

自然语言处理

第5讲：支持向量机

刘洋



内容提要

分类问题

支持向量机

松弛变量

核函数

一维分类问题

数据	类别
0	负
1	负
3	正
4	正

表1：一维分类数据集。

表1给出了一个一维分类数据集。输入数据点是一个实数。对于每个数据点 x ，可以标上类别 y 。表1中共有两个类别：正 ($y = +1$) 和负 ($y = -1$)。

给定一个新的数据点5，其类别应该是正还是负？

我们可以建立一个分类器：

$$f(x) = \text{sign}(x - 2)$$

其中， x 表示输入数据。当 $x \geq 0$ 时， $\text{sign}(x) = +1$ 。当 $x < 0$ 时， $\text{sign}(x) = -1$ 。

显然，该分类器可以拟合训练集。由于 $f(5) = +1$ ，因此数据点5的类别应该为正。

一维分类问题的图形化表示

数据	类别
0	负
1	负
3	正
4	正

表1：一维分类数据集。



图1：一维特征空间中的数据点。

二维分类问题

- 在二维空间中将属于不同类别的数据点进行分隔

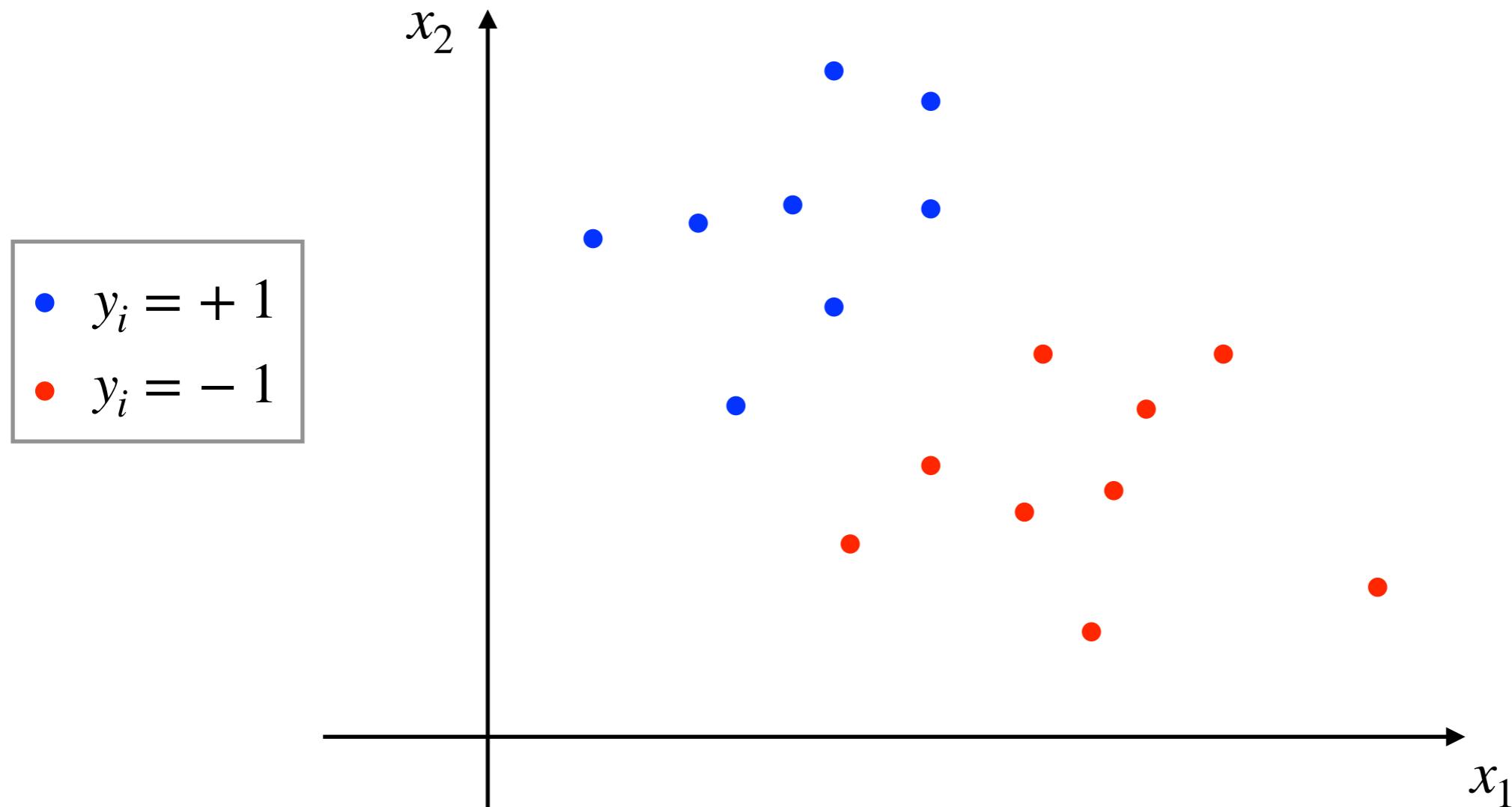


图2：二维空间中的数据点。

分类平面

- 分类平面是指分类器的决策边界，能够将不同类别的数据点分开。

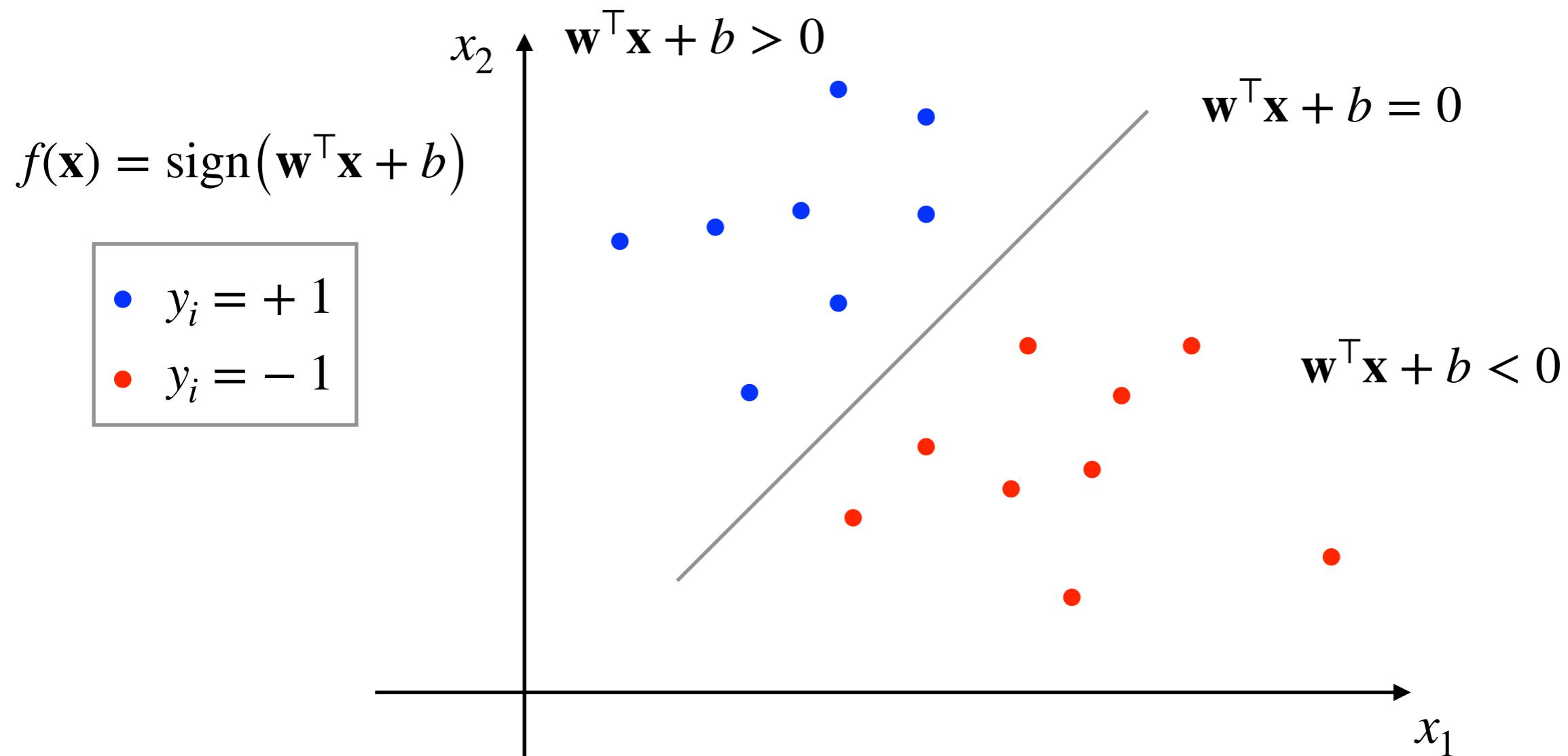


图3：二维空间中的分类平面。

分类平面的选择

- 对于一个数据集，存在着多种不同的分类平面，如何选择？

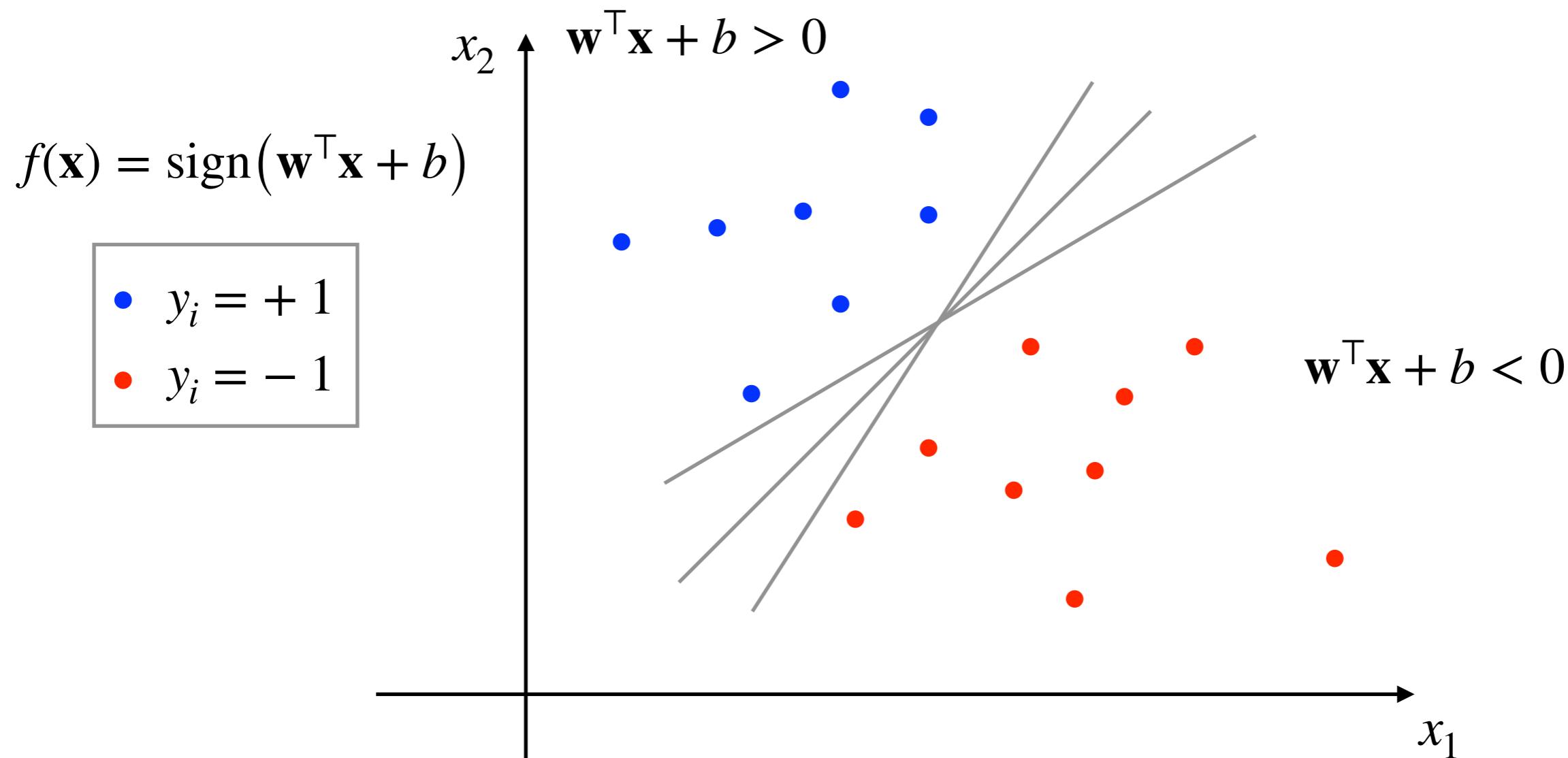


图4：多种分类平面。

分类距离：函数距离

- 函数距离 $y_i(\mathbf{w}^\top \mathbf{x} + b)$ 的正负可以表示分类的正确性和信心。

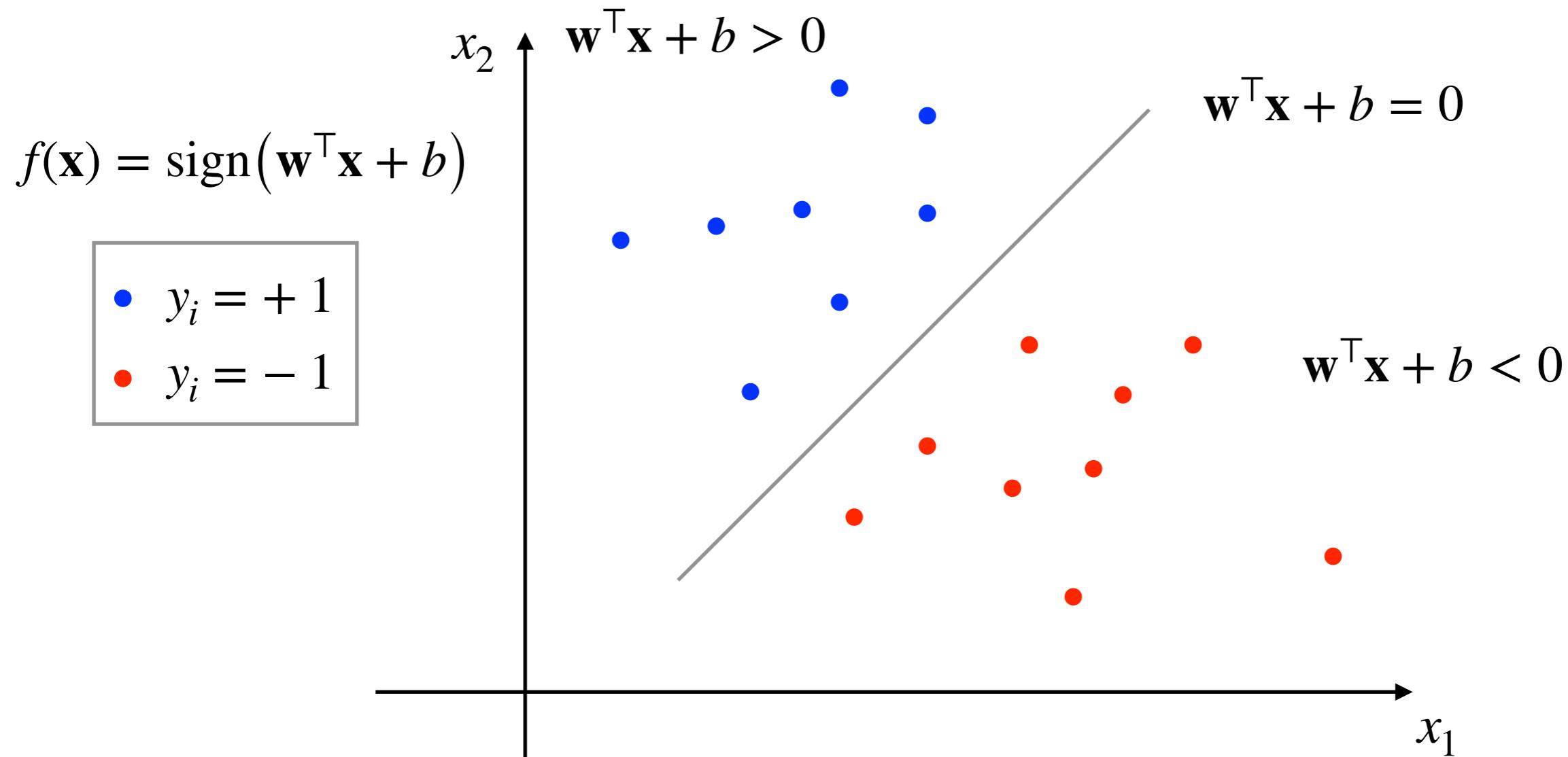


图5：函数距离。

分类距离：几何距离

- 几何距离 $|\mathbf{w}^\top \mathbf{x}_i + b| / \|\mathbf{w}\|^2$ 表示样本点 \mathbf{x}_i 到分类平面的距离。

