FDU 统计机器学习 4. 支持向量机

本文参考以下教材:

- 机器学习 (周志华) 第 6 章
- 统计学习方法 (第2版, 李航) 第 7 章
- 神经网络与深度学习 (邱锡鹏) 第 3 章

欢迎批评指正!

4.1 线性支持向量机

4.1.1 问题描述

支持向量机 (Support Vector Machine, SVM) 是一个经典的二分类算法.
给定一个二分类训练集 $D_{\text{train}} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$，其中 $y^{(i)} \in \{-1, 1\}$ 而 $x^{(i)} \in \mathbb{R}^d$
若存在一个超平面 $w^{\mathrm{T}}x + b = 0$ 能够将两类样本分开 (即 $y^{(i)}(w^{\mathrm{T}}x^{(i)} + b) > 0 \ (i = 1, \dots, n)$),
则我们称这两类样本**线性可分** (linearly separable)，称 $w^{\mathrm{T}}x + b = 0$ 为**分割超平面**.

训练集 $D_{\text{train}} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$ 中每个点到分割超平面 $w^{\mathrm{T}}x + b = 0$ 的距离为:

$$d^{(i)} = \frac{|w^{\mathrm{T}}x^{(i)} + b|}{\|w\|_2} = \frac{y^{(i)}(w^{\mathrm{T}}x^{(i)} + b)}{\|w\|_2} \ (i = 1, \dots, n)$$

我们称 $y^{(i)}(w^{\mathrm{T}}x^{(i)} + b)$ 为样本点 $(x^{(i)}, y^{(i)})$ 被分量正确的置信度 (又称**分类得分**)
我们定义**间隔** (margin) d 为训练集 $D_{\text{train}} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$ 中的样本到分割超平面的最短距离:

$$d := \min_{1 \le i \le n} d^{(i)} = \min_{1 \le i \le n} \frac{y^{(i)}(w^{\mathrm{T}}x^{(i)} + b)}{\|w\|_2}$$

间隔 d 越大，则对应的分割超平面对两类样本的划分越稳定 (即不容易受噪声等因素的影响)
因此支持向量机的目的是寻找最优的分割超平面 $(w_\star, b_\star)$ 最大化间隔 d:

$$\begin{aligned} \max_{w,b} \quad & d \\ \text{s.t.} \quad & \frac{y^{(i)}(w^{\mathrm{T}}x^{(i)} + b)}{\|w\|_2} \ge d \ (i = 1, \dots, n) \end{aligned}$$

不失一般性，我们可以限制 w 满足 $\|w\|_2 \cdot d = 1$，则上述优化问题等价于:

$$\begin{aligned} \max_{w,b} \quad & \frac{1}{\|w\|_2} \\ \text{s.t.} \quad & y^{(i)}(w^{\mathrm{T}}x^{(i)} + b) \ge 1 \ (i = 1, \dots, n) \end{aligned}$$

进而写成标准形式的**凸优化问题** (具体来说是**二次规划问题**):

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1 - y^{(i)}(w^{\mathrm{T}}x^{(i)} + b) \le 0 \ (i = 1, \dots, n) \end{aligned}$$

其 KKT 点即为最优解，记为 $(w_\star, b_\star)$
我们称训练集 $D_{\text{train}} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$ 中满足 $y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) = 1$ 的样本 $(x^{(i)}, y^{(i)})$ 称为**支持向量** (support vector)
从某种意义上说，支持向量相当于是最难分类的样本.

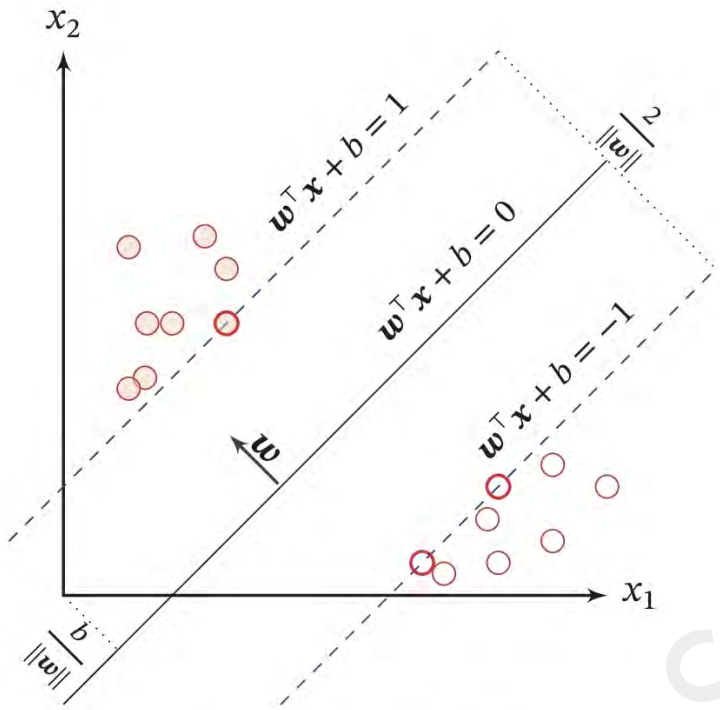

图 3.6　支持向量机示例

4.1.2 问题求解

给定二分类训练集 $D_{\text{train}} = \{x^{(i)}, y^{(i)}\}_{i=1}^n$，其中 $y^{(i)} \in \{-1, 1\}$ 而 $x^{(i)} \in \mathbb{R}^d$
考虑求解以下标准形式的凸优化问题 (具体来说是**二次规划问题**):

