机器学习与数据挖掘

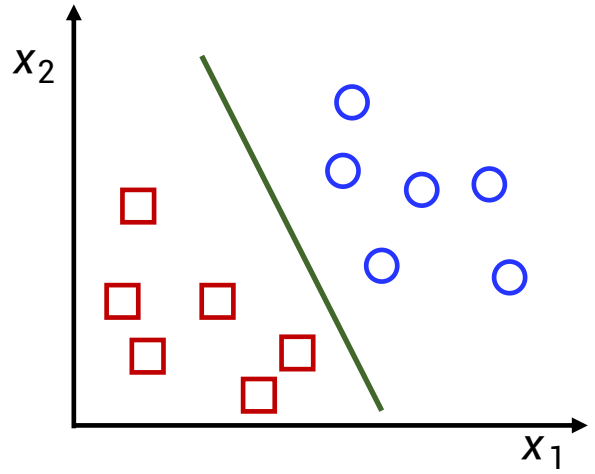
支持向量机

课程大纲

- 线性分类器的决策边界
- 线性最大间隔分类器
- 软线性最大间隔分类器
- 支持向量机
- 与逻辑回归的关系

线性分类器中的决策边界

- 在线性分类器中，决策边界总是一个超平面。目标是找到能够分离不同类型样本的超平面

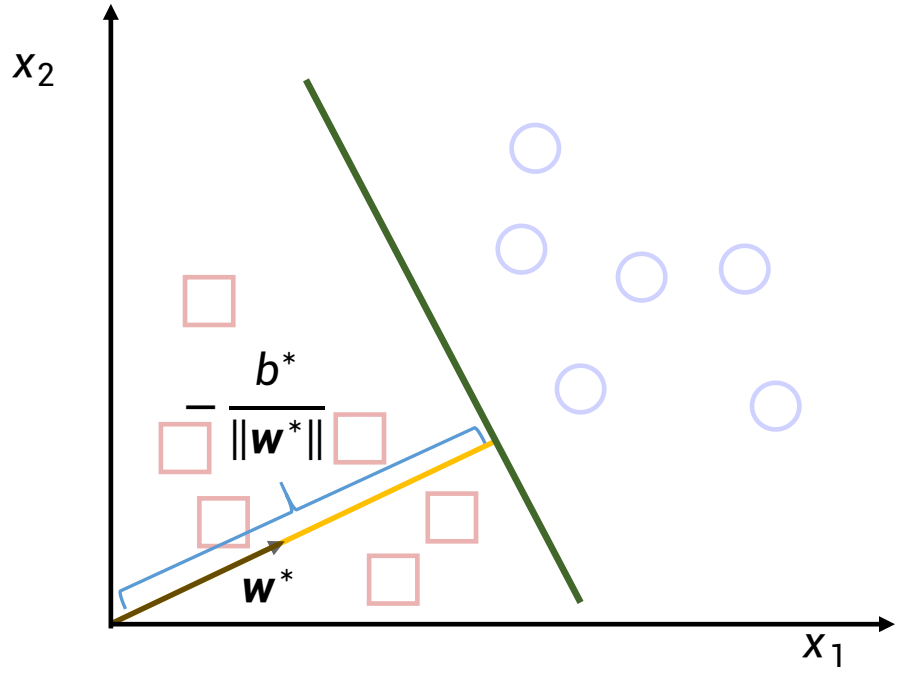


- 逻辑回归
  - 决策边界超平面是通过最小化交叉熵损失来找到的

$$L(\boldsymbol{w}, b) = -y\log\left(\sigma(\boldsymbol{w}^T\boldsymbol{x}+b)\right) - (1-y)\log\left(1-\sigma(\boldsymbol{w}^T\boldsymbol{x}+b)\right)$$

- 当得到最优的 $\boldsymbol{w}^*$ 和 b^* 时，超平面由下式的 $\boldsymbol{x}$ 组成：

$$\{\boldsymbol{x}|\boldsymbol{w}^{*T}\boldsymbol{x}+b^*=0\}$$



1) 超平面与向量 $\boldsymbol{w}^*$ 垂直

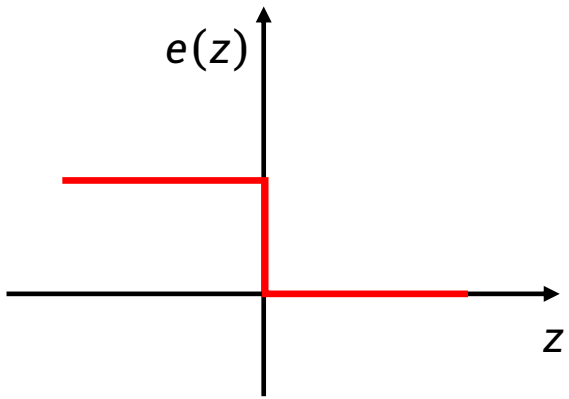
2) 从原点到平面的距离是 $-\frac{b^*}{\|\boldsymbol{w}^*\|}$

- 理想分类器
  - 超平面是通过最小化损失来确定的

$$L(\boldsymbol{w},\, b) = \sum_{\ell=1}^{N} e\left(y^{(\ell)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(\ell)} + b\right)\right)$$

