}, i = 1, 2$ 的取值，有

$$\alpha_2 = \alpha_1 - \xi. \tag{1.5}$$

已知 $0 \le \alpha_1 \le C$。若 $\xi > 0$, 则由 (1.5) 式可知，α_1 不能取 0(若取 0，则会导致 $\alpha_2 < 0$)，同时，α_1 可以取得最大值 C。此时，由 (1.4) 式，我们可以得到 ξ 的表达式，$\xi = \alpha_1^{old} - \alpha_2^{old}$。所以，

$$0 \le \alpha_2 = \alpha_1 - \xi \le C - \xi = C + \alpha_2^{old} - \alpha_1^{old}.$$

若 $\xi < 0$, 则 α_2 的最大值只能为 C, 而 α_2 的最小值只在 $\alpha_1 = 0$ 的时候取到，即 $\alpha_2 = -\xi$。所以

$$\alpha_2^{old} - \alpha_1^{old} \le \alpha_2 \le C.$$

从上面的讨论，我们得到了两个 α_2 的区间，我们把它写为一种形式。经观察，可以令

$$L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\}$$
$$H = \min\{C, C + \alpha_2^{old} - \alpha_1^{old}\}$$

所以，可以得到

$$L \le \alpha_2 \le H.$$

- 情形二. 根据 $y^{(i)}, i = 1, 2$ 的取值，有

$$\alpha_2 = \alpha_1 + \xi. \tag{1.6}$$

已知 $0 \le \alpha_1 \le C$。若 $\xi > 0$, 则由 (1.6) 式可知，α_1 不能取 C(若取 C，则会导致 $\alpha_2 > C$)，同时，α_1 可以取得最小值 0。所以，α_2 的最小值为 ξ, 最大值为 C。此时，由 (1.4) 式，我们可以得到 ξ 的表达式，$\xi = \alpha_2^{old} - \alpha_1^{old}$。所以，

$$\alpha_2^{old} - \alpha_1^{old} \le \alpha_2 \le C.$$

若 $\xi < 0$, 则 α_2 的最小值为 0，最大值为 $C + \xi$。所以

$$0 \le \alpha_2 \le C + \alpha_2^{old} - \alpha_1^{old}.$$

关于情形二，仿照前面的讨论，我们也可以得到

$$L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\}$$
$$H = \min\{C, C + \alpha_2^{old} - \alpha_1^{old}\}$$

所以，情形二也可以利用与情形一完全相同的符号写为:

$$L \le \alpha_2 \le H.$$

由关于情形一、二的讨论，在优化问题 W 的约束条件之下，我们得到 α_2 的取值范围:

$$L \le \alpha_2 \le H,$$

其中，$L = \max\{0, \alpha_2^{old} - \alpha_1^{old}\}, H = \min\{C, C + \alpha_2^{old} - \alpha_1^{old}\}$.

关于情形三、四，思路与情形一、二是类似的。

$$L = \max\{0, \alpha_2^{old} + \alpha_1^{old} - C\} \quad H = \min\{C, \alpha_2^{old} + \alpha_1^{old}\}$$

如前所述，由式子 (1.3), 我们可以将两变量问题，变为单变量优化问题，我们以 α_2 为自变量。$\alpha_1 = y^{(1)}\xi - \alpha_2 y^{(1)} y^{(2)}$。并且令

$$B_1 = \sum_{i=3}^{m} \alpha_j y^{(j)} A_{1j}$$
$$B_2 = \sum_{i=3}^{m} \alpha_j y^{(j)} A_{2j}$$

则目标函数 W 可以写为:

$$\begin{aligned} W(\alpha_2) =& \alpha_2^2(-\frac{1}{2}A_{11} - \frac{1}{2}A_{22} + A_{12}) \\ &+ \alpha_2(1 - y^{(1)}y^{(2)} + \xi y^{(2)} A_{11} - \xi y^{(2)} A_{12} + y^{(2)} B_1 - y^{(2)} B_2) \\ &+ (y^{(1)}\xi - \frac{1}{2}\xi^2 A_{11} - \xi B_1). \end{aligned}$$