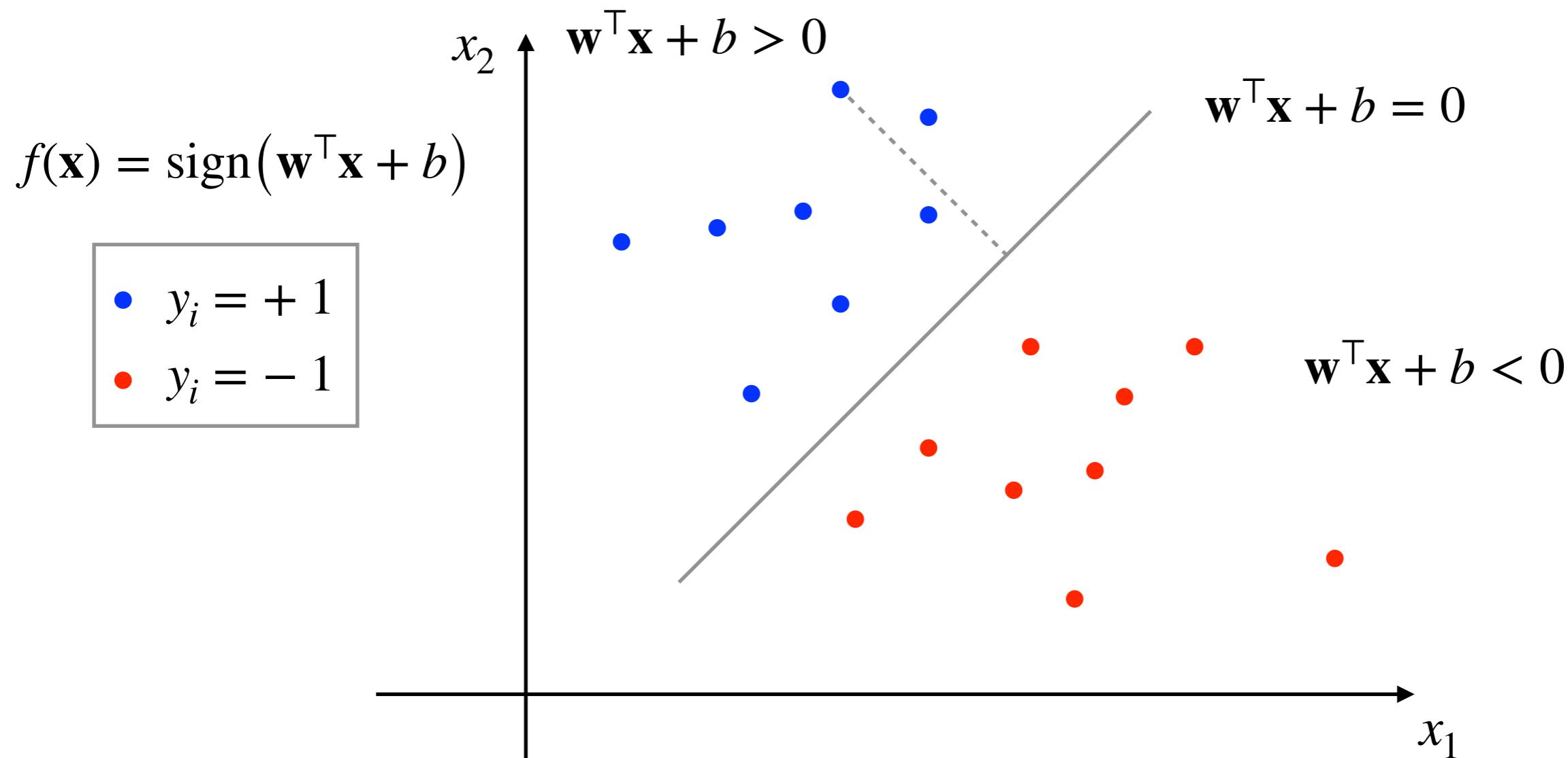


图6：几何距离。

支持向量

- 与分类平面距离最近的样本点称为支持向量，进而构成支持平面。

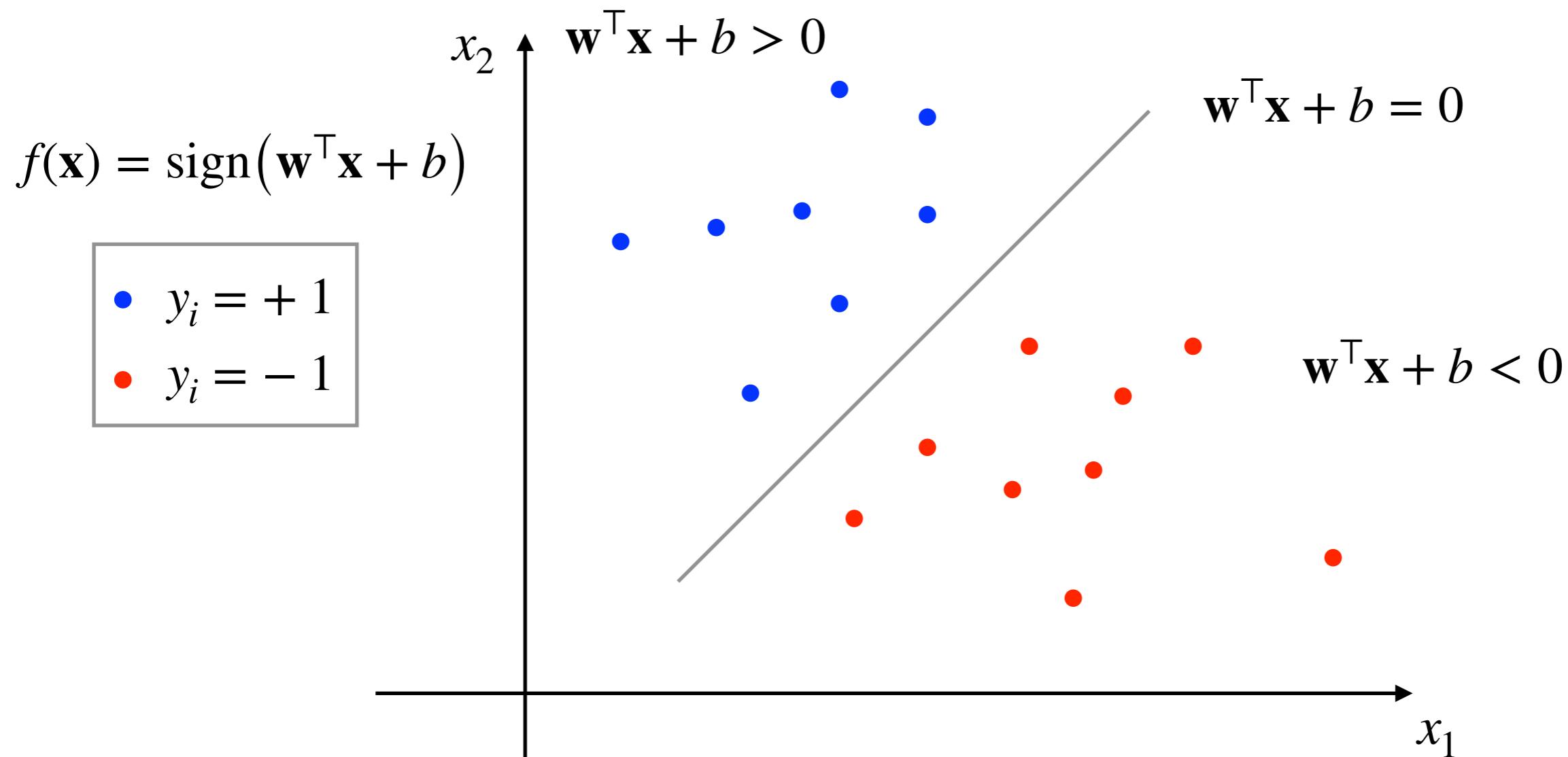


图7：支持向量与支持平面。

支持向量

- 与分类平面距离最近的样本点称为支持向量，进而构成支持平面。

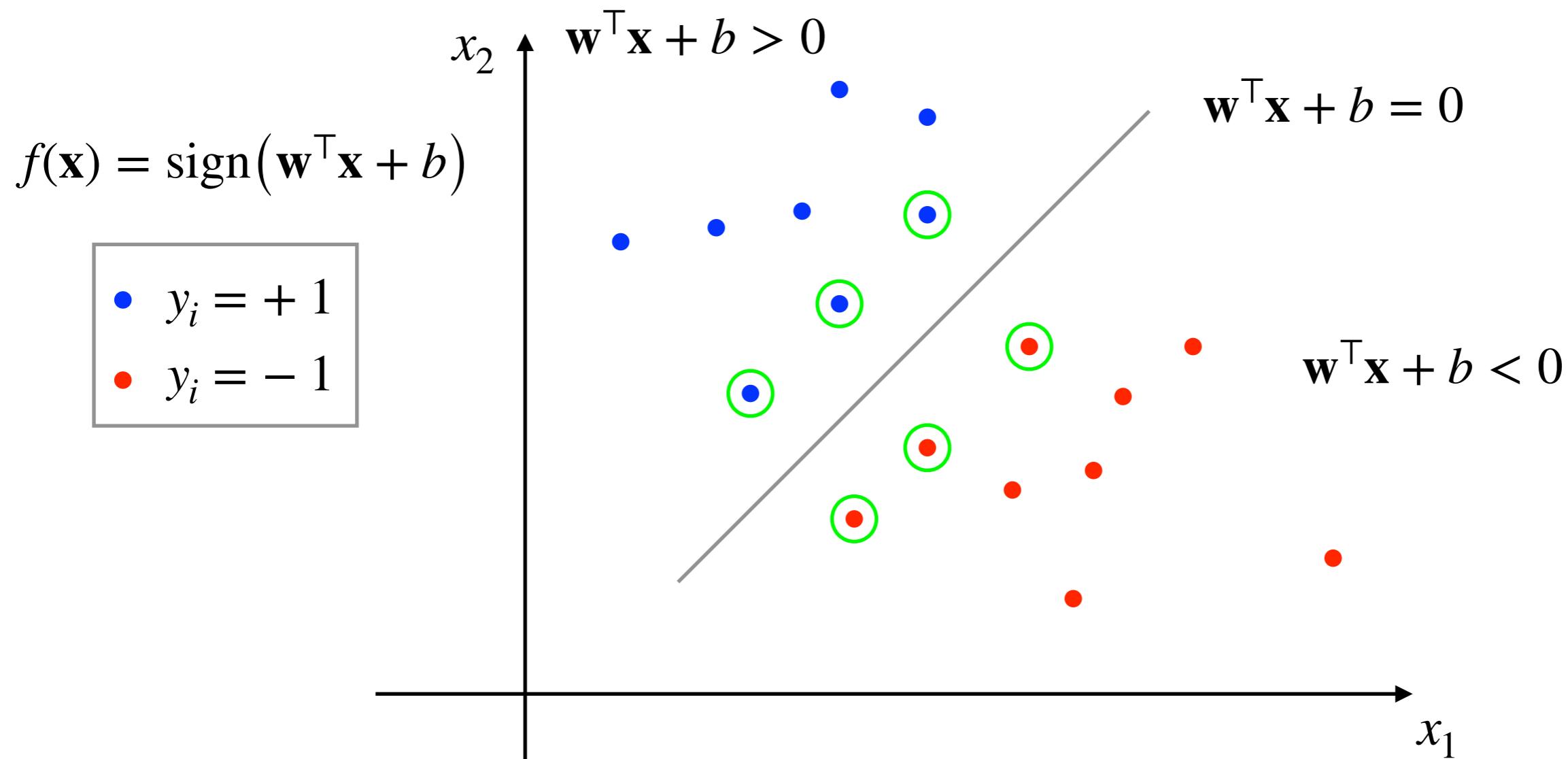


图7：支持向量与支持平面。

支持向量

- 与分类平面距离最近的样本点称为支持向量，进而构成支持平面。

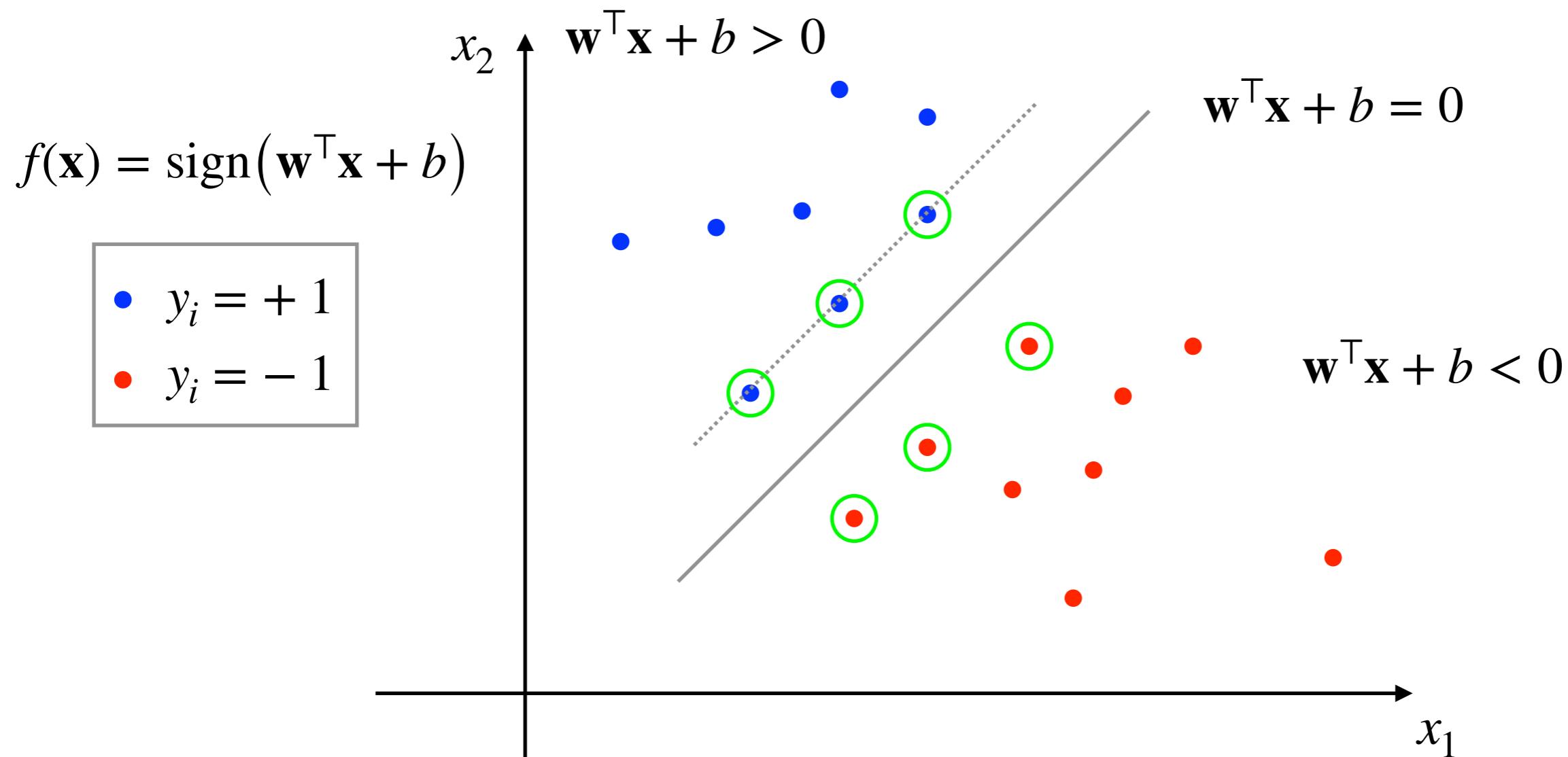


图7：支持向量与支持平面。

支持向量

- 与分类平面距离最近的样本点称为支持向量，进而构成支持平面。

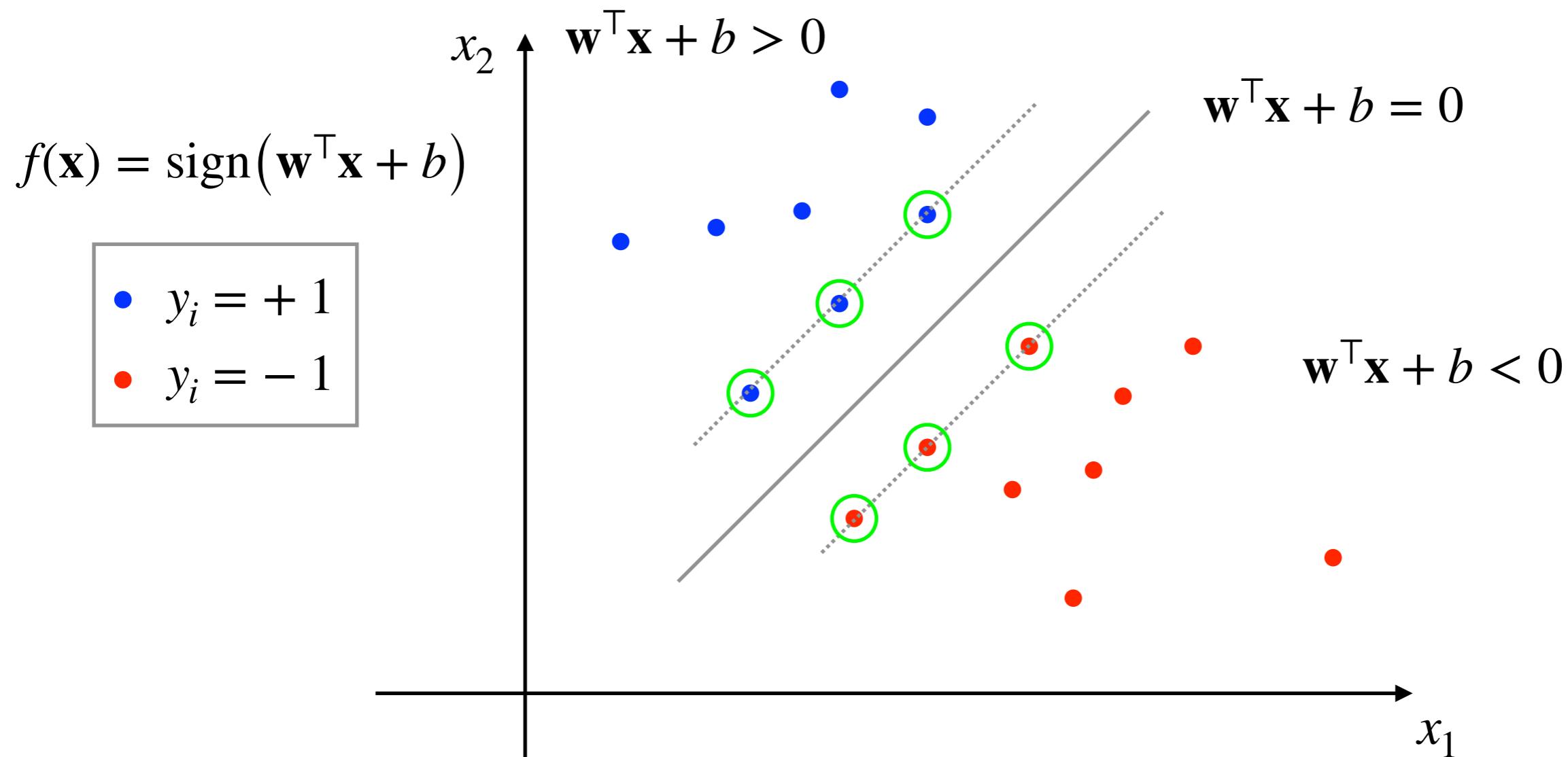


图7：支持向量与支持平面。

分类间距

- 分类器的分类间距 ρ 指的是支持平面之间的距离

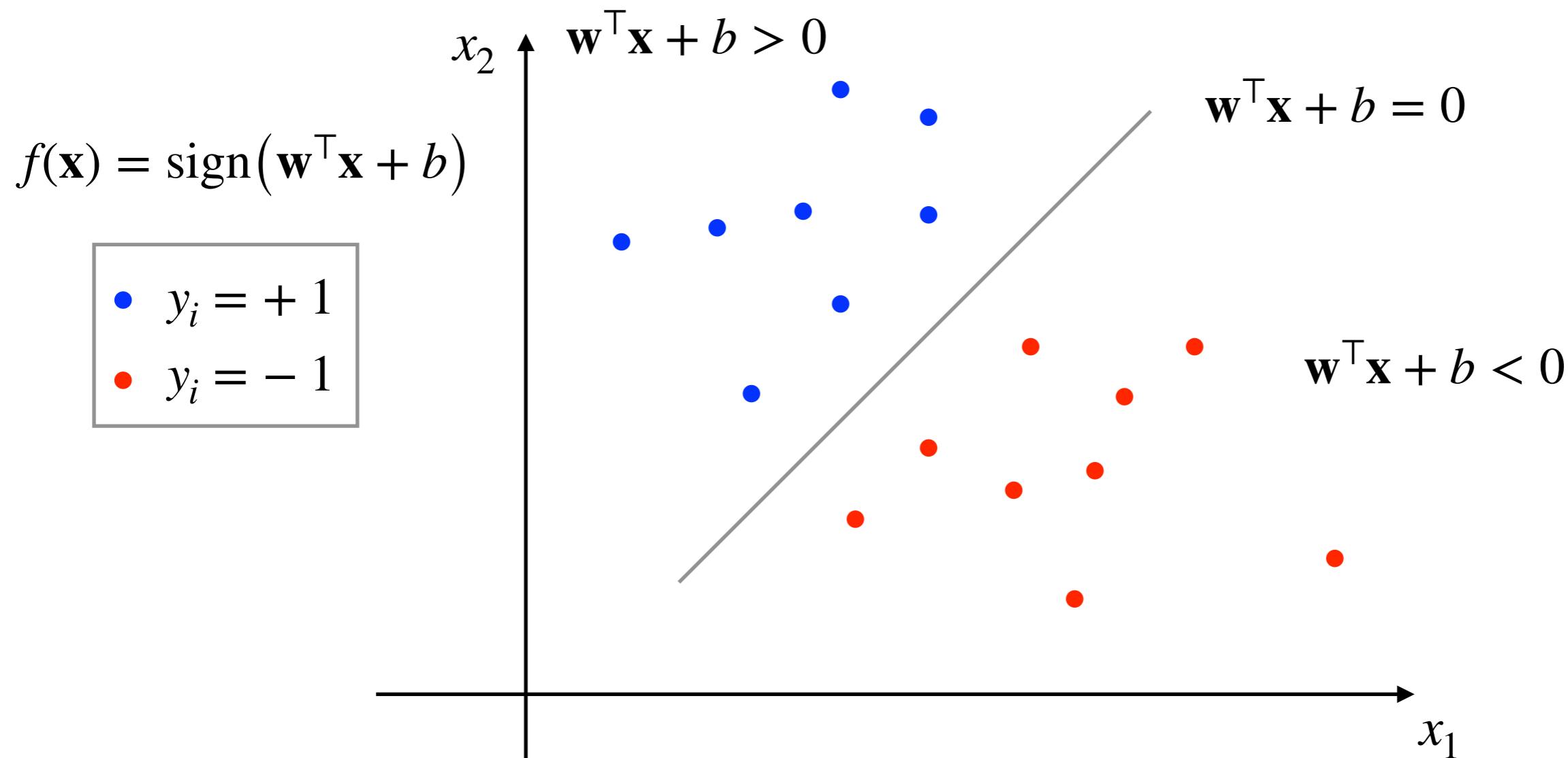


图7：分类间距。

分类间距

- 分类器的分类间距 ρ 指的是支持平面之间的距离

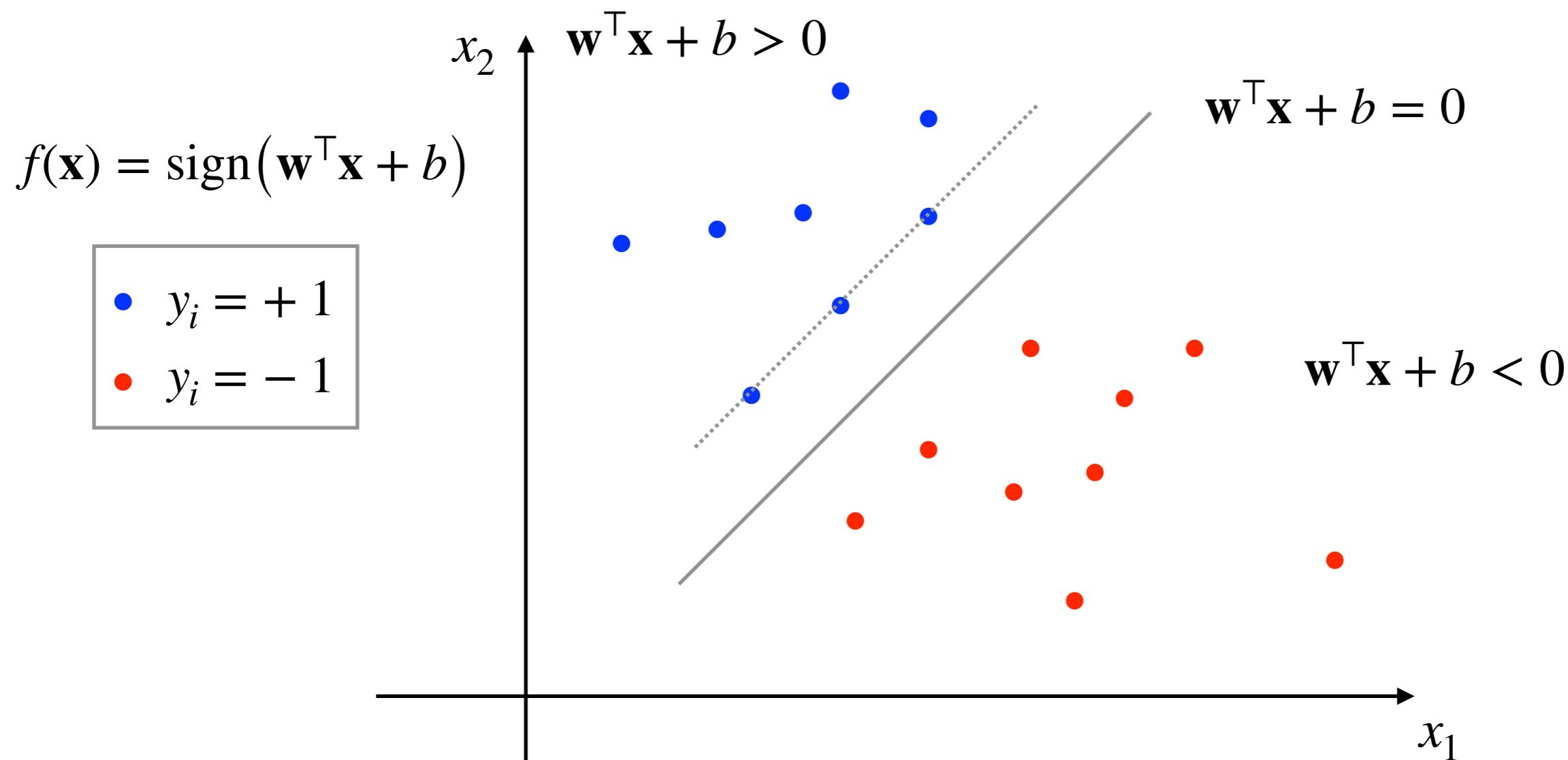


图7：分类间距。

分类间距

- 分类器的分类间距 ρ 指的是支持平面之间的距离

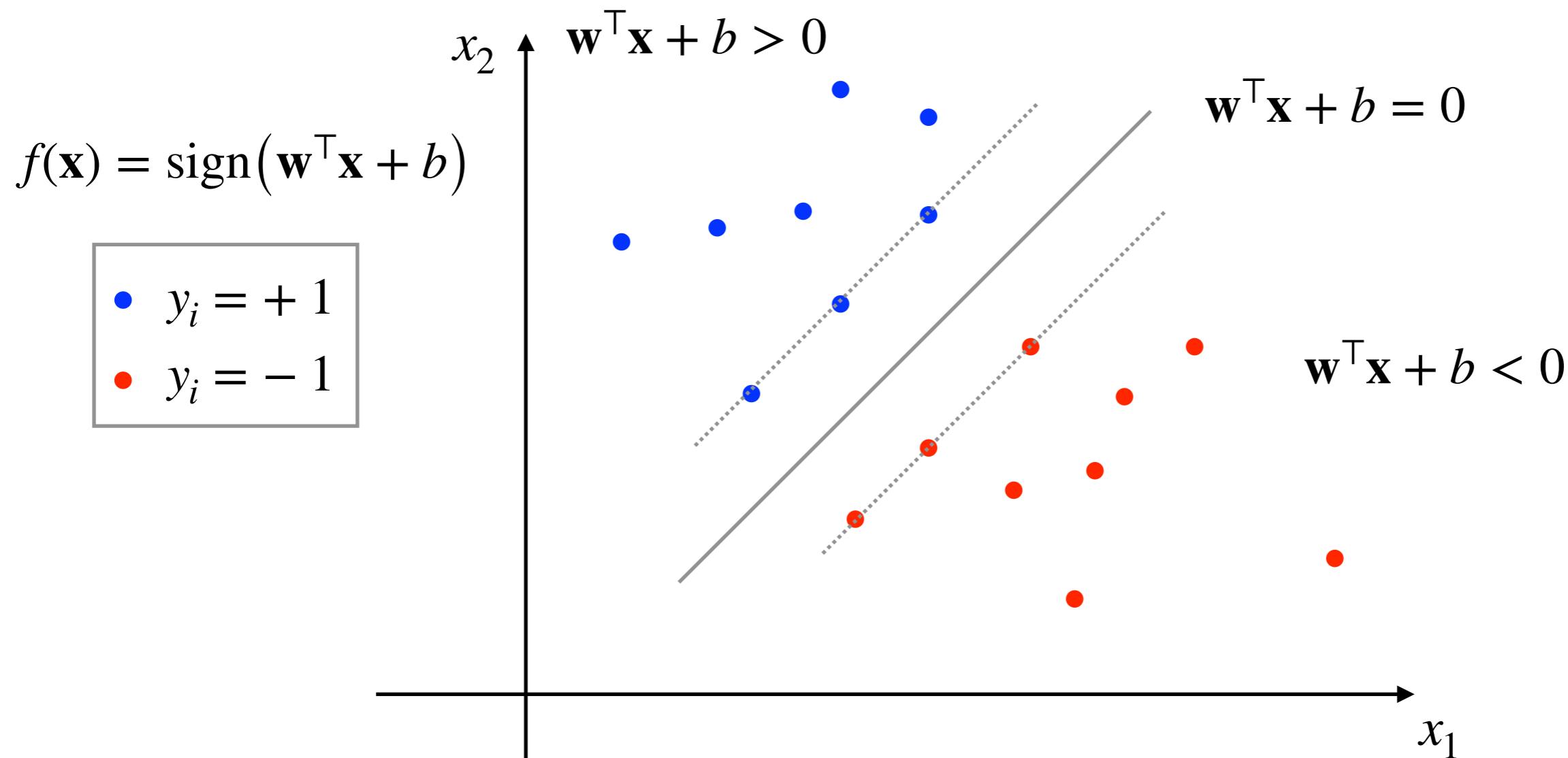


图7：分类间距。

分类间距

- 分类器的分类间距 ρ 指的是支持平面之间的距离

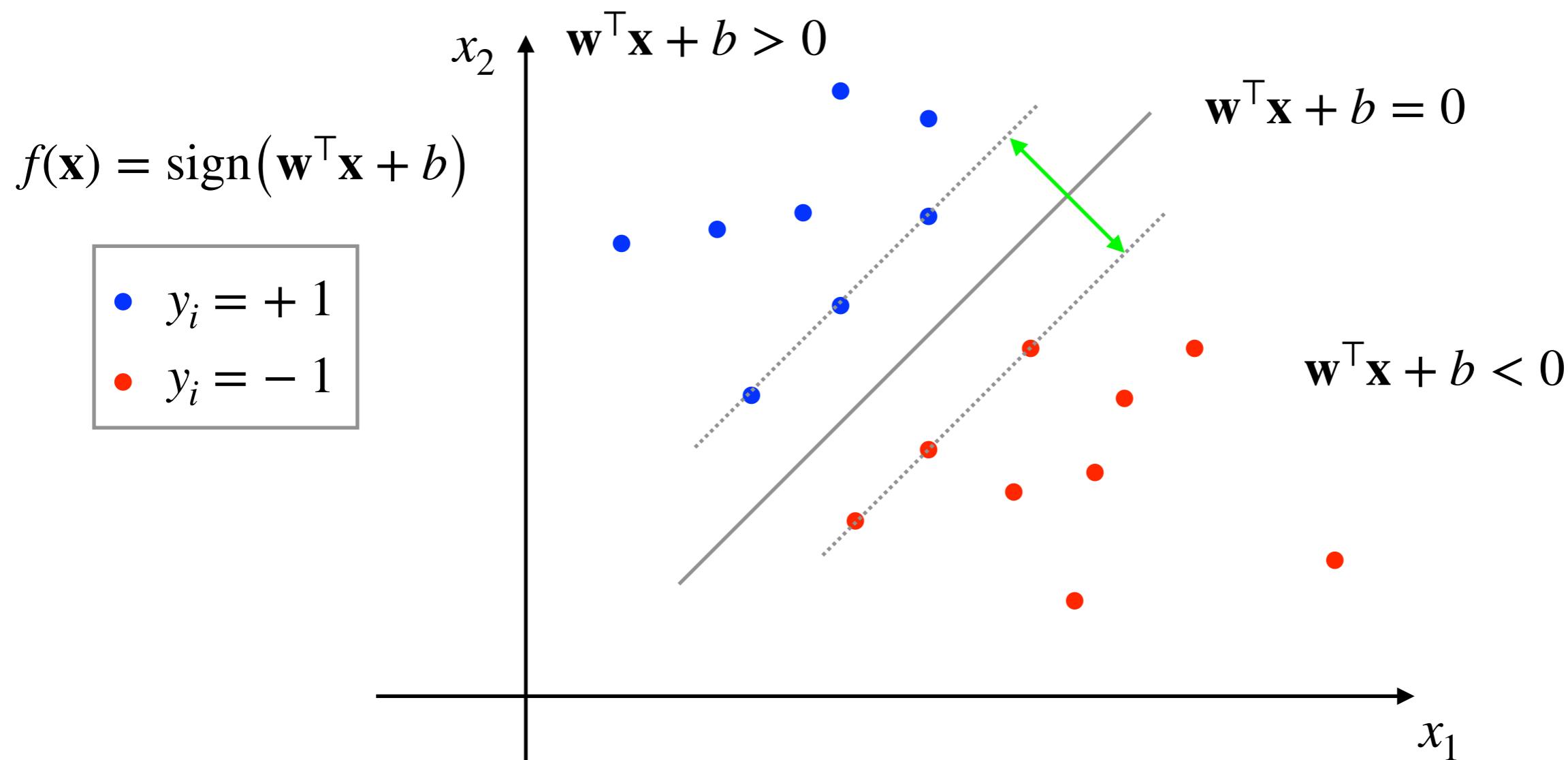


图7：分类间距。

分类间距

- 分类器的分类间距 ρ 指的是支持平面之间的距离

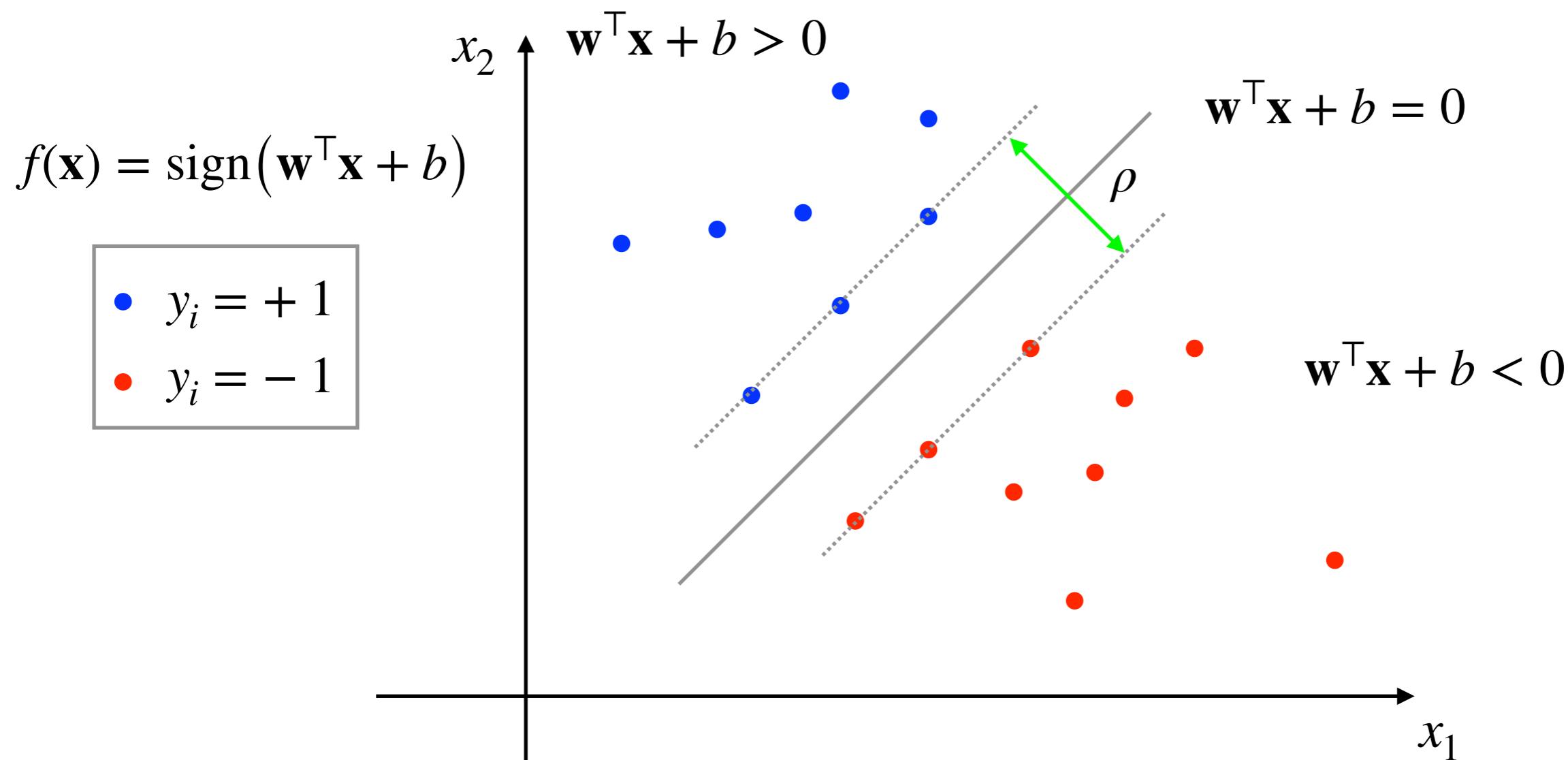


图7：分类间距。

内容提要

分类问题

支持向量机

松弛变量

核函数

支持向量机

- 支持向量机的核心思想：最大化分类间距 ρ 。

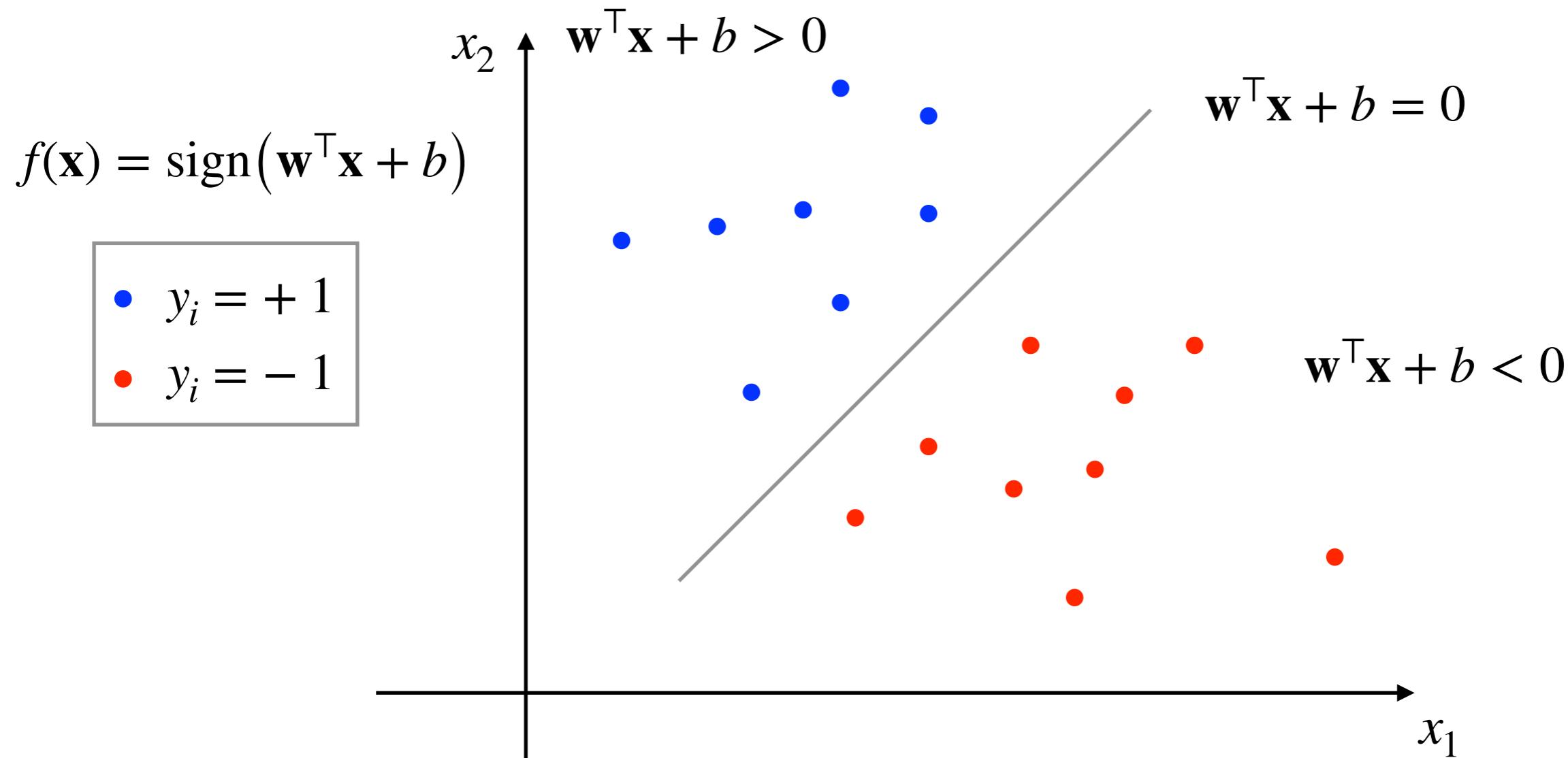


图7：分类间距。

支持向量机

- 支持向量机的核心思想：最大化分类间距 ρ 。

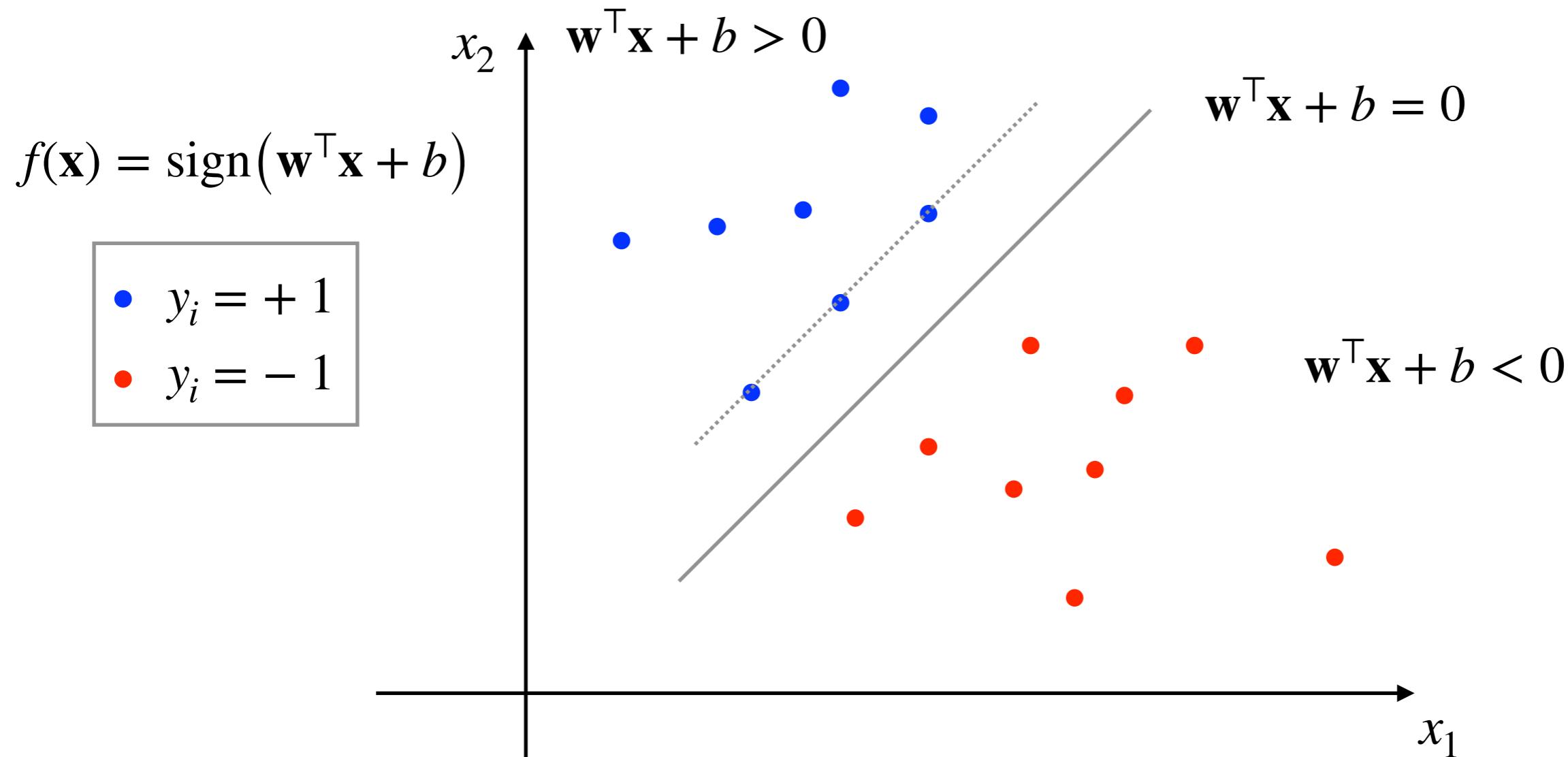


图7：分类间距。

支持向量机

- 支持向量机的核心思想：最大化分类间距 ρ 。

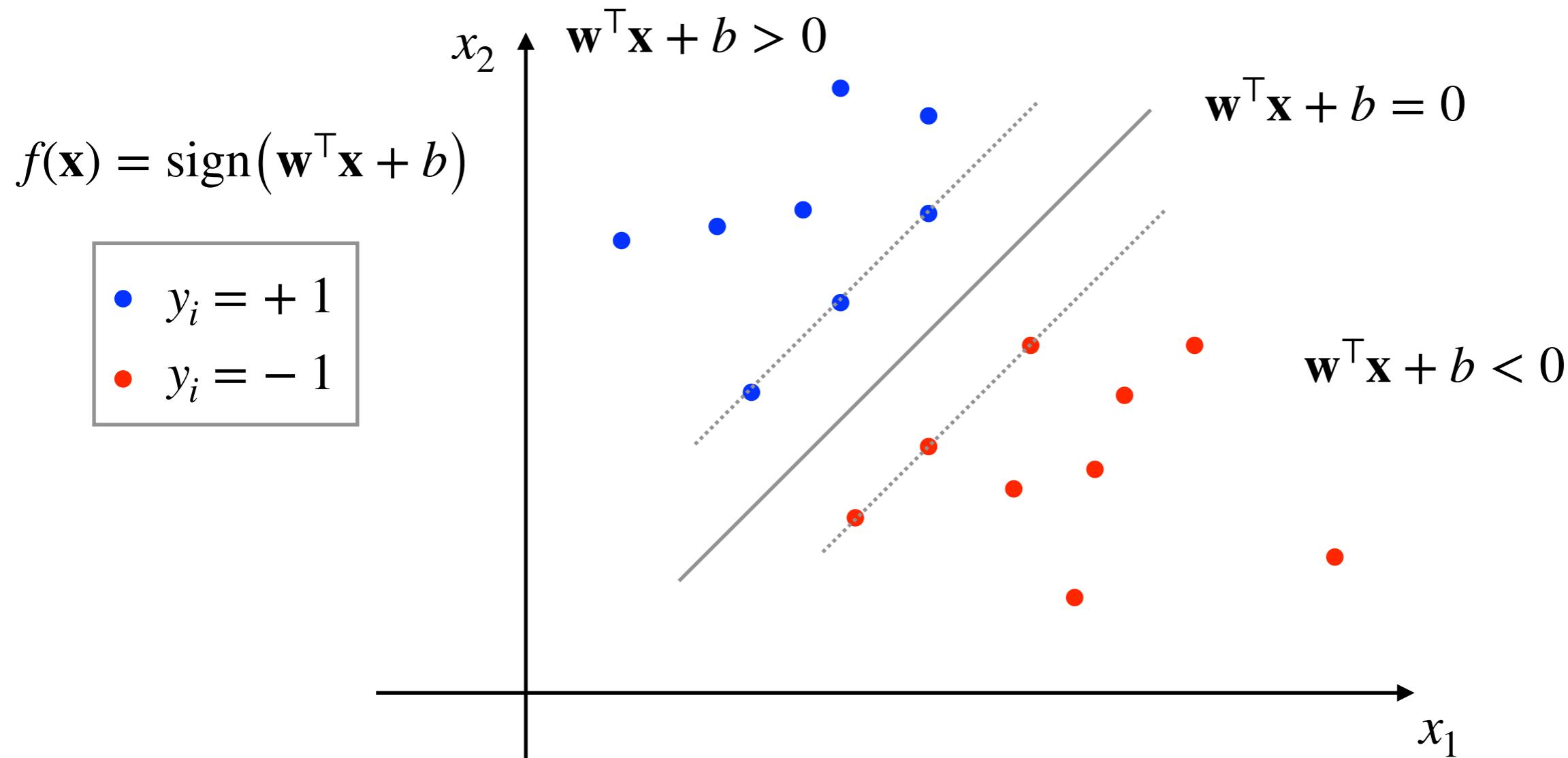


图7：分类间距。

支持向量机

- 支持向量机的核心思想：最大化分类间距 ρ 。

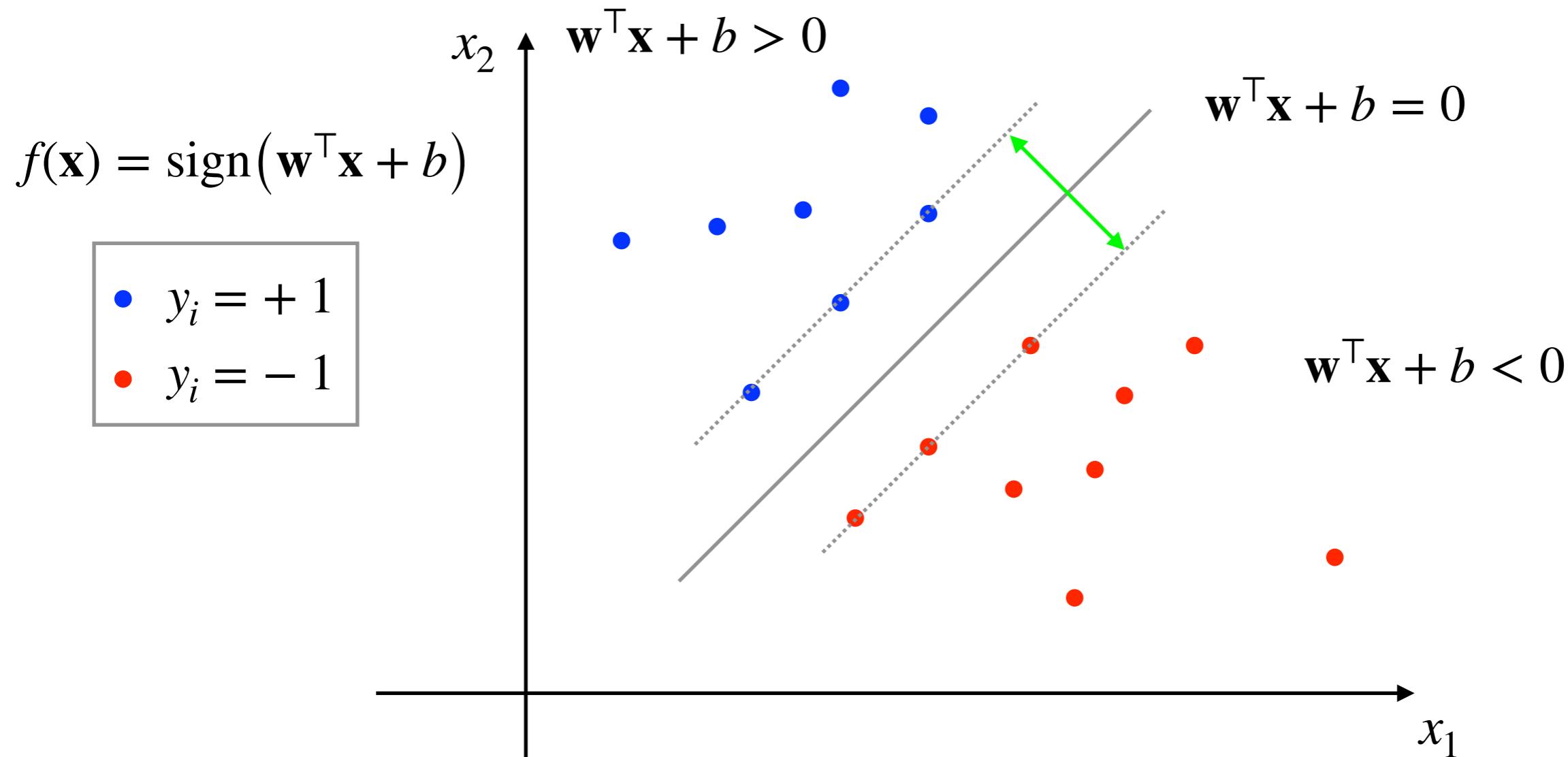