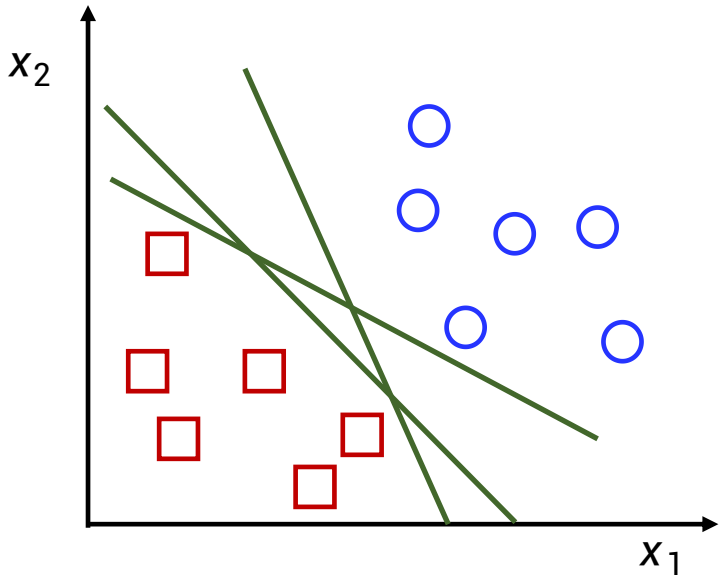
$L(\boldsymbol{w},\, b)$代表被错误分类的样本数量

- $y \in \{-1,\, 1\}$
- *如果* $z \geq 0$，*则* $e(z) = 0$；*否则* $E[z] = 1$



- 如果样本是线性可分的，将会有无数个理想分类器，它们由 $\boldsymbol{w}^*$ 和 b^* 决定

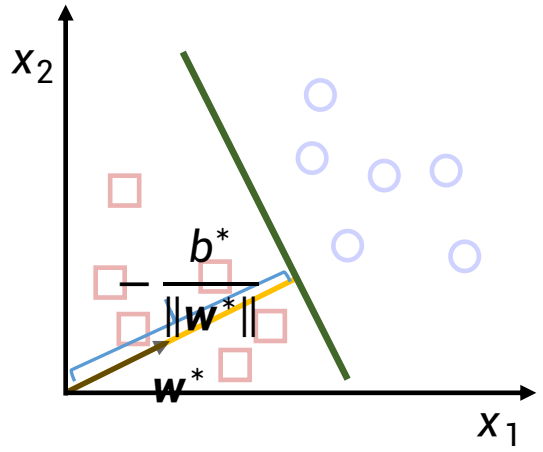
- 每一对 $\boldsymbol{w}^*$ 和 b^* 都对应一个超平面



上述所有超平面都能使损失降为零

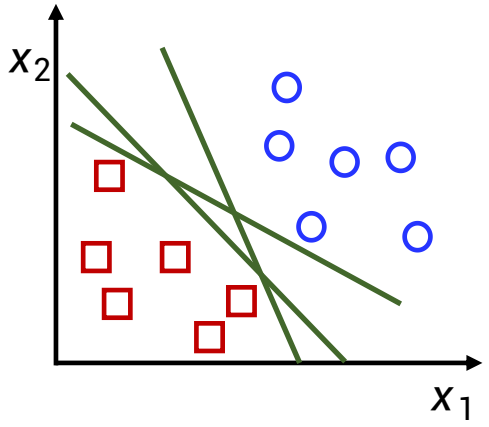
哪个超平面是最好的？

- 下面的超平面从*最小化交叉熵损失* 的角度来看是最优的



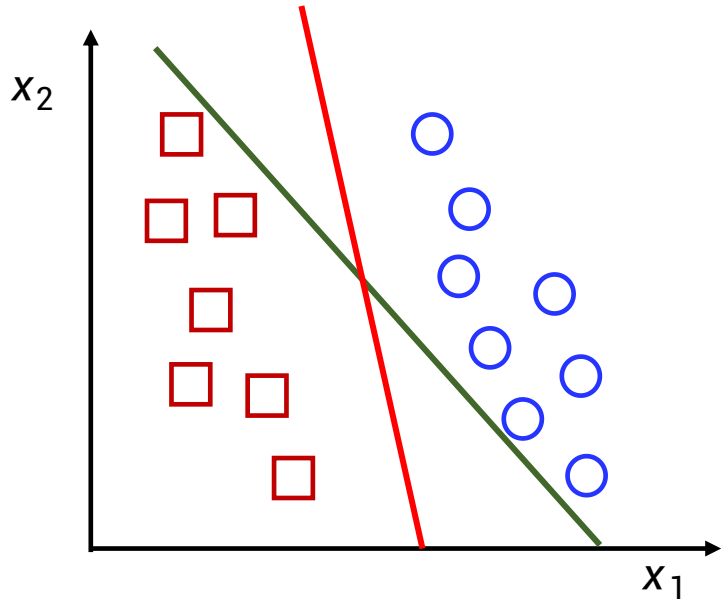
哪个超平面是最好的？

- 下面的超平面从*最小化误分类样本数量* 的角度来看是最优的



- 这些超平面都是在*训练样本*上进行评估的
- 但我们真正需要的是在*（未见过的）测试数据*上的表现

根据我们的直觉，下面的哪个超平面更有可能在测试数据上产生更好的结果？

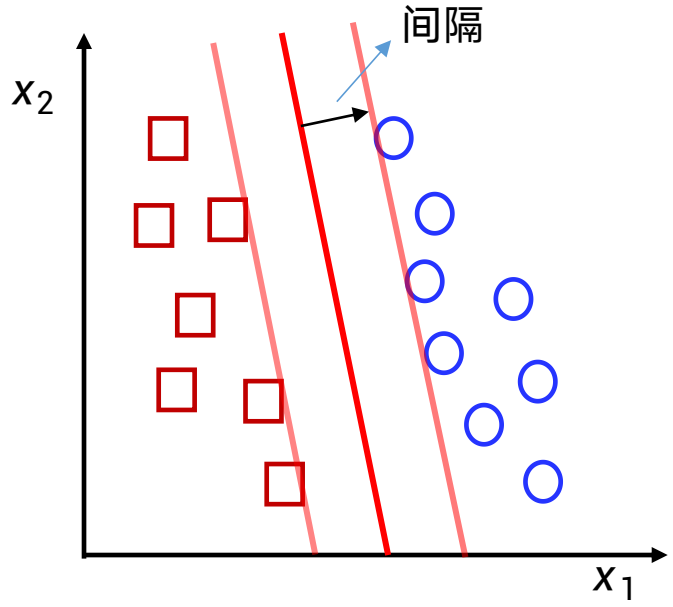


课程大纲

- 线性分类器的决策边界
- 线性最大间隔分类器
- 软线性最大间隔分类器
- 支持向量机
- 与逻辑回归的关系

最大间隔目标

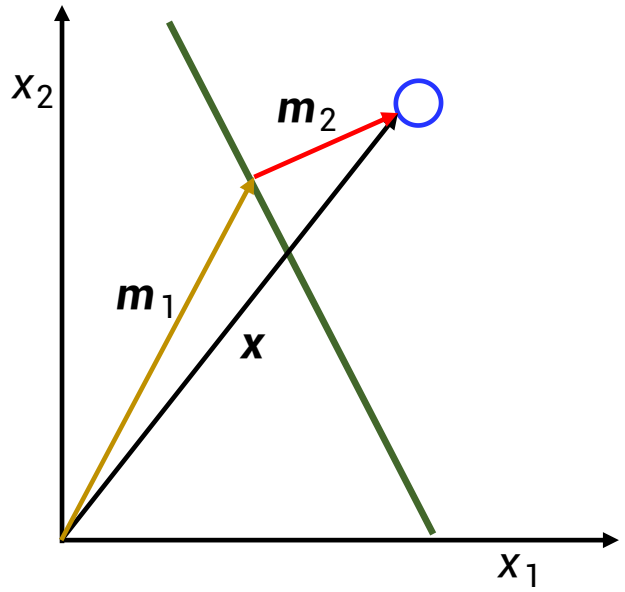
- 为了在未见过的数据上表现良好，直觉是找到一个能尽可能扩大间隔的超平面



当间隔较大时，我们可以预期*未见过的样本有更高的几率被正确分类*

如何表示间隔？

- 样本 $\boldsymbol{x}$ 到超平面 $\mathcal{H}$ 的距离。记 $\boldsymbol{w}^T\boldsymbol{x}+b=\mathrm{h}(\boldsymbol{x})$



  - 每个 $\boldsymbol{x}$ 都可以被分解为：

$$\boldsymbol{x}=\boldsymbol{m}_1+\boldsymbol{m}_2$$

    - $\boldsymbol{m}_1$ 在超平面 $\mathcal{H}$ 上，即 $\boldsymbol{w}^T\boldsymbol{m}_1+b=0$
    - $\boldsymbol{m}_2\perp\mathcal{H}$ 及 $\boldsymbol{m}_2\parallel\boldsymbol{w}$
  - 因此，我们有：

$$\boldsymbol{w}^T\boldsymbol{x}+b=\boldsymbol{w}^T(\boldsymbol{m}_1+\boldsymbol{m}_2)+b=\boldsymbol{w}^T\boldsymbol{m}_2=\mathrm{h}(\boldsymbol{x})$$

  - 由于 $\boldsymbol{m}_2\parallel\boldsymbol{w}$，我们可以写成：

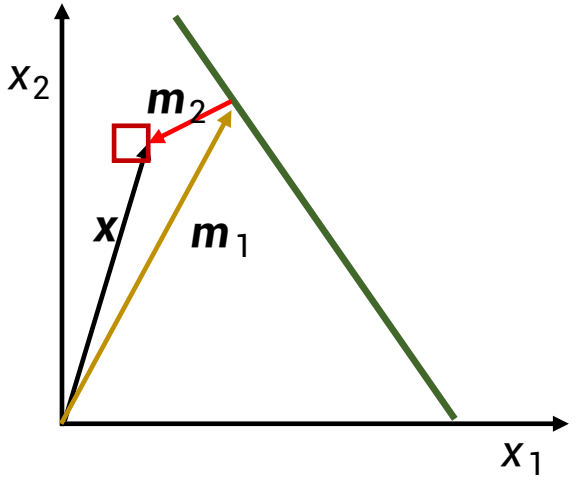
$$\boldsymbol{m}_2=\gamma\cdot\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},$$

其中 $|\gamma|$ *代表* $\boldsymbol{m}_2$ *的长度*

- 将 $\boldsymbol{m}_2 = \gamma \cdot \frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$ 代入 $\mathrm{h}(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{m}_2$ 得到：

$$\mathrm{h}(\boldsymbol{x}) = \gamma \cdot \frac{\boldsymbol{w}^T \boldsymbol{w}}{\|\boldsymbol{w}\|} \quad \Longrightarrow \quad \gamma = \frac{\mathrm{h}(\boldsymbol{x})}{\|\boldsymbol{w}\|}$$

- 位于超平面上方的样本 $\boldsymbol{x}$ 到超平面的距离为 $\frac{\mathrm{h}(\boldsymbol{x})}{\|\boldsymbol{w}\|}$
- 位于超平面下方的样本 $\boldsymbol{x}$ 到超平面的距离为 $-\frac{\mathrm{h}(\boldsymbol{x})}{\|\boldsymbol{w}\|}$