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b) \leq 0 \ (i = 1, \ldots, n) \end{aligned}$$

记 $\begin{cases} y = [y^{(1)}, \ldots, y^{(n)}]^{\mathrm{T}} \in \mathbb{R}^n \\ X = [x^{(1)}, \ldots, x^{(n)}] \in \mathbb{R}^{d \times n} \end{cases}$ 则上述凸优化问题可以表示为向量形式:

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1_n - y \odot (X^{\mathrm{T}} w + b 1_n) \preceq 0_n \end{aligned} \tag{P}$$

其中 $\odot$ 代表 Hadamard 乘积 (即逐元素乘积)

定义 **Lagrange 函数**为:

$$\begin{aligned} L(w, b, \lambda) &:= \frac{1}{2}\|w\|_2^2 + \lambda^{\mathrm{T}}(1_n - y \odot (X^{\mathrm{T}} w + b 1_n)) \\ &= \frac{1}{2}\|w\|_2^2 + (1_n - by)^{\mathrm{T}} \lambda - \lambda^{\mathrm{T}} (\mathrm{diag}(y) X^{\mathrm{T}}) w \end{aligned}$$

其中 $\lambda \in \mathbb{R}_+^n$ 为 Langrange 乘子.
KKT 条件为:

$$\begin{cases} \nabla_w L(w, b, \lambda) = w - X\mathrm{diag}(y)\lambda = 0_d \\ \nabla_b L(w, b, \lambda) = -y^{\mathrm{T}} \lambda = 0 \\ 1_n - y \odot (X^{\mathrm{T}} w + b 1_n) \preceq 0_n \\ \lambda \succeq 0_n \\ \lambda_i (1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b)) = 0 \ (i = 1, \ldots, n) \end{cases}$$

上述 KKT 系统本身是可以解的，但其直接求解非常复杂.
注意到这个问题是标准形式的凸优化问题，我们可将其化为对偶问题进行求解.
可以证明对偶问题 (D) 的最优值等于原问题 (P) 的最优值 (即**强对偶性**成立)
(这是因为原问题的只拥有线性不等式约束，这要原问题可行，**Slater 条件**就一定满足)

而且原问题最优解 $w_\star, b_\star$ 与对偶最优解 $\lambda_\star$的关系为 ($(w_\star, b_\star, \lambda_\star)$ 构成鞍点):
(存疑: 严格互补松弛性?)

$$
\begin{aligned}
I_{\text{useless}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = 0\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}} x^{(i)} + b_\star) > 1\} \\
I_{\text{support}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} > 0\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}} x^{(i)} + b_\star) = 1\} \\
w_\star &= X^{\mathrm{T}}\mathrm{diag}(y)\lambda_\star = \sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} x^{(i)} = \sum_{i \in I_{\text{support}}} \lambda_\star^{(i)} y^{(i)} x^{(i)} \\
b_\star &= \frac{1}{y^{(i_0)}} - w^{\mathrm{T}} x^{(i_0)} = y^{(i_0)} - w^{\mathrm{T}} x^{(i_0)} \text{ for any } i_0 \in I_{\text{support}}
\end{aligned}
$$

其中积极指标集 I_{support} 是支持向量对应的指标集，我们可以任选一个支持向量计算最优偏置 $b_\star$

定义 **Lagrange 对偶函数**为:

$$
\begin{aligned}
\Gamma(\lambda) &:= \inf_{w,b} L(w, b, \lambda) \\
&= \inf_{w,b} \left\{ \frac{1}{2}\|w\|_2^2 + (1_n - by)^{\mathrm{T}}\lambda - \lambda^{\mathrm{T}}(\mathrm{diag}(y)X^{\mathrm{T}})w \right\} \quad \left(\text{substitute } \begin{cases} w = X\mathrm{diag}(y)\lambda \\ y^{\mathrm{T}}\lambda = 0_n \end{cases}\right) \\
&= 1_n^{\mathrm{T}}\lambda + \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2 - \lambda^{\mathrm{T}}(\mathrm{diag}(y)X^{\mathrm{T}})X\mathrm{diag}(y)\lambda \\
&= 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2
\end{aligned}
$$

于是我们得到硬间隔的**对偶问题** (D):

$$
\begin{aligned}
\max \quad & \Gamma(\lambda) := 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2 \\
\text{s.t.} \quad & y^{\mathrm{T}}\lambda = 0 \\
& \lambda \succeq 0_n
\end{aligned}
\tag{D}
$$

注意到 Lagrange 对偶函数 $\Gamma(\lambda)$ 是一个关于 λ 的凹函数，且等式约束 $y^{\mathrm{T}}\lambda = 0$ 和不等式约束 $\lambda \succeq 0_n$ 都是仿射的.
故上述对偶问题也是标准形式的凸优化问题 (具体来说是二次规划问题)，可以通过多种凸优化方法来进行求解.
但由于其优化变量 λ 的维度为训练样本数量 n，
而且最终的对偶最优解 $\lambda_\star$ 可能很稀疏 (即 $\lambda_\star$ 有很多分量为 0，只有非零分量对应支持向量，零分量对应的样本是不重要的)
因此一般的优化方法代价比较高.
在实践中通常采用更加高效的优化方法，例如**序列最小优化** (Sequential Minimal Optimization, SMO) 算法.