图7：分类间距。

支持向量机

- 支持向量机的核心思想：最大化分类间距 ρ 。

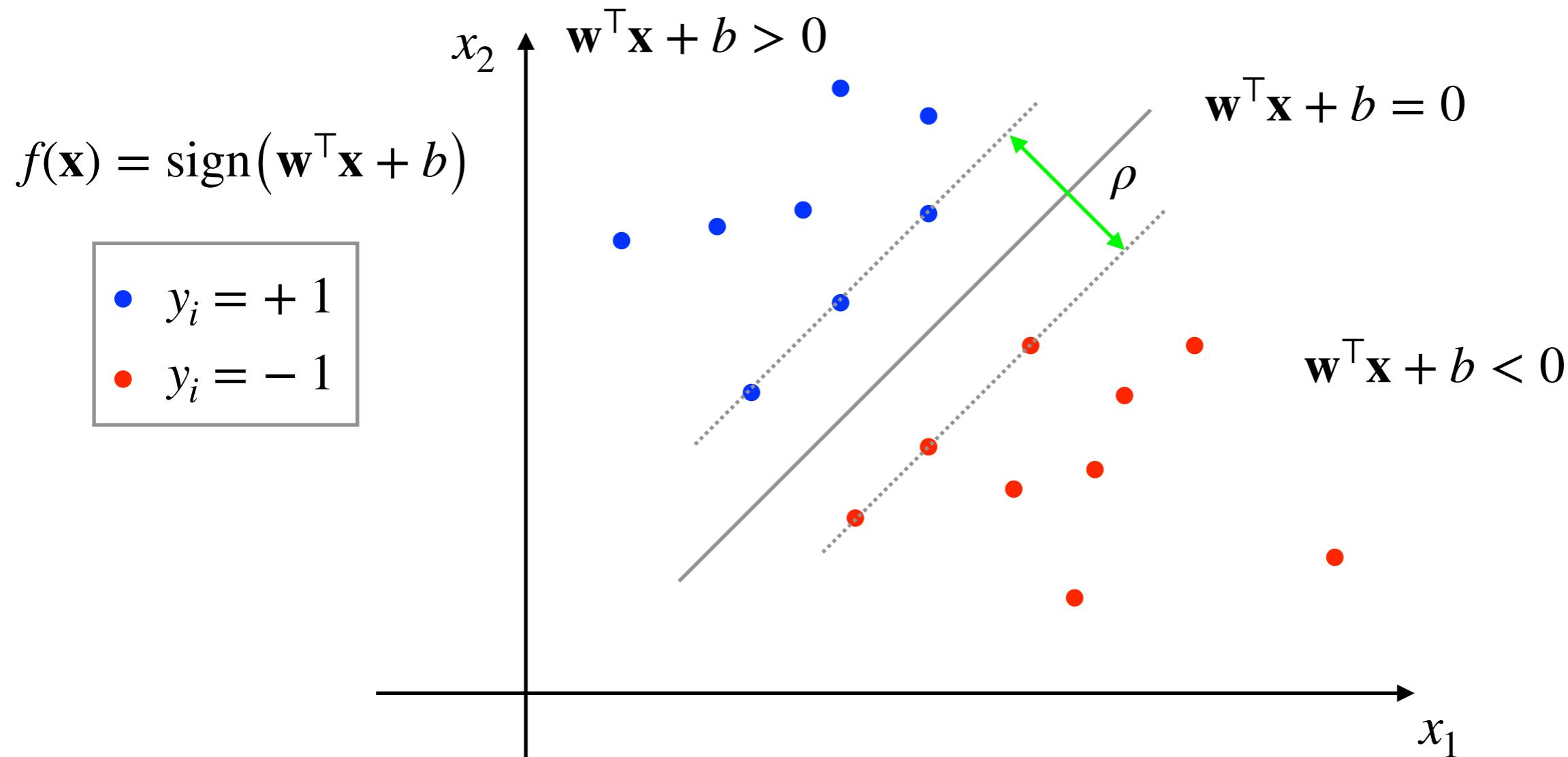


图7：分类间距。

支持向量机的形式化表示

令 $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$ 表示训练集，其中 $\mathbf{x}_i \in \mathbb{R}^d$ 是一个 d 维实数向量的输入数据点， $y_i \in \{-1, +1\}$ 是输出的类别。如果训练集可以被一个分类平面分开，根据分类间隔 ρ 的定义，对于任意样本 (\mathbf{x}_i, y_i) 都满足：

- ① 如果 $y_i = -1$ ，则 $\mathbf{w}^\top \mathbf{x}_i + b \leq -\rho/2$
- ② 如果 $y_i = +1$ ，则 $\mathbf{w}^\top \mathbf{x}_i + b \geq +\rho/2$

换句话说，函数距离满足以下不等式：

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq \rho/2$$

由于参数除以一个常数不影响结果（如 $\text{sign}(2x - 2) \equiv \text{sign}(x - 1)$ ），可以进一步简化为

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

支持向量机的形式化表示

根据支持向量的定义，作为支持向量的样本点满足分类距离最小，因此可以对上述不等式取等：

$$y_s(\mathbf{w}^\top \mathbf{x}_s + b) = 1$$

其中， \mathbf{x}_s 表示支持向量， y_s 表示对应的类别。

由于 $|y_s| = 1$ ，必然有 $|\mathbf{w}^\top \mathbf{x}_s + b| = 1$ ，因此支持向量到分类平面的几何距离为

$$\frac{\rho}{2} = \frac{|\mathbf{w}^\top \mathbf{x}_s + b|}{\|\mathbf{w}\|^2} = \frac{1}{\|\mathbf{w}\|^2}$$

因此，最大化分类间隔 ρ 的问题转化为最小化 $\|\mathbf{w}\|^2$ 的问题，消除了非参数变量 ρ 。

约束条件

支持向量机的目标是最大化分类间隔（即几何距离），但一个重要的前提条件是必须保证分类正确（即函数距离）：

$$y_i(\mathbf{w}^\top \mathbf{x}_i + b) \geq 1$$

可以做一个变换消除掉偏置项 b