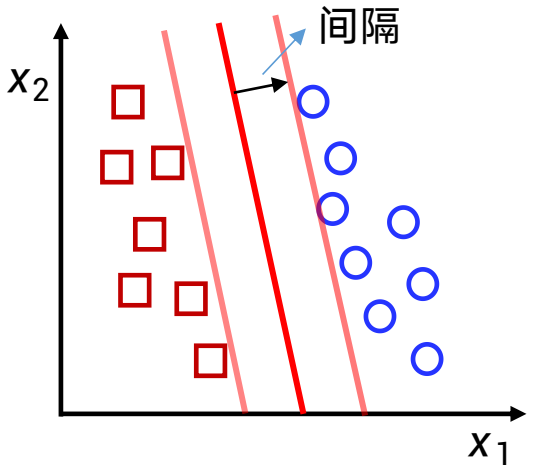
- 样本 $(\boldsymbol{x}, y)$ 到超平面的距离由下式给出：

$$\frac{y \cdot \mathrm{h}(\boldsymbol{x})}{\|\boldsymbol{w}\|} = \frac{y \cdot (\boldsymbol{w}^T \boldsymbol{x} + b)}{\|\boldsymbol{w}\|}$$

其中 $y \in \{-1, 1\}$

- 一个超平面在一个数据集上的间隔由最小距离给出，即：

$$Margin = \min_{\ell} \frac{y^{(\ell)} \cdot (\boldsymbol{w}^T \boldsymbol{x}^{(\ell)} + b)}{\|\boldsymbol{w}\|}$$



- 因此，最大间隔分类器的目标是找到最优的 $\boldsymbol{w}^*$ 和 b^* 来使间隔最大化，即：

$$\boldsymbol{w}^*, b^* = \arg\max_{\boldsymbol{w}, b} \left\{ \frac{1}{\|\boldsymbol{w}\|} \min_{\ell} \left[y^{(\ell)} \cdot (\boldsymbol{w}^T \boldsymbol{x}^{(\ell)} + b) \right] \right\}$$

但是，如何优化是未知的

变换后的目标函数

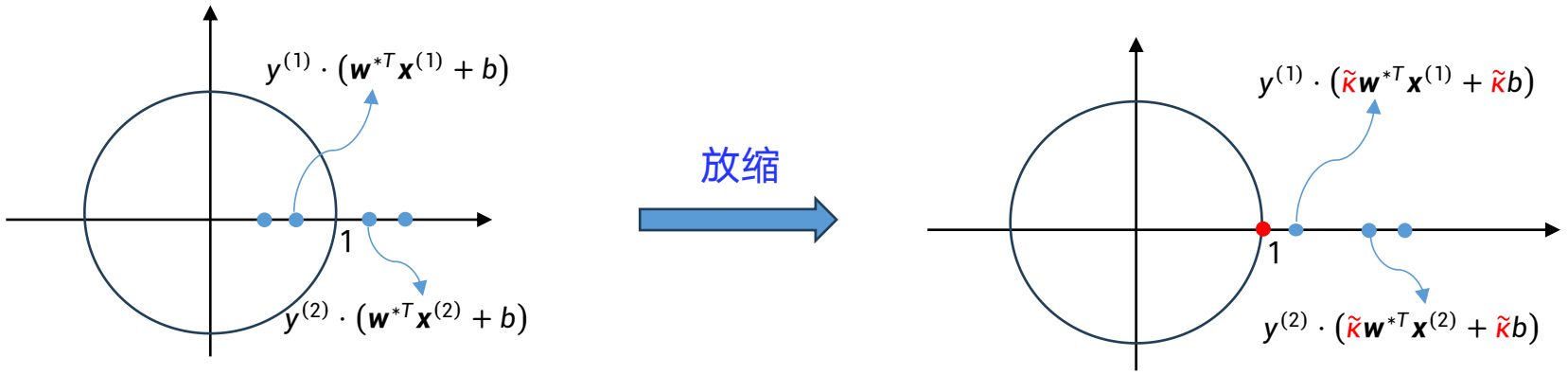
- 基本思想：优化一个与原问题具有相同最优解的目标函数

  - 假设 $\boldsymbol{w}^*$ 和 b^* 是 $\frac{1}{\|\boldsymbol{w}\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\boldsymbol{w}^T\boldsymbol{x}^{(\ell)}+b\right)\right]$ 的最优解。那么，对于所有 $\kappa\neq 0$， $\kappa\boldsymbol{w}^*$ 和 κb^* 也必定是最优解

  - 此外，总存在一个特定的 $\tilde{\kappa}$ 使得

$$\text{对于所有 } \ell=1,2,\cdots,n,\ y^{(\ell)}\cdot\left(\tilde{\kappa}\boldsymbol{w}^{*T}\boldsymbol{x}^{(\ell)}+\tilde{\kappa}b^*\right)\geq 1$$

并且至少有一个等号成立



- 因此，可以通过求解无约束优化问题来找到 $\frac{1}{\|\boldsymbol{w}\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\boldsymbol{w}^T\boldsymbol{x}^{(\ell)}+b\right)\right]$ 的最大值：

$$\max_{\boldsymbol{w},b}\left[\frac{1}{\|\boldsymbol{w}\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\boldsymbol{w}^T\boldsymbol{x}^{(\ell)}+b\right)\right]\right]$$

或者通过求解带约束的优化问题来找到：

$$\max_{\widetilde{\boldsymbol{w}},\tilde{b}}\left[\frac{1}{\|\widetilde{\boldsymbol{w}}\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\widetilde{\boldsymbol{w}}^T\boldsymbol{x}^{(\ell)}+\tilde{b}\right)\right]\right]$$

$$s.t.:\quad \text{对于所有 } \ell=1,2,\cdots,n,\ y^{(\ell)}\cdot\left(\widetilde{\boldsymbol{w}}^T\boldsymbol{x}^{(\ell)}+\tilde{b}\right)\geq 1$$

并且至少有一个等号成立

- 最优解 $\boldsymbol{w}^*$ 和 $\widetilde{\boldsymbol{w}}^*$ 可能不相同，但它们所得到的 $\frac{1}{\|\boldsymbol{w}^*\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(\ell)}+b^*\right)\right]$ 和 $\frac{1}{\|\widetilde{\boldsymbol{w}}^*\|}\min_{\ell}\left[y^{(\ell)}\cdot\left(\widetilde{\boldsymbol{w}}^{*T}\boldsymbol{x}^{(\ell)}+\tilde{b}^*\right)\right]$ 的值必定相等

- 在第二个优化问题中，由于对于所有 $\ell = 1,2,\cdots,n$ ， $y^{(\ell)} \cdot \left(\widetilde{\boldsymbol{w}}^T \boldsymbol{x}^{(\ell)} + \tilde{b}\right) \geq 1$ 并且至少有一个等号成立，我们可以很容易地得到：

$$\min_{\ell} \left[y^{(\ell)} \cdot \left(\widetilde{\boldsymbol{w}}^T \boldsymbol{x}^{(\ell)} + \tilde{b}\right) \right] = 1$$

因此，优化目标可以简化为 $\frac{1}{\|\widetilde{\boldsymbol{w}}\|}$

- 最大化 $\frac{1}{\|\widetilde{\boldsymbol{w}}\|}$ 可以等价于最小化 $\|\widetilde{\boldsymbol{w}}\|^2$ 。因此，该问题可以等价地写为：

$$\min_{\widetilde{\boldsymbol{w}},\tilde{b}} \|\widetilde{\boldsymbol{w}}\|^2$$

$$s.t. \quad \text{对于所有 } \ell = 1,2,\cdots,n, \quad y^{(\ell)} \cdot \left(\widetilde{\boldsymbol{w}}^T \boldsymbol{x}^{(\ell)} + \tilde{b}\right) \geq 1$$

至少有一个等号成立

- 当最小化 $\|\widetilde{\boldsymbol{w}}\|^2$ 时，“至少有一个等号成立”的约束条件将自动满足（为什么？？）。因此，可以在不影响结果的情况下将其移除

- 因此，最大间隔超平面可以通过求解下面的优化问题来找到：

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

$$s.t.: \ y^{(\ell)} \cdot \left(\boldsymbol{w}^T \boldsymbol{x}^{(\ell)} + b\right) \geq 1, \quad \text{for } \ell = 1, 2, \cdots, N$$

- 这是一个二次规划问题。它的最优解可以高效地通过*数值方法* 找到

- 这是一个二次规划问题。它的最优解可以高效地通过*数值方法* 找到

二次规划问题：

$$\min_{x \in \mathbb{R}^n} \quad \frac{1}{2} x^T Q x + c^T x \text{ subject to} \quad Ax \leq b, \quad Ex = d, \quad l \leq x \leq u$$

$\frac{1}{2}\|\boldsymbol{w}\|^2$

x 是决策变量的向量
Q 是一个 n × n 的对称正定矩阵，定义了二次项
c 是线性项的系数向量
A 和 *b* 定义了不等式约束
E 和 *d* 定义了等式约束
l 和 *u* 定义了变量的下界和上界

- 这是一个二次规划问题。它的最优解可以高效地通过*数值方法* 找到

数值方法：牛顿法，
梯度下降法（神经网络学习），
共轭梯度下降法，
.....