得到对偶最优解 $\lambda_\star$ 后，我们可以计算原问题最优解 $w_\star, b_\star$:

$$
\begin{aligned}
I_{\text{useless}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = 0\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}} x^{(i)} + b_\star) > 1\} \\
I_{\text{support}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} > 0\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}} x^{(i)} + b_\star) = 1\} \\
w_\star &= X\mathrm{diag}(y)\lambda_\star = \sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} x^{(i)} = \sum_{i \in I_{\text{support}}} \lambda_\star^{(i)} y^{(i)} x^{(i)} \\
b_\star &= \frac{1}{y^{(i_0)}} - w^{\mathrm{T}} x^{(i_0)} = y^{(i_0)} - w^{\mathrm{T}} x^{(i_0)} \text{ for any } i_0 \in I_{\text{support}}
\end{aligned}
$$

其中积极指标集 I_{support} 是支持向量对应的指标集，我们可以任选一个支持向量计算最优偏置 $b_\star$
最终得到支持向量机的**决策函数**:

$$
\begin{aligned}
f(x) &:= \mathrm{sgn}(w_\star^{\mathrm{T}} x + b_\star) \\
&= \mathrm{sgn}\left((X\mathrm{diag}(y)\lambda_\star)^{\mathrm{T}} x + b_\star\right) \\
&= \mathrm{sgn}\left(\sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} (x^{(i)})^{\mathrm{T}} x + b_\star\right) \quad (\text{note that } \lambda_\star^{(i)} = 0 \text{ for all } i \notin I_{\text{support}}) \\
&= \mathrm{sgn}\left(\sum_{i \in I_{\text{support}}} \lambda_\star^{(i)} y^{(i)} (x^{(i)})^{\mathrm{T}} x + b_\star\right) \quad (\text{where } I_{\text{support}} = \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} > 0\})
\end{aligned}
$$

可以看出支持向量机的分类决策只依赖于支持向量 (即指标落在 I_{support} 中的样本，即 $\lambda_\star^{(i)} > 0$ 对应的样本)
与训练样本总数无关，分类速度比较快.
(而且只要保存了所有的支持向量，就相当于保存了模型)

4.1.3 软间隔

回顾 4.1.2 中的凸优化问题 (P):

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1_n - y \odot (X^{\mathrm{T}} w + b 1_n) \preceq 0_n \end{aligned} \tag{P}$$

其中 $\begin{cases} y = [y^{(1)}, \ldots, y^{(n)}]^{\mathrm{T}} \in \mathbb{R}^n \\ X = [x^{(1)}, \ldots, x^{(n)}] \in \mathbb{R}^{d \times n} \end{cases}$

在线性不可分的情况下，上述问题是不可行的 (即没有可行解)
此时我们的任务变为尽可能减少被错误分类的点的个数.
引入松弛变量 $s = [s_1, \ldots, s_n]^{\mathrm{T}}$ 得到等价问题:

$$\begin{aligned} \min_{w,b,s} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & 1_n - y \odot (X^{\mathrm{T}} w + b 1_n) = s \\ & s \preceq 0_n \end{aligned} \tag{P-equal}$$

现在我们抛弃约束 $s \preceq 0_n$ 以**允许出现点的错误分类**
同时引入**惩罚因子** $\mu > 0$ 来惩罚点的错误分类:

$$\begin{aligned} \min_{w,b,s} \quad & \frac{1}{2}\|w\|_2^2 + \mu \sum_{i=1}^{n} \max\{s_i, 0\} \\ \text{s.t.} \quad & 1_n - y \odot (X^{\mathrm{T}} w + b 1_n) = s \end{aligned} \tag{P-slack-and-punished}$$

最后消去松弛变量 s 即得到:

$$\begin{aligned} & \min_{w,b} \left\{ \frac{1}{2}\|w\|_2^2 + \mu \sum_{i=1}^{n} \max\{1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b), 0\} \right\} \\ \Leftrightarrow \quad & \min_{w,b} \left\{ \sum_{i=1}^{n} \max\{1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b), 0\} + \lambda \|w\|_2^2 \right\} \\ \Leftrightarrow \quad & \min_{w,b} \left\{ \sum_{i=1}^{n} \text{Hinge\_loss}(y^{(i)}(w^{\mathrm{T}} x^{(i)} + b)) + \lambda \|w\|_2^2 \right\} \\ \Leftrightarrow \quad & \begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}\|w\|_2^2 + \mu \sum_{i=1}^{n} \xi_i \\ \text{s.t.} \quad & \xi \succeq 1_n - y \odot (w^{\mathrm{T}} x^{(i)} + b) \\ & \xi \succeq 0_n \end{aligned} \end{aligned} \tag{P-punished}$$