注意到，此时我们的初始解满足 $\alpha_1^{old} y^{(1)} + \alpha_2^{old} y^{(2)} = \xi$，所以 ξ 是一个确定的常数，在对 α_2 求导数时，ξ 就作为一个常数来考虑，而不是作为 α_2 的函数。此时，对 $W(\alpha_2)$ 求导，并令其为 0:

$$\begin{aligned} 0 = \frac{\partial W}{\partial \alpha_2} =& 2\alpha_2(-\frac{1}{2}A_{11} - \frac{1}{2}A_{22} + A_{12}) \\ &+ (1 - y^{(1)}y^{(2)} + \xi y^{(2)} A_{11} - \xi y^{(2)} A_{12} + y^{(2)} B_1 - y^{(2)} B_2). \end{aligned}$$

所以，我们得到

$$
\begin{aligned}
\alpha_2(A_{11}+A_{22}-2A_{12}) &= 1-y^{(1)}y^{(2)}+\xi y^{(2)}A_{11}-\xi y^{(2)}A_{12}+y^{(2)}B_1-y^{(2)}B_2 \\
&= y^{(2)}\Big(y^{(2)}-y^{(1)}+\xi A_{11}-\xi A_{12}+B_1-B_2\Big) \\
&= y^{(2)}\underbrace{\Big(y^{(2)}-y^{(1)}+(\alpha_1^{old}y^{(1)}+\alpha_2^{old}y^{(2)})(A_{11}-A_{12})+B_1-B_2\Big)}_{B}.
\end{aligned}
$$

为了简化起见，我们将上面等式右端的 $(A_{11}-A_{12})$ 凑成 $(A_{11}+A_{22}-2A_{12})$。

$$
\begin{aligned}
B &= y^{(2)}-y^{(1)}+\alpha_1^{old}y^{(1)}(A_{11}-A_{12})+\alpha_2^{old}y^{(2)}(A_{11}-A_{12})+B_1-B_2 \\
&= y^{(2)}-y^{(1)}+\alpha_1^{old}y^{(1)}(A_{11}-A_{12})+\alpha_2^{old}y^{(2)}(A_{11}+A_{22}-2A_{12}) \\
&+\alpha_2^{old}(A_{12}-A_{22})+B_1-B_2 \\
&= \alpha_2^{old}y^{(2)}(A_{11}+A_{22}-2A_{12})+A
\end{aligned}
$$

由前面关于 B_1, B_2 的记号，我们得到

$$
\begin{aligned}
A &= y^{(2)}-y^{(1)}+\alpha_1^{old}y^{(1)}(A_{11}-A_{12})+\alpha_2^{old}(A_{12}-A_{22})+B_1-B_2 \\
&= y^{(2)}-y^{(1)}+\sum_{j=1}^{m}\alpha_j y^{(j)}A_{1j}-\sum_{j=1}^{m}\alpha_j y^{(j)}A_{2j}
\end{aligned}
$$

为了简便，记

$$
\begin{aligned}
E(x^{(1)}) &= \sum_{j=1}^{m}\alpha_j y^{(j)}A_{1j}-y^{(1)}, \\
E(x^{(2)}) &= \sum_{j=1}^{m}\alpha_j y^{(j)}A_{2j}-y^{(2)}.
\end{aligned}
$$

有了以上的记号，我们就得到

$$
\alpha_2(A_{11}+A_{22}-2A_{12}) = y^{(2)}\Big(\alpha_2^{old}y^{(2)}(A_{11}+A_{22}-2A_{12})+E(x^{(1)})-E(x^{(2)})\Big).
$$

所以我们就得到了 α_2^{new},

$$
\begin{aligned}
\alpha_2^{new} &= \frac{\alpha_2^{old}(A_{11}+A_{22}-2A_{12}+y^{(2)}\big(E(x^{(1)})-E(x^{(2)})\big)}{A_{11}+A_{22}-2A_{12}}. \\
&= \alpha_2^{old}+\frac{y^{(2)}\big(E(x^{(1)})-E(x^{(2)})\big)}{A_{11}+A_{22}-2A_{12}}.
\end{aligned}
$$