- 当得到最优的 $\boldsymbol{w}^*$ 和 b^* 时，一个未见过的数据 $\boldsymbol{x}$ 可以被分类为：

$$\hat{y}(\boldsymbol{x}) = sign\left(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*\right)$$

等价对偶形式

- 每个凸优化问题都对应一个等价的对偶形式

本节内容摘自凸优化这一学科

- 原始优化问题的拉格朗日函数

$$\mathcal{L}(\boldsymbol{w},\ b,\ \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{\ell=1}^{N} a_\ell \left(y^{(\ell)}\left(\boldsymbol{w}^T\boldsymbol{x}^{(\ell)} + b\right) - 1\right),$$

其中，拉格朗日乘子 a_ℓ 必须满足 $a_\ell \geq 0$

- 拉格朗日对偶函数

$$g(\boldsymbol{a}) = \min_{\boldsymbol{w},b} \mathcal{L}(\boldsymbol{w},\ b,\ \boldsymbol{a})$$

- 原始优化问题的对偶形式

$$\max_{\boldsymbol{a}}\ g(\boldsymbol{a})$$
$$s.t.:\ \boldsymbol{a} \geq \boldsymbol{0}$$

- 推导函数 $g(\boldsymbol{a})$ 的闭式表达式
  - 将梯度 $\frac{\partial \mathbb{1}}{\partial \boldsymbol{w}}$ 和 $\frac{\partial \mathbb{1}}{\partial b}$ 设置为 0，得到：

$$\boldsymbol{w} = \sum_{\ell=1}^{N} a_\ell y^{(\ell)} \boldsymbol{x}^{(\ell)} \qquad \sum_{\ell=1}^{N} a_\ell y^{(\ell)} = 0$$

$$\mathbb{1} = \frac{1}{2}\|\boldsymbol{w}\|^2 - \sum_{\ell=1}^{N} a_\ell \left(y^{(\ell)}(\boldsymbol{w}^T \boldsymbol{x}^{(\ell)} + b) - 1\right)$$

  - 将它们代入 $\mathbb{1}(\boldsymbol{w}, b, \boldsymbol{a})$，得到 $g(\boldsymbol{a}) = \min_{\boldsymbol{w},b} \mathbb{1}(\boldsymbol{w}, b, \boldsymbol{a})$ 为：

$$g(\boldsymbol{a}) = \underbrace{\sum_{\ell=1}^{N} a_\ell - \frac{1}{2}\sum_{\ell=1}^{N}\sum_{j=1}^{N} a_\ell a_j y^{(\ell)} y^{(j)} \boldsymbol{x}^{(\ell)T} \boldsymbol{x}^{(j)}}_{= \mathbf{1}^T \boldsymbol{a} - \frac{1}{2}\boldsymbol{a}^T \boldsymbol{M} \boldsymbol{a}}$$

其中 $[\boldsymbol{M}]_{\ell j} \triangleq y^{(\ell)} y^{(j)} \boldsymbol{x}^{(\ell)T} \boldsymbol{x}^{(j)}$

- 因此，对偶优化问题变为：

$$\max_{\boldsymbol{a}} \; g(\boldsymbol{a})$$
$$s.t.: \boldsymbol{a} \geq \mathbf{0} \text{ 及 } \sum_{\ell=1}^{N} a_\ell y^{(\ell)} = 0$$

其中 $g(\boldsymbol{a}) = \underbrace{\sum_{\ell=1}^{N} a_\ell - \frac{1}{2}\sum_{\ell=1}^{N}\sum_{j=1}^{N} a_\ell a_j y^{(\ell)} y^{(j)} \boldsymbol{x}^{(\ell)T} \boldsymbol{x}^{(j)}}_{= \mathbf{1}^T \boldsymbol{a} - \frac{1}{2}\boldsymbol{a}^T \boldsymbol{M} \boldsymbol{a}}$

它也是一个二次规划问题，可以通过*数值方法* 高效地求解

- 最优解 $\boldsymbol{w}^*$、b^* 与最优解 $\boldsymbol{a}^*$ 之间的关系

  - 给定最优解 $\boldsymbol{a}^*$，根据 $\boldsymbol{w} = \sum_{\ell=1}^{N} a_\ell y^{(\ell)} \boldsymbol{x}^{(\ell)}$，最优的 $\boldsymbol{w}^*$ 可以等价地表示为：

$$\boldsymbol{w}^* = \sum_{\ell=1}^{N} a_\ell^* y^{(\ell)} \boldsymbol{x}^{(\ell)}$$

  - 由于对于所有位于间隔边界上的样本 $(\boldsymbol{x}^{(\ell)}, y^{(\ell)})$，都有 $y^{(\ell)}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(\ell)} + b) = 1$，我们可以推导出：

$$b^* = \frac{1}{N_{\mathbb{S}}} \sum_{n \in \mathbb{S}} \left(y^{(n)} - \sum_{m} a_m^* y^{(m)} \boldsymbol{x}^{(n)T} \boldsymbol{x}^{(m)} \right)$$

其中 $\mathbb{S}$ 表示位于间隔边界上的样本集合

- 最大间隔分类器
  - 原问题形式

$$\hat{y}(\boldsymbol{x}) = sign(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

  - 对偶问题形式

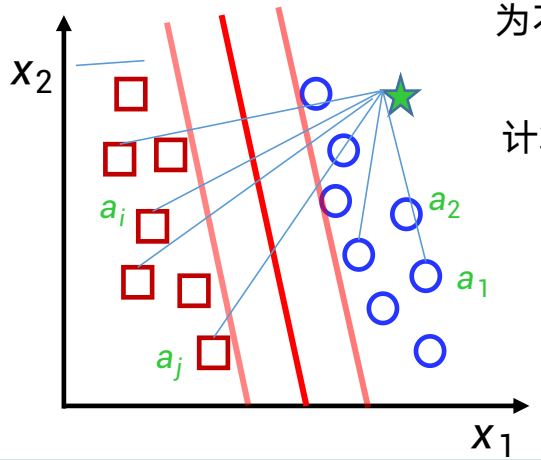
将 $\boldsymbol{w}^* = \sum_{n=1}^{N} a_n^* y^{(n)} \boldsymbol{x}^{(n)}$ 代入原问题形式，得到：

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^* y^{(n)} \boldsymbol{x}^{(n)T}\boldsymbol{x} + b^*\right)$$

这两个分类器是*等价的*

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^* \cdot (\boldsymbol{x}^{(n)T}\boldsymbol{x}) \cdot y^{(n)} + b^*\right)$$

- 如何理解对偶最大间隔分类器？
  - 对于一个测试样本 $\boldsymbol{x}$，通过计算它与所有训练样本 $\boldsymbol{x}^{(n)}$（$n = 1, \cdots, N$）的相似度 $\boldsymbol{x}^{(n)T}\boldsymbol{x}$
  - 将所有标签 $y^{(n)}$ 按样本相似度 $\boldsymbol{x}^{(n)T}\boldsymbol{x}$ 和乘子 a_n^* 进行加权求和



为不同的数据实例赋予权重 a_i

计算相似度 $\boldsymbol{x}^{(n)T}\boldsymbol{x}$

原问题与对偶问题比较

- 优化复杂度

原问题

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

$$s.t.: \ y^{(\ell)} \cdot \left(\boldsymbol{w}^T \boldsymbol{x}^{(\ell)} + b\right) \geq 1,$$

$$\text{for } \ell = 1, 2, \cdots, N$$

要优化的参数数量：特征维度

对偶问题

$$\max_{\boldsymbol{a}} \ g(\boldsymbol{a})$$

$$s.t.: \ \boldsymbol{a} \geq \boldsymbol{0}$$

$$\sum_{\ell=1}^{N} a_\ell y^{(\ell)} = 0$$

要优化的参数数量：训练样本数量

在高维特征情况下，求解对偶问题更高效

- 测试复杂度

原问题

$$\hat{y}(\boldsymbol{x}) = sign(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