其中正则化系数 $\lambda = \frac{1}{2\mu} > 0$
(惩罚因子 $\mu > 0$ 越大，正则化系数 $\lambda > 0$ 就越小)
我们可以把 $\max\{1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b), 0\}$ 视为损失函数，称为 **Hinge 损失函数**:
(它相当于 0-1 损失函数的替代)

$$\text{Hinge\_loss}(z) = [1 - z]_+ = \max\{1 - z, 0\}$$

- 当样本点 $(x^{(i)}, y^{(i)})$ 被正确分类时，我们有 $y^{(i)}(w^{\mathrm{T}} x^{(i)} + b) \geq 1$，此时惩罚为 0
- 当样本点 $(x^{(i)}, y^{(i)})$ 被错误分类时，我们有 $y^{(i)}(w^{\mathrm{T}} x^{(i)} + b) < 1$，此时惩罚为 $1 - y^{(i)}(w^{\mathrm{T}} x^{(i)} + b)$

除了 Hinge 损失函数，我们还可以使用指数损失函数或对数几率损失函数来替代 0-1 损失函数:

$$\text{exp\_loss}(z) := \exp(-z)$$
$$\text{logit\_loss}(z) := \log(1+\exp(-z))$$

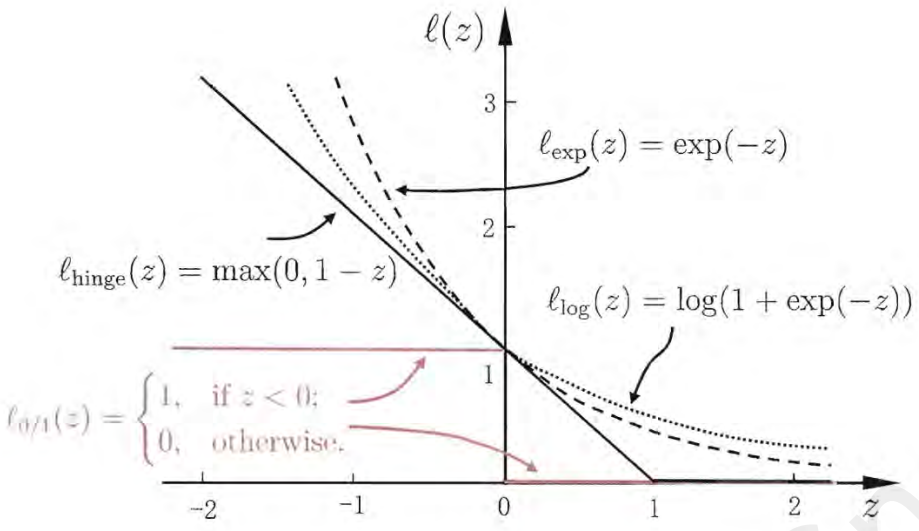

图 6.5 三种常见的替代损失函数: hinge损失、指数损失、对率损失

考虑求解**软阈值支持向量机问题**:

$$\min_{w,b}\left\{\frac{1}{2}\|w\|_2^2+\mu\sum_{i=1}^{n}\max\{1-y^{(i)}(w^{\mathrm{T}}x^{(i)}+b),0\}\right\}$$
$$\Leftrightarrow$$
$$\begin{aligned}\min_{w,b,\xi}\quad & \frac{1}{2}\|w\|_2^2+\mu\sum_{i=1}^{n}\xi_i \\ \text{s.t.}\quad & \xi \succeq 1_n - y\odot(w^{\mathrm{T}}x^{(i)}+b) \\ & \xi \succeq 0_n\end{aligned} \tag{P-punished}$$

定义 **Lagrange 函数**为:

$$\begin{aligned}L(w,b,\xi,\lambda,\nu) &:= \frac{1}{2}\|w\|_2^2+\mu 1_n^{\mathrm{T}}\xi+\lambda^{\mathrm{T}}(1_n-y\odot(X^{\mathrm{T}}w+b1_n)-\xi)-\nu^{\mathrm{T}}\xi \\ &= \frac{1}{2}\|w\|_2^2+(1_n-by)^{\mathrm{T}}\lambda-\lambda^{\mathrm{T}}(\mathrm{diag}(y)X^{\mathrm{T}})w+(\mu 1_n-\lambda-\nu)^{\mathrm{T}}\xi\end{aligned}$$