$$\mathbf{x}_i = [\mathbf{x}_i^\top, 1]^\top$$

$$\mathbf{w} = [\mathbf{w}^\top, -b]^\top$$

因此，约束条件可以简化为对于所有的 $i \in [1, N]$ ，满足

$$y_i \mathbf{w}^\top \mathbf{x}_i \geq 1$$

支持向量机的训练是一个受限条件下的求函数极值问题。

优化与推断

综上所述，支持向量机的优化目标是在保持分类正确的条件下最大化分类间隔，形式化表述为：

$$\hat{\mathbf{w}} = \operatorname{argmin}_{\mathbf{w}} \left\{ \frac{1}{2} \|\mathbf{w}\|^2 \right\}$$

$$\text{s. t. } y_i \mathbf{w}^\top \mathbf{x}_i \geq 1, \forall i = 1, \dots, N$$

参数 \mathbf{w} 可以理解为数据点 \mathbf{x}_i 的权重，对于 d 维中的每一维分量都有权重。

在获得最优参数 $\hat{\mathbf{w}}$ 后，可以使用下面的推断公式对未知数据点 \mathbf{x} 的类别进行预测：

$$f(\mathbf{x}) = \operatorname{sign}(\hat{\mathbf{w}}^\top \mathbf{x})$$

关键问题在于，如何在受限条件下获得支持向量机的最优参数？

拉格朗日乘子法

受限条件下的优化通常使用拉格朗日乘子法来解决。需要注意的是，这里的约束条件是不等式，而不是等式。因此，优化目标可定义为：

$$L(\mathbf{w}, \boldsymbol{\alpha}) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1) \quad (\text{公式1})$$

$$\text{s.t. } \alpha_i \geq 0, \forall i \in [1, N]$$

其中， $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_N)^\top$ 是拉格朗日乘子向量。

由于 $y_i \mathbf{w}^\top \mathbf{x}_i \geq 1$ 并且 $\alpha_i \geq 0$ ，则有

$$-\sum_{i=1}^N \alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1) \leq 0 \quad (\text{公式2})$$

拉格朗日乘子法

根据公式1和公式2，可以得到：