只需要一次内积运算 $\boldsymbol{w}^{*T}\boldsymbol{x}$

对偶问题

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^*(\boldsymbol{x}^{(n)T}\boldsymbol{x})y^{(n)} + b^*\right)$$

需要 N 次内积运算 $\boldsymbol{x}^{(n)T}\boldsymbol{x}$，其中 $n = 1, 2, \cdots, N$

乍一看，对偶分类器看起来比原分类器昂贵得多

- 幸运的是，可以证明绝大多数 a_n^* 都为 0

拉格朗日乘子 a^* 的稀疏性

- 对于任何凸优化问题，最优解都满足 *KKT 条件*，对于我们这个问题，KKT 条件为：

KKT条件：
1.原始可行性：约束被满足
2.对偶可行性：拉格朗日乘子不小于零
3.互补松弛性：拉格朗日乘子与约束的乘积为零
4.梯度条件：最优点的梯度为零

$$a_n^* \geq 0$$

$$y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + b^*\right) - 1 \geq 0$$

$$\boxed{a_n^*\left[y^{(n)}\left(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + b^*\right) - 1\right] = 0}$$

前两个条件来自原始的原问题和对偶问题

拉格朗日乘子 a^* 的稀疏性

- 从最后一个条件可以看出，只有当 $y^{(n)}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + b^*) = 1$ 时，$a_n^* \neq 0$

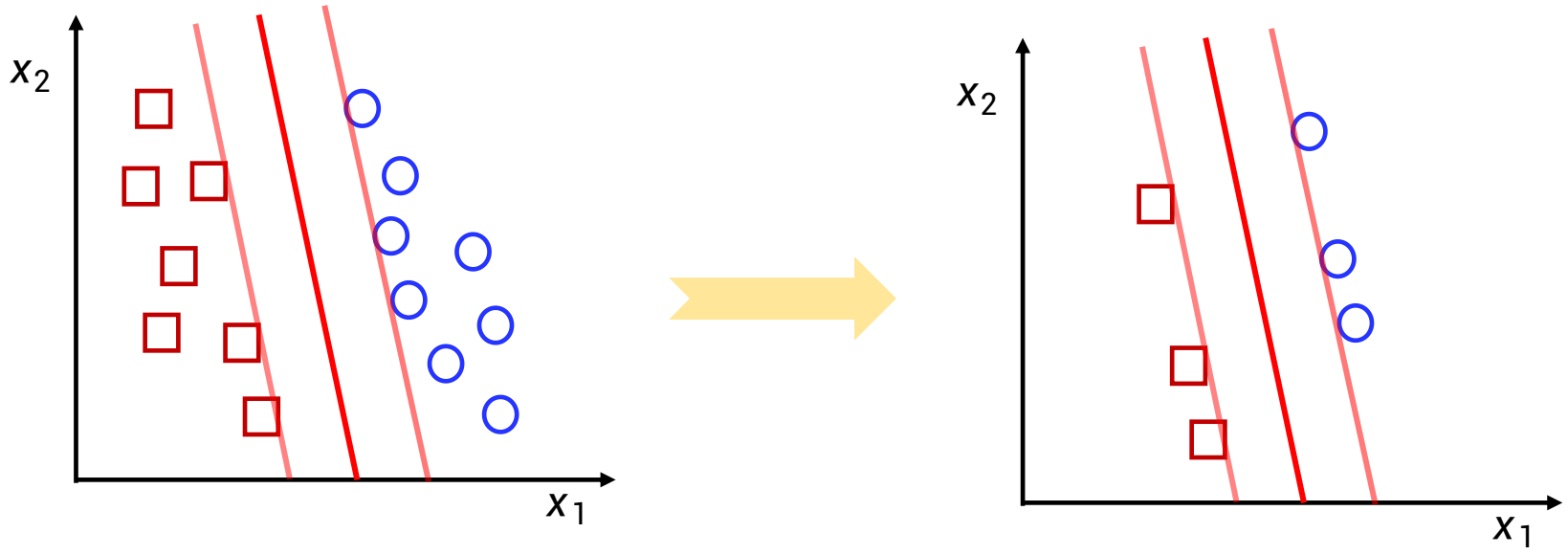
- 如果 $\boldsymbol{x}^{(n)}$ 满足 $y^{(n)}(\boldsymbol{w}^{*T}\boldsymbol{x}^{(n)} + b^*) = 1$，这意味着它位于间隔边界上

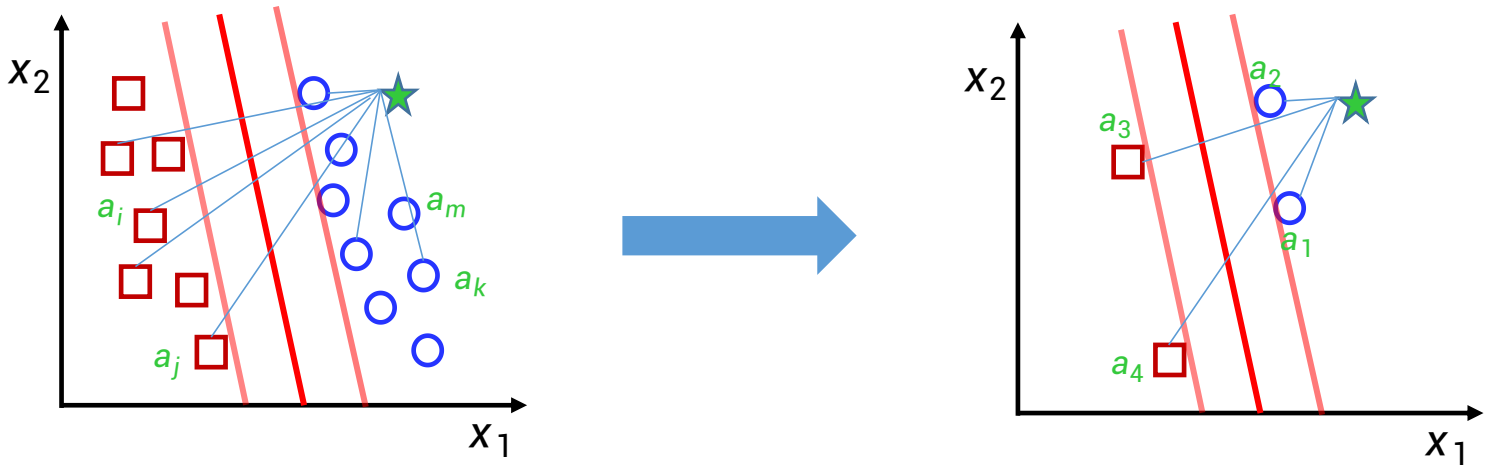
这类样本被称为支持向量

- 因此，当我们对一个未见过的样本 $\boldsymbol{x}$ 进行分类时：

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_n a_n^* y^{(n)} \boldsymbol{x}^{(n)T} \boldsymbol{x} + b^*\right),$$

我们只需要评估 $\boldsymbol{x}$ 与支持向量（样本）之间的相似度 $\boldsymbol{x}^{(n)T}\boldsymbol{x}$





$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N}\left(a_n^*(\boldsymbol{x}^{(n)T}\boldsymbol{x})\right)\cdot y^{(n)} + b^*\right)$$

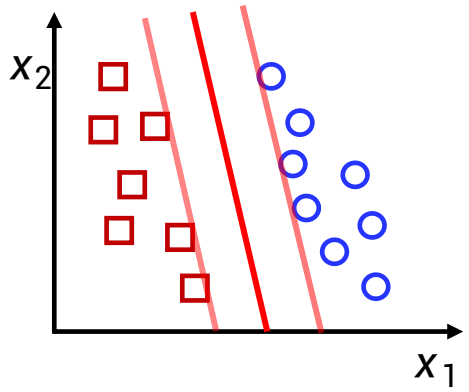
课程大纲

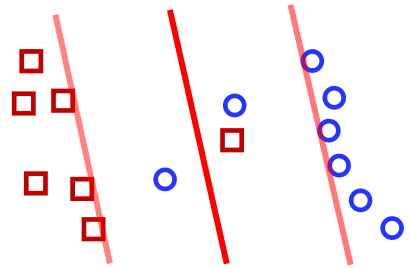
- 线性分类器的决策边界
- 线性最大间隔分类器
- 软线性最大间隔分类器
- 支持向量机
- 与逻辑回归的关系