其中 $\lambda,\nu\in\mathbb{R}_+^n$ 为 Langrange 乘子.
KKT 条件为:

$$\begin{cases}\nabla_w L(w,b,\xi,\lambda,\nu)=w-X\mathrm{diag}(y)\lambda=0_d \\ \nabla_b L(w,b,\xi,\lambda,\nu)=-y^{\mathrm{T}}\lambda=0 \\ \nabla_\xi L(w,b,\xi,\lambda,\nu)=\mu 1_n-\lambda-\nu=0_n \\ \xi\succeq 1_n-y\odot(w^{\mathrm{T}}x^{(i)}+b) \\ \xi\succeq 0_n \\ \lambda\succeq 0_n \\ \nu\succeq 0_n \\ \lambda_i(1-y^{(i)}(w^{\mathrm{T}}x^{(i)}+b)-\xi_i)=0\ (i=1,\dots,n) \\ \nu_i\xi_i=0\ (i=1,\dots,n)\end{cases} \Leftrightarrow \begin{cases}\nabla_w L(w,b,\xi,\lambda,\nu)=w-X\mathrm{diag}(y)\lambda=0_d \\ \nabla_b L(w,b,\xi,\lambda,\nu)=-y^{\mathrm{T}}\lambda=0 \\ \xi\succeq 1_n-y\odot(w^{\mathrm{T}}x^{(i)}+b) \\ \xi\succeq 0_n \\ 0_n\preceq\lambda\preceq\mu I_n \\ \lambda_i(1-y^{(i)}(w^{\mathrm{T}}x^{(i)}+b)-\xi_i)=0\ (i=1,\dots,n) \\ (\mu-\lambda_i)\xi_i=0\ (i=1,\dots,n)\end{cases}$$

上述 KKT 系统本身是可以解的，但其直接求解非常复杂.
定义 **Lagrange 对偶函数**为:

$$
\begin{aligned}
\Gamma(\lambda,\nu) &:= \inf_{w,b,\xi} L(w,b,\xi,\lambda,\nu) \\
&= \inf_{w,b,\xi} \left\{ \frac{1}{2}\|w\|_2^2 + (1_n - by)^{\mathrm{T}}\lambda - \lambda^{\mathrm{T}}(\mathrm{diag}(y)X^{\mathrm{T}})w + (\mu 1_n - \lambda - \nu)^{\mathrm{T}}\xi \right\} \quad \left(\text{substitute } \begin{cases} w = X\mathrm{diag}(y)\lambda \\ y^{\mathrm{T}}\lambda = 0_n \\ \mu 1_n - \lambda - \nu = 0_n \end{cases}\right) \\
&= 1_n^{\mathrm{T}}\lambda + \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2 - \lambda^{\mathrm{T}}(\mathrm{diag}(y)X^{\mathrm{T}})X\mathrm{diag}(y)\lambda \\
&= 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2
\end{aligned}
$$

注意到 Lagrange 对偶函数最终的形式与 ν 无关，
因此可将其简记为 $\Gamma(\lambda) = 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2$
于是我们得到软间隔的**对偶问题** (D-punished):

$$
\begin{aligned}
\max_{\lambda,\nu} \quad & \Gamma(\lambda) := 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2 \\
\text{s.t.} \quad & y^{\mathrm{T}}\lambda = 0 \\
& \mu 1_n - \lambda - \nu = 0_n \\
& \lambda \succeq 0_n \\
& \nu \succeq 0_n
\end{aligned}
\qquad \text{(D-punished)}
$$

$$
\Leftrightarrow
$$

$$
\begin{aligned}
\max_{\lambda} \quad & \Gamma(\lambda) := 1_n^{\mathrm{T}}\lambda - \frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2 \\
\text{s.t.} \quad & y^{\mathrm{T}}\lambda = 0 \\
& 0_n \preceq \lambda \preceq \mu 1_n
\end{aligned}
$$

可以看出软间隔的对偶问题与硬间隔的对偶问题的唯一区别就是约束 $\lambda \succeq 0_n$ 变成了 $0_n \preceq \lambda \preceq \mu 1_n$
它同样是一个标准形式的凸优化问题 (具体来说，是二次规划问题)，可由一般的凸优化求解包进行求解.
在实践中通常采用更加高效的优化方法，例如**序列最小优化** (Sequential Minimal Optimization, SMO) 算法.
记对偶最优解为 $\lambda_\star$，根据 KKT 条件可算得: (强对偶性保证了 KKT 点是鞍点)
(存疑: 严格互补松弛性?)

$$
\begin{aligned}
I_{\text{useless}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = 0\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) > 1\} \\
I_{\text{support}} &:= \{i \in \{1,\dots,n\} : 0 < \lambda_\star^{(i)} < \mu\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) = 1\} \\
I_{\text{punish}} &:= \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = \mu\} = \{i \in \{1,\dots,n\} : y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) < 1\} \\
w_\star &= X\mathrm{diag}(y)\lambda_\star = \sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} x^{(i)} = \sum_{i \in I_{\text{support}} \cup I_{\text{punish}}} \lambda_\star^{(i)} y^{(i)} x^{(i)} \\
b_\star &= \frac{1}{y^{(i_0)}} - w^{\mathrm{T}}x^{(i_0)} = y^{(i_0)} - w^{\mathrm{T}}x^{(i_0)} \text{ for any } i_0 \in I_{\text{support}}
\end{aligned}
$$