$$\max_{\alpha} L(\mathbf{w}, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2$$

因此，优化问题等价于求解：

$$(\hat{\mathbf{w}}, \hat{\alpha}) = \arg \min_{\mathbf{w}} \max_{\alpha} L(\mathbf{w}, \alpha) \quad (\text{公式3})$$

公式3可理解如下：首先以 α 为变量，计算 $L(\mathbf{w}, \alpha)$ 的最大值，然后再以 \mathbf{w} 为变量，进一步计算最小值。

相对于传统的直接求最大值或者最小值，支持向量机的优化问题要更加困难。这是因为在第一步中无法得到 α 与 \mathbf{w} 之间的转换关系（本质原因是 α 只是附在等式上的乘子），从而在第二步中无法消去 α 变成只有单个变量 \mathbf{w} 的求极值问题。

原始问题与对偶问题

为此，支持向量机采用了一种巧妙的方法，将原始问题的min-max求解

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\alpha}}) = \arg \min_{\mathbf{w}} \max_{\boldsymbol{\alpha}} L(\mathbf{w}, \boldsymbol{\alpha}) \quad (\text{公式3})$$

转换为对偶问题的max-min求解

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\alpha}}) = \arg \max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}) \quad (\text{公式4})$$

这样做好处在于，可以在第一步（即令 $L(\mathbf{w}, \boldsymbol{\alpha})$ 对 \mathbf{w} 的偏导为0）就得到 $\boldsymbol{\alpha}$ 与 \mathbf{w} 之间的转换关系，从而在第二步（即令 $L(\mathbf{w}, \boldsymbol{\alpha})$ 对 $\boldsymbol{\alpha}$ 的偏导为0）中消去 \mathbf{w} 得到单变量 $\boldsymbol{\alpha}$ 的函数。在求出最优拉格朗日乘子 $\hat{\boldsymbol{\alpha}}$ 之后，可以通过 $\boldsymbol{\alpha}$ 与 \mathbf{w} 之间的转换关系得到最优参数 $\hat{\mathbf{w}}$ 。

因此，我们可以改为求解对偶问题，将对偶问题的解作为原始问题的解。

支持向量机的对偶问题求解

对偶问题求解目标为：