非线性可分类别

- 之前最大间隔分类器中隐含的假设

训练样本是线性可分的！！！





非线性可分类别

- 在这种情况下，优化问题会发生什么？

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

$$s.t.: \ y^{(n)} \cdot \left(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + b\right) \geq 1, \ \text{for } n = 1, 2, \cdots, N$$

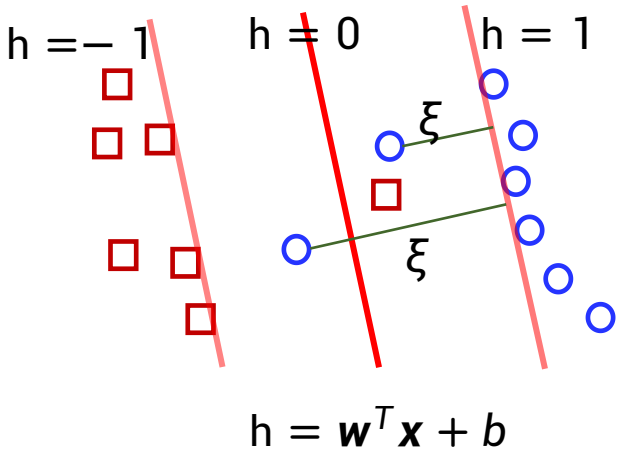
- 该优化问题没有可行解。也就是说，*不存在这样的超平面*

软最大间隔

- 为了解决这个问题，我们不再要求对于所有 $n = 1, \cdots, N$ 都必须满足 $y^{(n)} \cdot (\boldsymbol{w}^T \boldsymbol{x}^{(n)} + b) \geq 1$，而是只要求：

$$y^{(n)} \cdot (\boldsymbol{w}^T \boldsymbol{x}^{(n)} + b) \geq 1 - \xi_n$$

其中 ξ_n 是一个*松弛变量*，并且 $\xi_n \geq 0$



- 目标不再仅仅是最小化 $\frac{1}{2}\|\boldsymbol{w}\|^2$，还需要最小化 ξ_n 的总和，从而得到以下目标函数：

$$\frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N} \xi_n$$

其中 C 用于控制相对重要性

- 现在的优化问题变为：

$$\min_{\boldsymbol{w},\, b,\boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$s.t.: \quad y^{(n)} \cdot \left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + b\right) \geq 1 - \xi_n,$$

$$\xi_n \geq 0, \qquad \text{for } n = 1,\ 2,\ \cdots,\ N$$

- 使用之前相同的方法，可以推导出*对偶形式* 为：

$$\max_{\boldsymbol{a}} \; g(\boldsymbol{a})$$

$$s.t.: \quad a_n \geq 0, \quad a_n \leq C$$

$$\sum_{n=1}^{N} a_n y^{(n)} = 0$$

当 $a_n > C$ 时，可以证明$g(\boldsymbol{a}) = -\infty$

其中 $g(\boldsymbol{a}) = \sum_{n=1}^{N} a_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} \boldsymbol{x}^{(n)T} \boldsymbol{x}^{(m)}$

- 当得到最优的 $\boldsymbol{w}^*$ 和 b^* 时，样本 $\boldsymbol{x}$ 被分类为：

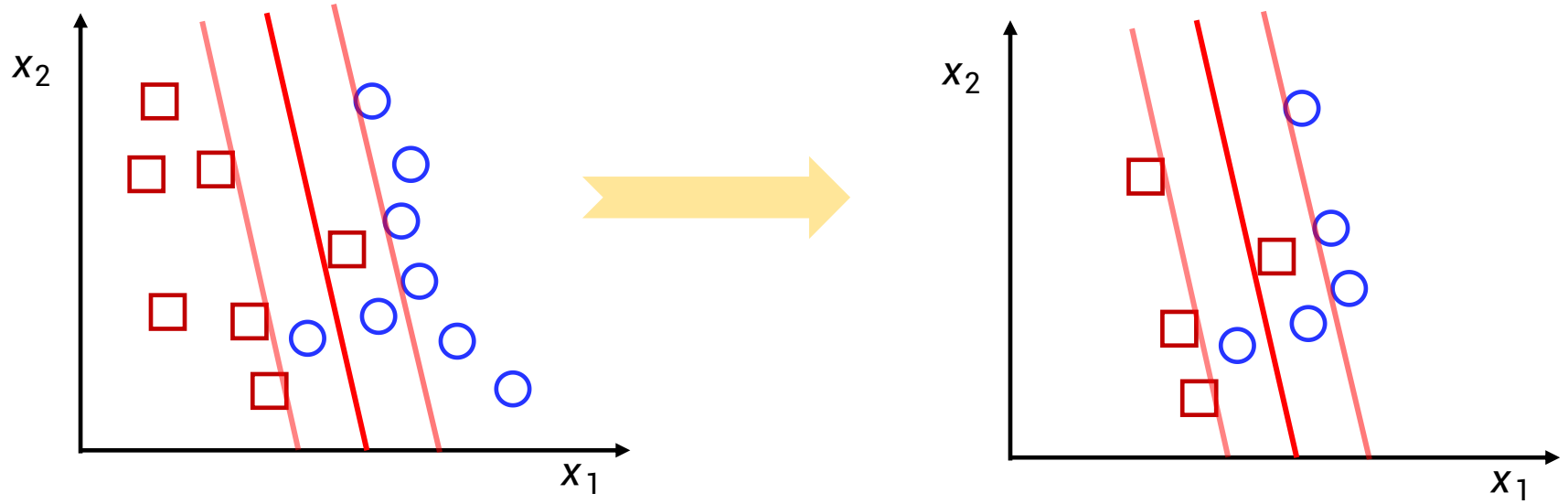
$$\hat{y}(\boldsymbol{x}) = sign(\boldsymbol{w}^{*T}\boldsymbol{x} + b^*)$$

- 当得到最优的 $\boldsymbol{a}^*$ 时，样本 $\boldsymbol{x}$ 被分类为：

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^* y^{(n)} \boldsymbol{x}^{(n)T}\boldsymbol{x} + b^*\right)$$

此外，上面的两个分类器是*等价的*

- 最优解 a^* 是稀疏的，只有间隔边界内的元素是非零的



课程大纲

- 线性分类器的决策边界
- 线性最大间隔分类器
- 软线性最大间隔分类器
- 支持向量机
- 与逻辑回归的关系

非线性化

- 到目前为止，最大间隔分类器仍然是*线性的*

- 为了将模型非线性化，我们可以通过基函数将原始数据 $\boldsymbol{x}$ 变换到特征空间

$$\boldsymbol{\phi}: \boldsymbol{x} \longrightarrow \boldsymbol{\phi}(\boldsymbol{x})$$

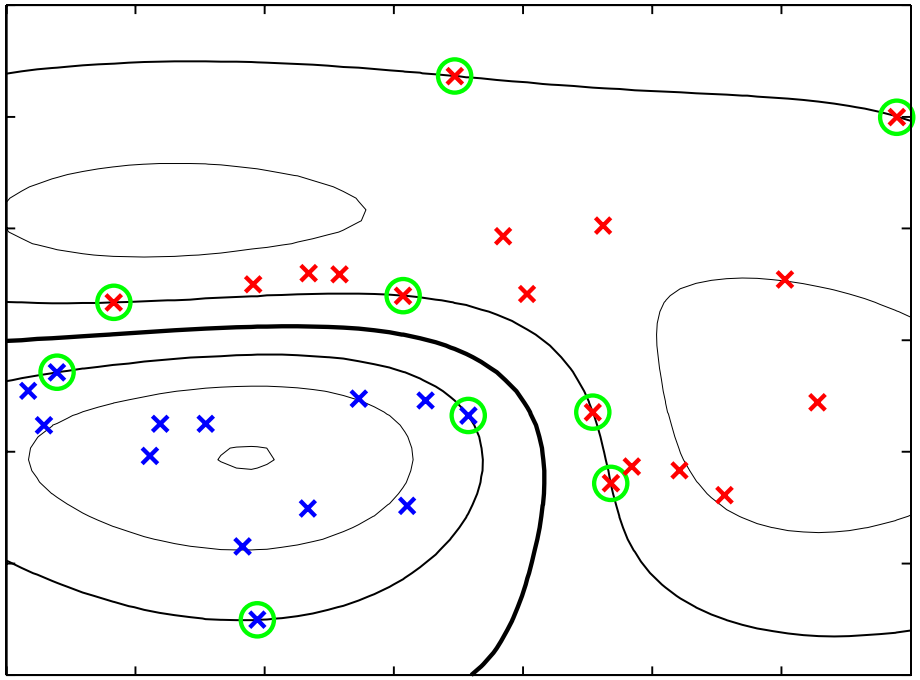
- 那么，原最大间隔优化问题变为：

$$\min_{\boldsymbol{w}, b, \boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$s.t.: \ \ y^{(n)} \cdot \left(\boldsymbol{w}^T \boldsymbol{\phi}\left(\boldsymbol{x}^{(n)}\right) + b\right) \geq 1 - \xi_n,$$

$$\xi_n \geq 0, \qquad \text{for } n = 1, 2, \cdots, N$$

分类器: $\hat{y}(\boldsymbol{x}) = sign(\boldsymbol{w}^{*T}\boldsymbol{\phi}(\boldsymbol{x}^{(n)}) + b^{*})$