- ① 当 $\lambda_\star^{(i)} = 0$ 时，样本 $(x^{(i)}, y^{(i)})$ 对支持向量机没有贡献.
- ② 当 $0 < \lambda_\star^{(i)} \le \mu$ 时，此时我们有 $1 - y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) - \xi_i = 0$ 成立.
  - 当 $0 < \lambda_\star^{(i)} < \mu$ 时，我们还有 $\xi_i = 0$ 成立，因此 $y^{(i)}(w^{\mathrm{T}}x^{(i)} + b) = 1 - \xi_i = 1$
    这表明样本 $(x^{(i)}, y^{(i)})$ 是支持向量.
  - 当 $\lambda_\star^{(i)} = \mu$ 时，我们有 $\xi_i > 0$ 成立 **(存疑: 严格互补松弛性?)**
    - 若 $0 < \xi_i \le 1$，则 $y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) = 1 - \xi_i \ge 0$
      这表明样本 $(x^{(i)}, y^{(i)})$ 虽然受到惩罚 (位于最大间隔内)，但仍然分类正确.
    - 若 $\xi_i > 1$，则 $y^{(i)}(w_\star^{\mathrm{T}}x^{(i)} + b_\star) = 1 - \xi_i < 0$
      这表明样本 $(x^{(i)}, y^{(i)})$ 分类错误，受到更严厉的惩罚.

最终得到软间隔支持向量机的**决策函数**:

$$
\begin{aligned}
f(x) &:= \mathrm{sgn}(w_\star^{\mathrm{T}}x + b_\star) \\
&= \mathrm{sgn}\left((X\mathrm{diag}(y)\lambda_\star)^{\mathrm{T}}x + b_\star\right) \\
&= \mathrm{sgn}\left(\sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} (x^{(i)})^{\mathrm{T}}x + b_\star\right) \quad \left(\text{note that } \lambda_\star^{(i)} = 0 \text{ for all } i \in I_{\text{useless}} = \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = 0\}\right) \\
&= \mathrm{sgn}\left(\sum_{i \in I_{\text{support}} \cup I_{\text{punish}}} \lambda_\star^{(i)} y^{(i)} (x^{(i)})^{\mathrm{T}}x + b_\star\right) \quad \left(\text{where } \begin{cases} I_{\text{support}} = \{i \in \{1,\dots,n\} : 0 < \lambda_\star^{(i)} < \mu\} \\ I_{\text{punish}} = \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = \mu\} \end{cases}\right)
\end{aligned}
$$

4.2 非线性支持向量机

支持向量机可以使用**核函数** (kernel function) 隐式地将样本从原始特征空间映射到更高维的空间，
以解决原始特征空间中的线性不可分问题.

4.2.1 核方法

核方法通过一个非线性算子将输入空间 $\mathcal{X}=\mathbb{R}^d$ 映射到特征空间 $\mathcal{H}$ (通常要求是 Hilbert 空间，即完备内积空间)
使得在输入空间 $\mathbb{R}^d$ 中的超曲面模型对应于特征空间 $\mathcal{H}$ 中的超平面模型.
这样二分类任务可通过在特征空间中求解线性支待向量机来完成.

设输入空间 $\mathcal{X}=\mathbb{R}^d$，特征空间是 Hilbert 空间 $\mathcal{H}$ (内积记为 $\langle\cdot,\cdot\rangle$)
给定映射 $\phi:\mathcal{X}\mapsto\mathcal{H}$
若函数 $k(\cdot,\cdot)$ 满足 $k(x,z)=\langle\phi(x),\phi(z)\rangle$ $(\forall\ x,z\in\mathcal{X})$，则我们称 $k(\cdot,\cdot)$ 为映射 ϕ 的**核函数**

在学习和预测时我们使用核函数 $k(\cdot,\cdot)$，则不显式地定义映射 $\phi(\cdot)$
通常来说，直接计算 $k(x,z)$ 比较容易，而显式地计算 $\phi(x),\phi(z)$ 再计算内积 $\langle\phi(x),\phi(z)\rangle$ 并不容易.
以 $\mathcal{X}=\mathbb{R}^2$ 为例:

- ① 取 $\phi(x):=[x_1^2,\sqrt{2}x_1x_2,x_2^2]^{\mathrm{T}}$，将 $\mathcal{X}=\mathbb{R}^2$ 映射到 $\mathcal{H}=\mathbb{R}^3$
  可以验证 $k(x,z)=(x^{\mathrm{T}}z)^2=\phi(x)^{\mathrm{T}}\phi(z)$ 是对应的核函数.
- ② 取 $\phi(x):=\frac{1}{\sqrt{2}}[x_1^2-x_2^2,2x_1x_2,x_1^2+x_2^2]^{\mathrm{T}}$，将 $\mathcal{X}=\mathbb{R}^2$ 映射到 $\mathcal{H}=\mathbb{R}^3$
  可以验证 $k(x,z)=(x^{\mathrm{T}}z)^2=\phi(x)^{\mathrm{T}}\phi(z)$ 是对应的核函数.
- ③ 取 $\phi(x):=[x_1^2,x_1x_2,x_1x_2,x_2^2]^{\mathrm{T}}$，将 $\mathcal{X}=\mathbb{R}^2$ 映射到 $\mathcal{H}=\mathbb{R}^4$
  可以验证 $k(x,z)=(x^{\mathrm{T}}z)^2=\phi(x)^{\mathrm{T}}\phi(z)$ 是对应的核函数.
- ④ 取 $\phi(x):=[1,\sqrt{2}x_1,\sqrt{2}x_2,\sqrt{2}x_1x_2,x_1^2,x_2^2]^{\mathrm{T}}$，将 $\mathcal{X}=\mathbb{R}^2$ 映射到 $\mathcal{H}=\mathbb{R}^6$
  可以验证 $k(x,z)=(1+x^{\mathrm{T}}z)^2=\phi(x)^{\mathrm{T}}\phi(z)$ 是对应的核函数.

通常所说的核函数 $k(\cdot,\cdot)$ 默认是**正定核函数** (positive definite kernel function)
一个对称函数 $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ 构成正定核函数的充要条件如下:
(Mercer 定理, 统计学习方法, 定理 7.5)
设 $k:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}$ 是对称函数 (即满足 $k(z,x)=k(x,z)$ $(\forall\ x,z\in\mathcal{X})$)
则 $k(\cdot,\cdot)$ 为正定核函数当且仅当对于任意正整数 $m\in\mathbb{Z}_+$ 和 $x_1,\dots,x_m\in\mathcal{X}$ 都有 Gram 矩阵半正定.
其中 $x_1,\dots,x_m\in\mathcal{X}$ 的 Gram 矩阵的定义如下:

$$K:=[k(x_i,x_j)]_{i,j=1}^m$$

4.2.2 应用

考虑软间隔支持向量机的对偶问题 $(\mathrm{D\text{-}punished})$:

$$\begin{aligned}\max_{\lambda}\quad&\Gamma(\lambda):=1_n^{\mathrm{T}}\lambda-\frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2\\ \text{s.t.}\quad&y^{\mathrm{T}}\lambda=0\\&0_n\preceq\lambda\preceq\mu 1_n\end{aligned}$$

其中 $\begin{cases}y=[y^{(1)},\dots,y^{(n)}]^{\mathrm{T}}\in\mathbb{R}^n\\X=[x^{(1)},\dots,x^{(n)}]\in\mathbb{R}^{d\times n}\end{cases}$
目标函数可以展开为:

$$\begin{aligned}\Gamma(\lambda)&=1_n^{\mathrm{T}}\lambda-\frac{1}{2}\|X\mathrm{diag}(y)\lambda\|_2^2\\&=1_n^{\mathrm{T}}\lambda-\frac{1}{2}\lambda^{\mathrm{T}}\mathrm{diag}(y)X^{\mathrm{T}}X\mathrm{diag}(y)\lambda\\&=1_n^{\mathrm{T}}\lambda-\frac{1}{2}(y\odot\lambda)^{\mathrm{T}}X^{\mathrm{T}}X(y\odot\lambda)\end{aligned}$$

注意到 $X^{\mathrm{T}}X\in\mathbb{R}^n$ 是一个 Euclid 内积定义的 Gram 矩阵:

$$X^{\mathrm{T}}X=[(x^{(i)})^{\mathrm{T}}x^{(j)}]_{i,j=1}^{n}=\begin{bmatrix}(x^{(1)})^{\mathrm{T}}x^{(1)} & \cdots & (x^{(1)})^{\mathrm{T}}x^{(n)}\\ \vdots & & \vdots\\ (x^{(n)})^{\mathrm{T}}x^{(1)} & \cdots & (x^{(n)})^{\mathrm{T}}x^{(n)}\end{bmatrix}$$

给定映射 $\phi:\mathcal{X}=\mathbb{R}^d\mapsto\mathcal{H}$ 和核函数 $k(\cdot,\cdot)$，我们定义:

$$\begin{aligned}\phi(X)&=[\phi(x^{(1)}),\dots,\phi(x^{(n)})]\\ K(X,X)&=(\phi(X))^{\mathrm{T}}\phi(X)\\ &=[(\phi(x^{(i)}))^{\mathrm{T}}\phi(x^{(j)})]_{i,j=1}^{n}\\ &=[k(x^{(i)},x^{(j)})]_{i,j=1}^{n}\\ &=\begin{bmatrix}k(x^{(1)},x^{(1)}) & \cdots & k(x^{(1)},x^{(n)})\\ \vdots & & \vdots\\ k(x^{(n)},x^{(1)}) & \cdots & k(x^{(n)},x^{(n)})\end{bmatrix}\end{aligned}$$