由约束条件 $\alpha_1^{new}y^{(1)}+\alpha_2^{new}y^{(2)}=\xi$, 以及 $\alpha_1^{old}y^{(1)}+\alpha_2^{old}y^{(2)}=\xi$，可得

$$\begin{aligned}\alpha_1^{new} &= \xi y^{(1)} - \alpha_2^{new}y^{(1)}y^{(2)} \\ &= (\alpha_1^{old}y^{(1)}+\alpha_2^{old}y^{(2)})y^{(1)} - \alpha_2^{new}y^{(1)}y^{(2)} \\ &= \alpha_1^{old} + y^{(1)}y^{(2)}\left(\alpha_2^{old}-\alpha_2^{new}\right)\end{aligned}$$

到这里，我们就得到最优解 $(\alpha_1^{new},\alpha_2^{new})$, 但是此时得到的解只满足约束条件 (1.3), 结合前面的结果 $L\le\alpha_2^{new}\le H$, 我们得到

$$\alpha_2^{new}=\begin{cases} L & \alpha_2^{new}<L \\ \alpha_2^{new} & L\le\alpha_2^{new}\le H \\ H & \alpha_2^{new}>H\end{cases}$$

关于该问题，利用 SMO 方法求解完毕。

1.8 Summary of SVM

1. 最后需要优化求解的模型是:

$$\begin{aligned}\max_{\alpha}\quad & W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}(x^{(j)})^T x^{(i)} \\ s.t.\quad & 0\le\alpha_i\le C, i=1,2,...,m \\ & \sum_{i=1}^{m}\alpha_i y^{(i)}=0\end{aligned}$$

2. SVM 的预测函数是:

$$f(x)=\mathrm{sign}(w^Tx+b),$$

   且

$$\omega^Tx+b=\sum_{i=1}^{m}\alpha_i y^{(i)}\langle x^{(i)},x\rangle+b$$

   确定了 α 和 b 就得到了模型.

3. 求解 α 的算法是序列最小优化算法 SMO. 挑选 α_i 的标准是 KKT 条件, 即选择不符合 KKT 条件的 α_i, KKT 条件是:

$$\begin{aligned}\alpha_i=0 &\iff y_i g(x_i)\ge 1 \\ 0<\alpha_i<C &\iff y_i g(x_i)=1 \\ \alpha_i=C &\iff y_i g(x_i)\le 1\end{aligned}$$

上面的第一个和第二个等式可以合并为

$$\alpha_i < C \iff y_i g(x_i) \geq 1,$$

第二、三个等式可以合并为

$$\alpha_i > 0 \iff y_i g(x_i) \leq 1,$$

所以, KKT 条件可以写为

$$\begin{aligned} \alpha_i < C &\iff y_i g(x_i) \geq 1 \\ \alpha_i > 0 &\iff y_i g(x_i) \leq 1 \end{aligned}$$

4. 距离分离超平面的长度为 1 的向量是支撑向量, 由 KKT 条件:

$$0 < \alpha_i < C \iff y_i g(x_i) = 1,$$

若 $\alpha_i > 0$, 那么 α_i 对应的向量 x_i 就是支撑向量.

5. α 在初始化为 0 向量, 是为了保证条件:

$$\sum_{i=1}^{m} \alpha_i y^{(i)} = 0.$$

6. 关于 KKT 条件的误差 ε, 一般取 10^{-3}. 更新的 α^{new} 与原来的 α^{old} 相差较大, 即 $| \alpha^{new} - \alpha^{old} | < 10^{-5}$ 时, 才更新 α^{new}.

7. 在不可分离情形中, $C > 0$ 是罚项. 即对原模型的惩罚:

   - 若 C 比较小, 这就表明, 可以接受相对多一些的点成为 outlier,(总体的最小值不会增加过快), 所以最后的决策“带”就会比较宽, 也就是会有正 sample 也会有负 sample 落在“带”中;
   - 若 C 值较大, 则每产生一个 outlier, 最小值增加的幅度会很大, 这就使得决策“带”会比较窄, 只允许很少的 sample 落在“带”中.