- 直观上，数据在高维空间中更容易被分离
- 为了获得更好的性能，我们应该将变换后特征空间 $\boldsymbol{\phi}(\boldsymbol{x}^{(n)})$ *的维度设置得尽可能高*
- 然而，基函数 $\boldsymbol{\phi}(\boldsymbol{x})$ 的维度*不能设置得太高*，因为原问题的计算成本会变得非常昂贵

$$\min_{\boldsymbol{w},\, b,\boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$s.t.: \ \ y^{(n)} \cdot \left(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x}^{(n)}) + b\right) \geq 1 - \xi_n,$$

$$\xi_n \geq 0, \qquad \text{for } n = 1, 2, \cdots, N$$

- 该问题可以通过其对偶形式求解

$$\max_{\boldsymbol{a}} \; g(\boldsymbol{a})$$

$$s.t.: \quad a_n \geq 0, \quad a_n \leq C$$

$$\sum_{n=1}^{N} a_n y^{(n)} = 0$$

其中 $g(\boldsymbol{a}) = \underbrace{\sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} \boldsymbol{\phi}\left(\boldsymbol{x}^{(n)}\right)^T \boldsymbol{\phi}\left(\boldsymbol{x}^{(m)}\right)}_{=\mathbf{1}^T\boldsymbol{a}-\frac{1}{2}\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a}}$

分类器: $\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^* y^{(n)} \boldsymbol{\phi}\left(\boldsymbol{x}^{(n)}\right)^T \boldsymbol{\phi}(\boldsymbol{x}) + b^*\right)$

- $\boldsymbol{a}$ 的*维度与 $\boldsymbol{\phi}(\cdot)$ 的维度无关*，因此对偶形式能够在非常大的特征空间 $\boldsymbol{\phi}(\cdot)$ 中工作

对偶形式需要计算内积

$$\boldsymbol{\phi}\left(\boldsymbol{x}^{(n)}\right)^{T}\boldsymbol{\phi}(\boldsymbol{x}),$$

这在高维情况下计算成本很高

这个问题可以通过使用*核技巧*来解决

核函数

- 核函数是一个双变量函数 $k(\boldsymbol{x}, \boldsymbol{x}')$，可以表示为某个函数 $\boldsymbol{\phi}(\cdot)$ 的内积

$$k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$$

显然，$\boldsymbol{x}^T \boldsymbol{x}'$ 和 $\boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$ 是核函数

- 默瑟定理：如果一个函数 $k(\boldsymbol{x}, \boldsymbol{x}')$是对称正定的，即

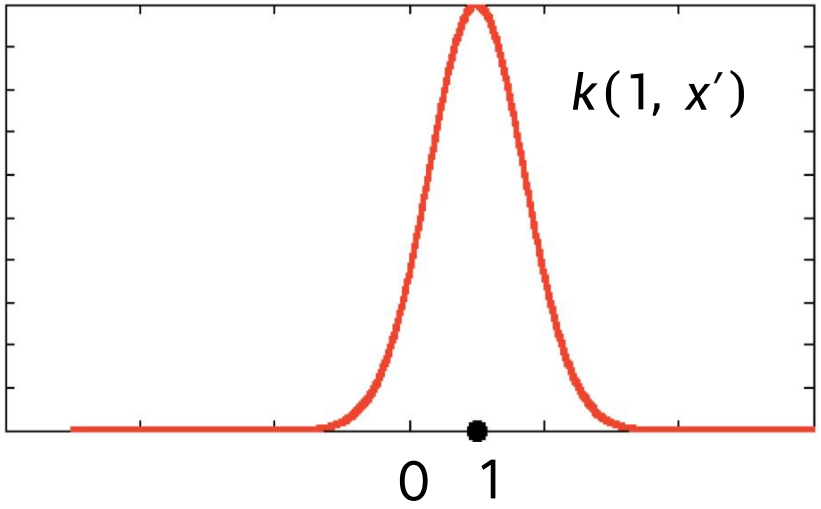
$$\int \int g(\boldsymbol{x}) k(\boldsymbol{x}, \boldsymbol{y}) g(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \geq 0, \quad \forall\, g(\cdot) \in L^2,$$

那么必然存在一个函数 $\boldsymbol{\phi}(\cdot)$ 使得 $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T \boldsymbol{\phi}(\boldsymbol{x}')$

如果一个函数 $k(\boldsymbol{x}, \boldsymbol{x}')$ 满足对称正定条件，那么它必然是一个核函数

- 最广泛使用的核函数之一是高斯核，其形式为：

$$k(\boldsymbol{x},\ \boldsymbol{x}') = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x}-\boldsymbol{x}'\|^2\right\}$$



- 高斯核的函数 $\boldsymbol{\phi}(\cdot)$ 具有无限维度

$$\boldsymbol{\phi}(x) = e^{-x^2/2\sigma^2}\left[1, \sqrt{\frac{1}{1!\sigma^2}}x, \sqrt{\frac{1}{2!\sigma^4}}x^2, \sqrt{\frac{1}{3!\sigma^6}}x^3, \cdots\right]^T$$

核技巧

- 借助核函数，对偶最大间隔分类器可以等价地重写为：

$$\max_{\boldsymbol{a}} \; g(\boldsymbol{a})$$

$$s.t.: \quad a_n \geq 0, \quad a_n \leq C$$

$$\sum_{n=1}^{N} a_n y^{(n)} = 0$$

$$\text{其中 } g(\boldsymbol{a}) = \underbrace{\sum_{n=1}^{N} a_n - \frac{1}{2}\sum_{n=1}^{N}\sum_{m=1}^{N} a_n a_m y^{(n)} y^{(m)} k\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}^{(m)}\right)}_{=\mathbf{1}^T\boldsymbol{a} - \frac{1}{2}\boldsymbol{a}^T\boldsymbol{M}\boldsymbol{a}}$$

- 导出的分类器

$$\hat{y}(\boldsymbol{x}) = sign\left(\sum_{n=1}^{N} a_n^* y^{(n)} k\left(\boldsymbol{x}^{(n)}, \boldsymbol{x}\right) + b^*\right)$$

核技巧：用核函数 $k(\boldsymbol{x}, \boldsymbol{x}')$ 替代 $\boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$

- 结论可以总结如下：

  - 如果 $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{x}^T\boldsymbol{x}'$，这是一个线性最大间隔分类器

  - 如果 $k(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\phi}(\boldsymbol{x})^T\boldsymbol{\phi}(\boldsymbol{x}')$，这是一个基于基函数的*有限维*非线性最大间隔分类器

  - 如果 $k(\boldsymbol{x}, \boldsymbol{x}') = \exp\left\{-\frac{1}{2\sigma^2}\|\boldsymbol{x} - \boldsymbol{x}'\|^2\right\}$，这是一个*无限维*非线性最大间隔分类器

课程大纲

- 线性分类器的决策边界
- 线性最大间隔分类器
- 软线性最大间隔分类器
- 支持向量机
- 与逻辑回归的关系

- 在逻辑回归中，我们最小化损失：

$$L(\boldsymbol{w}, b) = -\sum_{n=1}^{N}\left[\tilde{y}^{(n)}\log\sigma(\mathrm{h}^{(n)}) + (1-\tilde{y}^{(n)})\log\left(1-\sigma(\mathrm{h}^{(n)})\right)\right] + \lambda\|\boldsymbol{w}\|^2$$

$$= \sum_{n=1}^{N}\log\left(1+\exp\left(-y^{(n)}\mathrm{h}^{(n)}\right)\right) + \lambda\|\boldsymbol{w}\|^2$$

$$= \sum_{n=1}^{N} E_{LR}(y^{(n)}\mathrm{h}^{(n)}) + \lambda\|\boldsymbol{w}\|^2$$