$$(\hat{\mathbf{w}}, \hat{\boldsymbol{\alpha}}) = \arg \max_{\boldsymbol{\alpha}} \min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha}) \quad (\text{公式4})$$

首先考虑内层的计算最小值问题： $\min_{\mathbf{w}} L(\mathbf{w}, \boldsymbol{\alpha})$ 。计算目标函数 $L(\mathbf{w}, \boldsymbol{\alpha})$ 关于 \mathbf{w} 的偏导，可以得到

$$\frac{\partial L(\mathbf{w}, \boldsymbol{\alpha})}{\partial \mathbf{w}} = \mathbf{w} - \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (\text{公式5})$$

令公式5中的偏导为0，可以得到 \mathbf{w} 的极值点：

$$\hat{\mathbf{w}} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i \quad (\text{公式6})$$

支持向量机的对偶问题求解

将公式4代入公式1，得到

$$\begin{aligned} L(\hat{\mathbf{w}}, \boldsymbol{\alpha}) &= \frac{1}{2} \|\hat{\mathbf{w}}\|^2 - \sum_{i=1}^N \alpha_i (y_i \hat{\mathbf{w}}^\top \mathbf{x}_i - 1) \\ &= \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{Y} \mathbf{G} \mathbf{Y} \boldsymbol{\alpha} \end{aligned} \quad (\text{公式7})$$

其中， $\mathbf{1} \in \mathbb{R}^{N \times 1}$ 是一个元素均为1的N维列向量， $\mathbf{Y} \in \mathbb{R}^{N \times N}$ 是一个对角矩阵 $\text{diag}(y_1, \dots, y_N)$ ， $\mathbf{G} \in \mathbb{R}^{N \times N}$ 是一个方阵，其元素 $G_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ 。

可以使用SMO算法来求解 $\hat{\boldsymbol{\alpha}}$ ：