我们可以定义新的目标函数:

$$\begin{aligned}\Gamma(\lambda)&:=1_n^{\mathrm{T}}\lambda-\frac{1}{2}\|\phi(X)\mathrm{diag}(y)\lambda\|^2\\ &=1_n^{\mathrm{T}}\lambda-\frac{1}{2}(\mathrm{diag}(y)\lambda)^{\mathrm{T}}K(X,X)(\mathrm{diag}(y)\lambda)\\ &=1_n^{\mathrm{T}}\lambda-\frac{1}{2}(y\odot\lambda)^{\mathrm{T}}K(X,X)(y\odot\lambda)\end{aligned}$$

这样我们就得到了新的对偶问题:

$$\begin{aligned}\max_{\lambda}\quad&\Gamma(\lambda):=1_n^{\mathrm{T}}\lambda-\frac{1}{2}(y\odot\lambda)^{\mathrm{T}}K(X,X)(y\odot\lambda)\\ \text{s.t.}\quad&y^{\mathrm{T}}\lambda=0\\ &0_n\preceq\lambda\preceq\mu 1_n\end{aligned}$$

其中 $\begin{cases}y=[y^{(1)},\dots,y^{(n)}]^{\mathrm{T}}\in\mathbb{R}^n\\ X=[x^{(1)},\dots,x^{(n)}]\in\mathbb{R}^{d\times n}\\ K(X,X)=[k(x^{(i)},x^{(j)})]_{i,j=1}^{n}\in\mathbb{R}^{n\times n}\end{cases}$

解得对偶最优解 $\lambda_\star$ 后，我们便可计算:
(由于我们不需要显式计算 $w_\star$，故我们只需知道核函数 $k(\cdot,\cdot)$ 即可，而无需知道映射 $\phi:\mathcal{X}=\mathbb{R}^d\mapsto\mathcal{H}$)

$$\begin{aligned}I_{\text{useless}}&:=\{i\in\{1,\dots,n\}:\lambda_\star^{(i)}=0\}=\{i\in\{1,\dots,n\}:y^{(i)}(w_\star^{\mathrm{T}}\phi(x^{(i)})+b_\star)>1\}\\ I_{\text{support}}&:=\{i\in\{1,\dots,n\}:0<\lambda_\star^{(i)}<\mu\}=\{i\in\{1,\dots,n\}:y^{(i)}(w_\star^{\mathrm{T}}\phi(x^{(i)})+b_\star)=1\}\\ I_{\text{punish}}&:=\{i\in\{1,\dots,n\}:\lambda_\star^{(i)}=\mu\}=\{i\in\{1,\dots,n\}:y^{(i)}(w_\star^{\mathrm{T}}\phi(x^{(i)})+b_\star)<1\}\\ w_\star&=X\mathrm{diag}(y)\lambda_\star=\sum_{i=1}^{n}\lambda_\star^{(i)}y^{(i)}\phi(x^{(i)})=\sum_{i\in I_{\text{support}}\cup I_{\text{punish}}}\lambda_\star^{(i)}y^{(i)}\phi(x^{(i)})\\ b_\star&=\frac{1}{y^{(i_0)}}-w^{\mathrm{T}}\phi(x^{(i_0)})\\ &=y^{(i_0)}-w^{\mathrm{T}}\phi(x^{(i_0)})\\ &=\sum_{i\in I_{\text{support}}\cup I_{\text{punish}}}\lambda_\star^{(i)}y^{(i)}(\phi(x^{(i)}))^{\mathrm{T}}\phi(x^{(i_0)})\\ &=\sum_{i\in I_{\text{support}}\cup I_{\text{punish}}}\lambda_\star^{(i)}y^{(i)}k(x^{(i)},x^{(i_0)})\text{ for any }i_0\in I_{\text{support}}\end{aligned}$$

最终得到非线性支持向量机的决策函数:

$$
\begin{aligned}
f(x) &:= \operatorname{sgn}(w_\star^{\mathrm{T}}\phi(x) + b_\star) \\
&= \operatorname{sgn}\left((\phi(X)\operatorname{diag}(y)\lambda_\star)^{\mathrm{T}}\phi(x) + b_\star\right) \\
&= \operatorname{sgn}\left(\sum_{i=1}^{n} \lambda_\star^{(i)} y^{(i)} (\phi(x^{(i)}))^{\mathrm{T}}\phi(x) + b_\star\right) \quad (\text{note that } \lambda_\star^{(i)} = 0 \text{ for all } i \in I_{\text{useless}} = \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = 0\}) \\
&= \operatorname{sgn}\left(\sum_{i \in I_{\text{support}} \cup I_{\text{punish}}} \lambda_\star^{(i)} y^{(i)} (\phi(x^{(i)}))^{\mathrm{T}}\phi(x) + b_\star\right) \quad \left(\text{where } \begin{cases} I_{\text{support}} = \{i \in \{1,\dots,n\} : 0 < \lambda_\star^{(i)} < \mu\} \\ I_{\text{punish}} = \{i \in \{1,\dots,n\} : \lambda_\star^{(i)} = \mu\} \end{cases}\right) \\
&= \operatorname{sgn}\left(\sum_{i \in I_{\text{support}} \cup I_{\text{punish}}} \lambda_\star^{(i)} y^{(i)} k(x^{(i)}, x) + b_\star\right)
\end{aligned}
$$

The End