注：$\tilde{y} \in \{0, 1\}$, $y \in \{-1, 1\}$

其中，$E_{LR}(z) = \log(1+\exp(-z))$

- 在理想分类器中，我们最小化损失

$$L(\boldsymbol{w}, b) = \sum_{n=1}^{N} E_{Ideal}\left(y^{(n)} \mathrm{h}^{(n)}\right) + \lambda \|\boldsymbol{w}\|^2$$

注：$y \in \{-1, 1\}$，
$\mathrm{h} \in \{-\infty, \infty\}$

其中，如果 $z \geq 0$， $E_{Ideal}(z) = 0$；否则为 1

公式定义： $E_{Ideal}(z) = 0 \; if \; z \geq 0 \; else \; 1$

阶跃函数，不可导

- 回顾“线性最大间隔分类器”：

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

$$s.t.: \ y^{(n)} \cdot \left(\boldsymbol{w}^T \boldsymbol{x}^{(n)} + b\right) \geq 1, \ \text{for } n = 1, 2, \cdots, N$$

- 在线性最大间隔分类器中，我们等价地最小化损失：

$$L(\boldsymbol{w}, b) = \sum_{n=1}^{N} E_\infty\left(y^{(n)} \mathrm{h}^{(n)} - 1\right) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

其中，如果 $z \geq 0$，则 $E_\infty(z) = 0$；否则为 $+\infty$

硬约束：所有样本都满足约束，否则损失无穷大

公式定义： $E_\infty(z) = 0 \ if \ z \geq 0 \ else \ +\infty$

也为阶跃函数，可以发现仍然不可导

- 回顾“软线性最大间隔分类器”：

$$\min_{\boldsymbol{w},\, b,\boldsymbol{\xi}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{n=1}^{N}\xi_n$$

$$s.t.: \quad y^{(n)} \cdot \left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + b\right) \geq 1 - \xi_n,$$

$$\xi_n \geq 0, \qquad \text{for } n = 1,\ 2,\ \cdots,\ N$$

- 从约束推导损失函数：

$$\xi_n \geq 1 - y^{(n)} \cdot \left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + b\right),$$

$$\xi_n \geq 0,$$

$$\xi_n = \max\left(0,\ 1 - y^{(n)} \cdot \left(\boldsymbol{w}^T\boldsymbol{x}^{(n)} + b\right)\right)$$

注：$\max(0,\ 1 - z)$

- 在软线性最大间隔分类器中，我们等价地最小化损失：

$$L(\boldsymbol{w}, b) = C\sum_{n=1}^{N} E_{SV}(y^{(n)}\mathrm{h}^{(n)}) + \frac{1}{2}\|\boldsymbol{w}\|^2$$

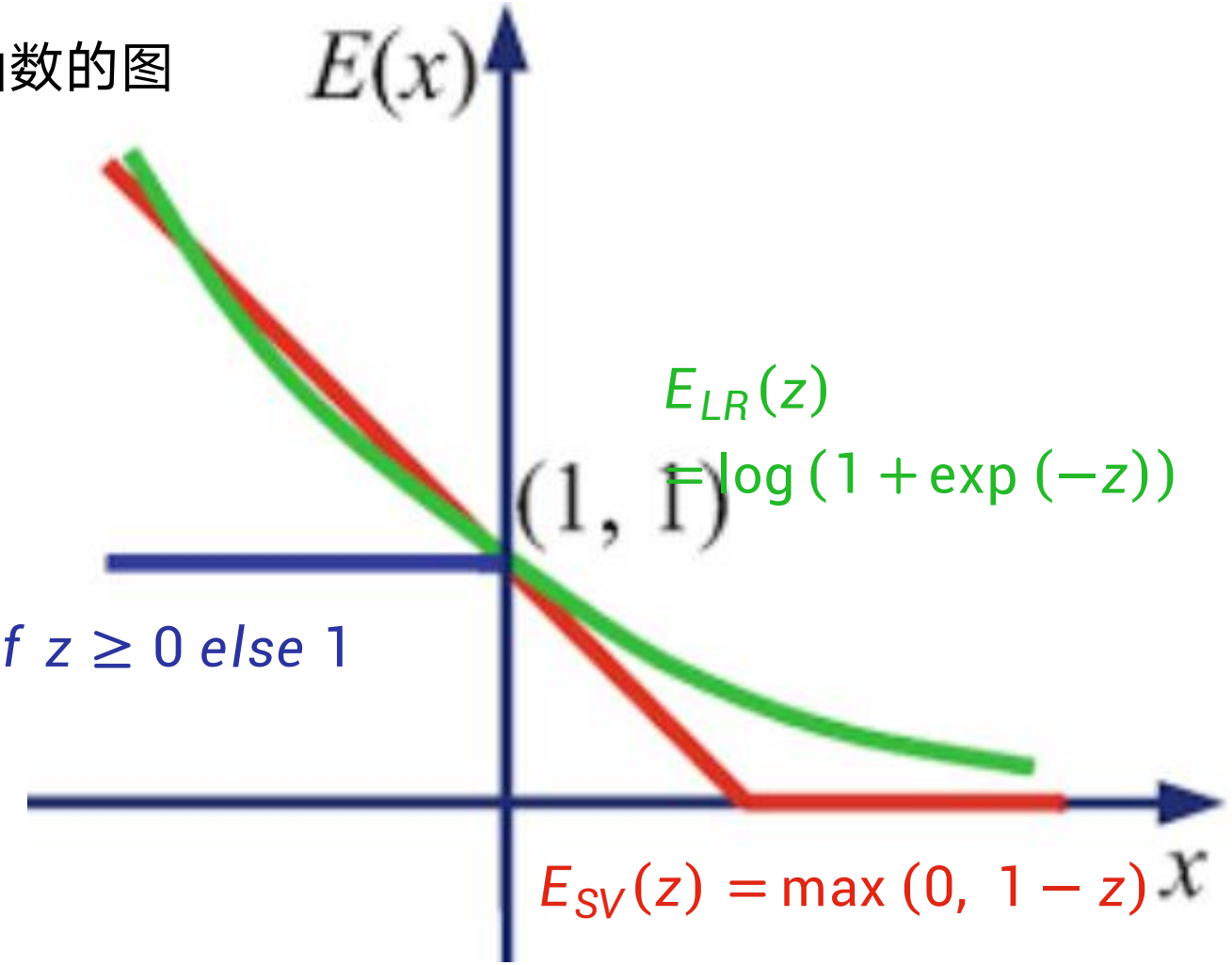
$$= \sum_{n=1}^{N} E_{SV}(y^{(n)}\mathrm{h}^{(n)}) + \lambda\|\boldsymbol{w}\|^2$$

其中，$E_{SV}(z) = \max(0,\ 1-z)$，这被称为*合页损失（hinge loss）*

理想损失像“台阶”——对了 0，错了 1，不可导;
而在软线性最大间隔分类器里，我们用 hinge 函数把它磨平：
max(0,1−z) —— *训练时可导，测试时近似理想*

- 我们可以看到，这四种分类器可以在同一个框架下进行建模，唯一的区别在于所选择的误差函数

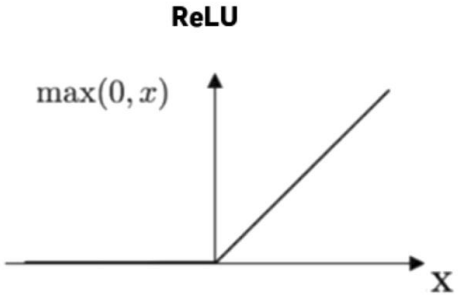
- 四种误差函数的图



$E_{LR}(z) = \log(1 + \exp(-z))$

$E_{Ideal}(z) = 0 \; if \; z \geq 0 \; else \; 1$

$E_{SV}(z) = \max(0, \; 1 - z)$



注：$E_{SV}(z)$类似 ReLU
训练时可导：
可以使用次梯度求导

$$f'(x) = \begin{cases} 1 & if \; x > 0 \\ 0 & if \; x \leq 0 \end{cases}$$