$$\hat{\boldsymbol{\alpha}} = \operatorname{argmax}_{\boldsymbol{\alpha}} \left\{ \boldsymbol{\alpha}^\top \mathbf{1} - \frac{1}{2} \boldsymbol{\alpha}^\top \mathbf{Y} \mathbf{G} \mathbf{Y} \boldsymbol{\alpha} \right\} \quad (\text{公式8})$$

KKT条件

原始问题与对偶问题同时取得最优解需要满足一定条件，这个条件就是 Karush-Kuhn-Tucker 条件，简称 KKT 条件。

考虑一个通用的约束优化问题： $\min f(\mathbf{x}) \text{ s.t. } g(\mathbf{x}) \leq 0$ ，KKT 条件主要包括以下四个子条件：

- ① 定常方程式： $\frac{\partial f}{\partial \mathbf{x}} + \lambda \frac{\partial g}{\partial \mathbf{x}} = 0$
- ② 原始可行性： $g(\mathbf{x}) \leq 0$
- ③ 对偶可行性： $\lambda \geq 0$
- ④ 互补松弛性： $\lambda g(\mathbf{x}) = 0$

支持向量机中的KKT条件

$$L(\mathbf{w}, \alpha) = \frac{1}{2} \|\mathbf{w}\|^2 - \sum_{i=1}^N \alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1)$$

$$\text{s.t. } \alpha_i \geq 0, \forall i \in [1, N]$$

- ① 定常方程式: $\frac{\partial L}{\partial \mathbf{w}} = 0$
- ② 原始可行性: $y_i \mathbf{w}^\top \mathbf{x}_i \geq 1, \forall i = 1, \dots, N$
- ③ 对偶可行性: $\alpha_i \geq 0, \forall i = 1, \dots, N$
- ④ 互补松弛性: $\alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1) = 0, \forall i = 1, \dots, N$

第4点的含义是，支持向量的样本点满足 $y_i \mathbf{w}^\top \mathbf{x}_i = 1$ 而且其对应的 $\alpha_i > 0$ ，而非支持向量的样本点对应的 $\alpha_i = 0$ 。

预测只与支持向量相关

- 支持向量机的预测只与支持向量相关，非支持向量的样本点并不起作用

$$f(\mathbf{x}) = \text{sign}(\hat{\mathbf{w}}^\top \mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \hat{a}_i y_i \mathbf{x}_i^\top \mathbf{x} \right)$$

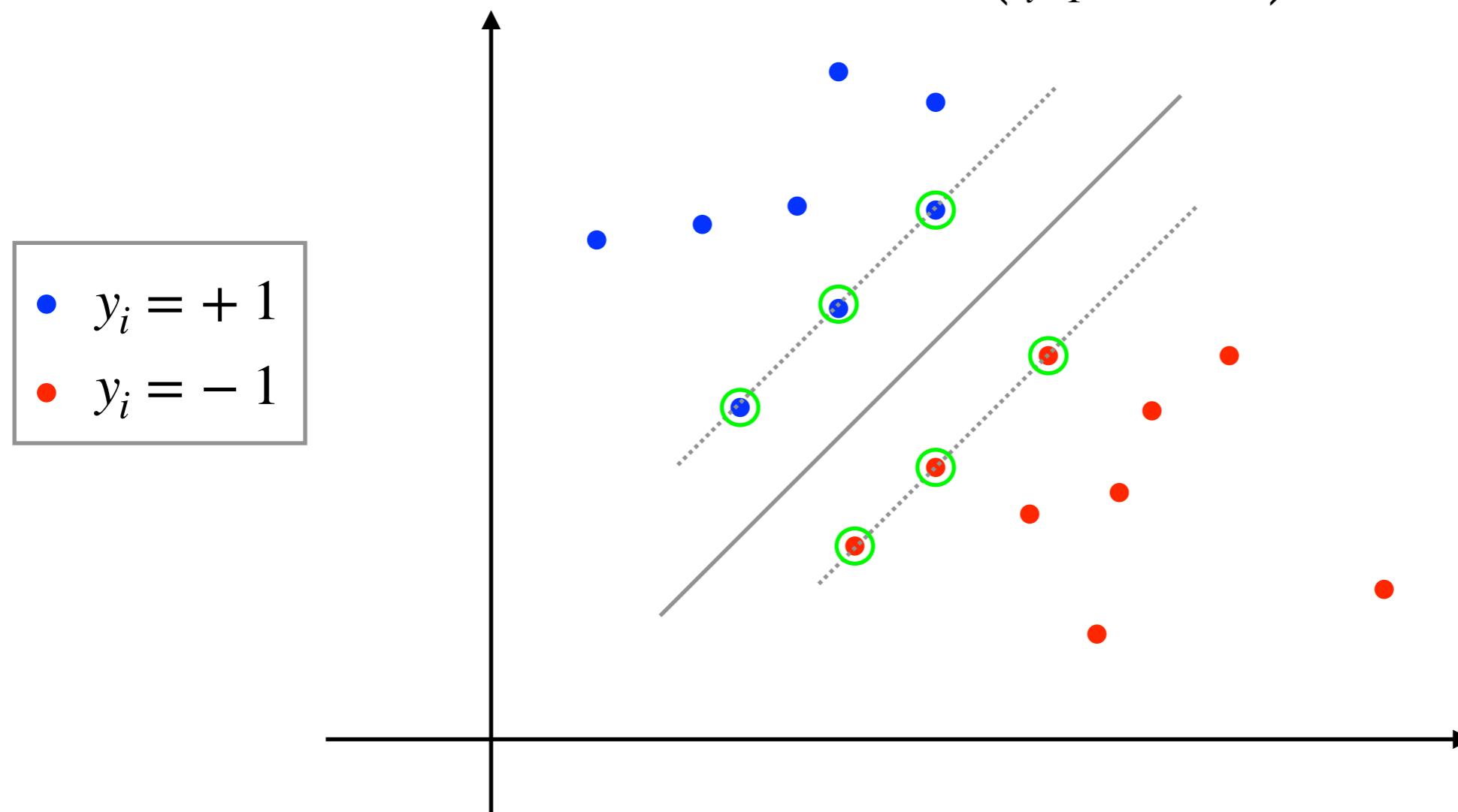


图8：支持向量。

内容提要

分类问题

支持向量机

松弛变量

核函数

线性不可分

- 上面讨论的都是线性可分数据。在实际应用中，存在着大量线性不可分的数据。该如何处理？



图9：一维线性不可分数据。

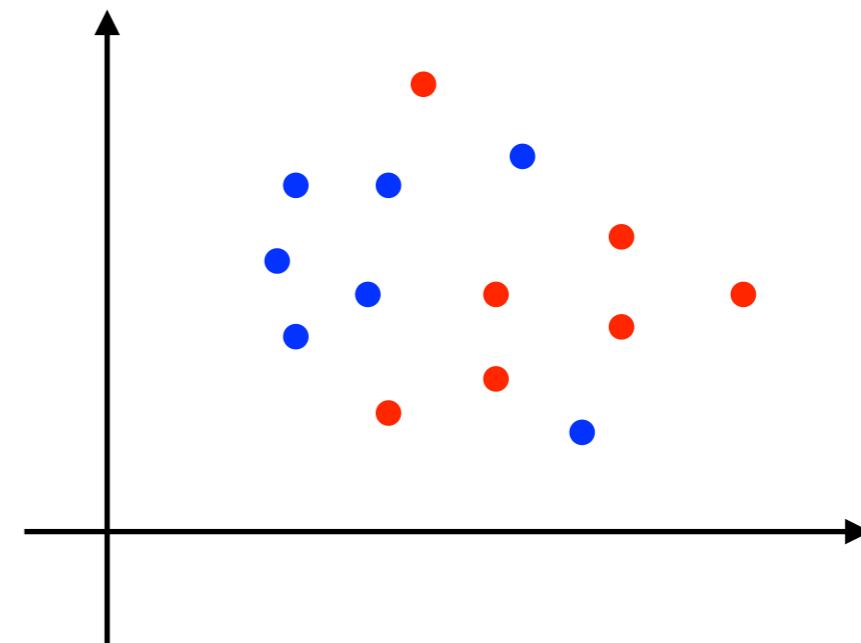


图10：二维线性不可分数据。

松弛变量

- 引入松弛变量，容忍部分不可分数据。

$$y_i \mathbf{w}^\top \mathbf{x}_i \geq 1 - \xi_i$$

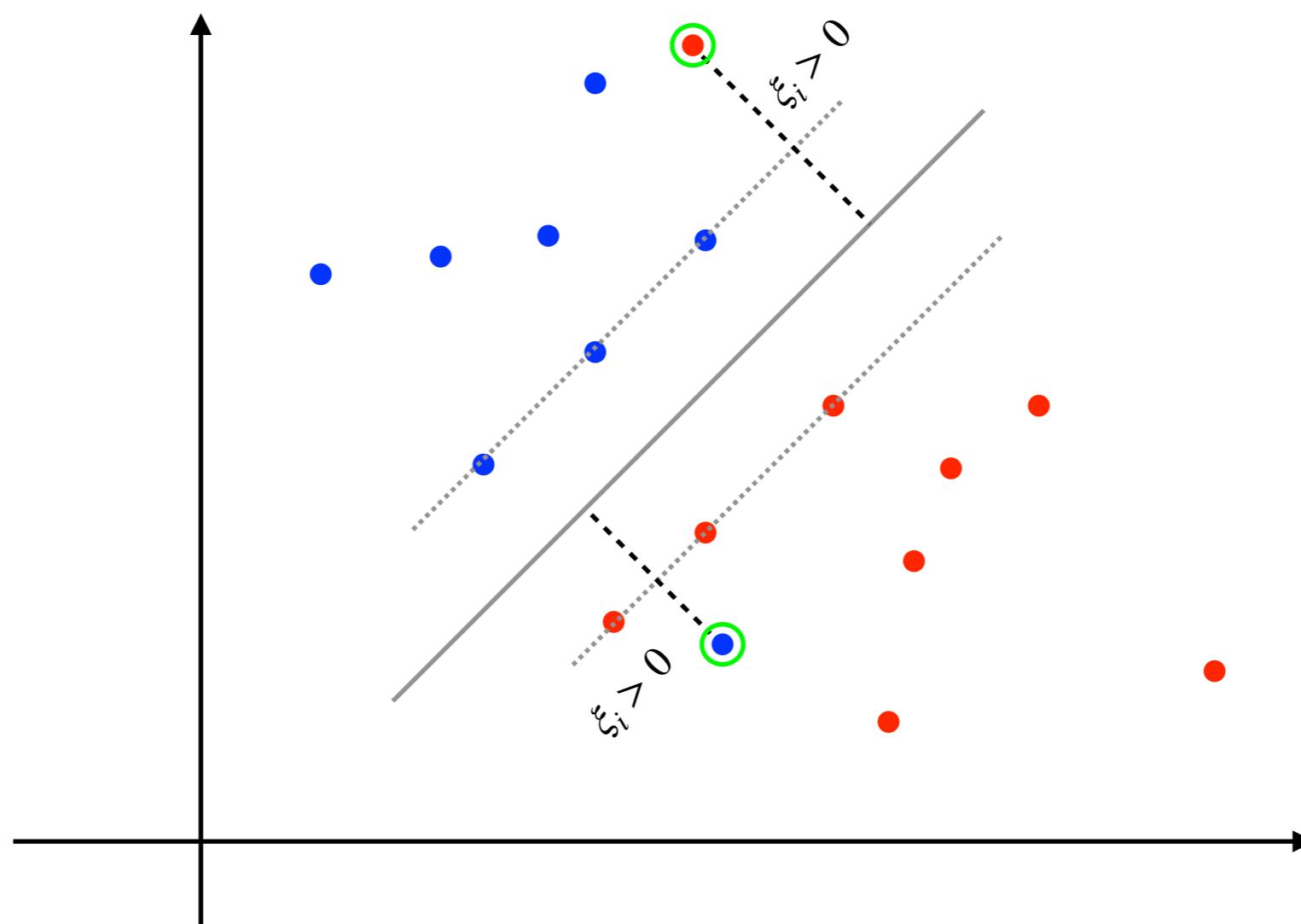


图8：支持向量。

引入松弛变量的支持向量机

引入松弛变量后，支持向量机的优化目标可以扩展为

$$\min_{\mathbf{w}, \xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i$$

$$\text{s.t. } y_i \mathbf{w}^\top \mathbf{x}_i \geq 1 - \xi_i, \forall i = 1, \dots, N$$

$$\xi_i \geq 0, \forall i = 1, \dots, N$$

我们将远离群体的样本点称为离群点。例如，图8中左上方蓝色样本点群中的红色样本点是一个离群点，因为它远离了右下方红色样本点群。

每个离群点 \mathbf{x}_i 都有一个 $\xi_i > 0$ ，表示其远离的程度，而非离群点对应的 $\xi_i = 0$ 。 C 是一个超参数，其取值越大，表示越重视离群点，不希望舍弃这些样本点。

因此，新的优化的目标是最大化分类间隔，同时最小化离群距离。

拉格朗日乘子法

使用拉格朗日乘子法，可以得到目标函数如下：

$$L(\mathbf{w}, \xi, \alpha, \beta) = \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^N \xi_i - \sum_{i=1}^N \alpha_i (y_i \mathbf{w}^\top \mathbf{x}_i - 1 + \xi_i) - \sum_{i=1}^N \beta_i \xi_i$$

原始问题可以表述为

$$\min_{\mathbf{w}, \xi} \max_{\alpha, \beta} L(\mathbf{w}, \xi, \alpha, \beta)$$

对偶问题可以表述为

$$\max_{\alpha, \beta} \min_{\mathbf{w}, \xi} L(\mathbf{w}, \xi, \alpha, \beta)$$

接下来，我们可以求解对偶问题获得原始问题的解。

对偶问题求解

首先计算目标函数关于 w 和 ξ 的极值点，可以得到

$$0 = \frac{\partial L}{\partial w} \Big|_{\hat{w}} \Rightarrow \hat{w} = \sum_{i=1}^N \alpha_i y_i \mathbf{x}_i$$

$$0 = \frac{\partial L}{\partial \xi} \Big|_{\hat{\xi}} \Rightarrow 0 \leq \alpha \leq C1$$

代入目标函数，可以得到以下优化公式：

$$\hat{\alpha} = \operatorname{argmax}_{0 \leq \alpha \leq C1} \left\{ \alpha^\top \mathbf{1} - \frac{1}{2} \alpha^\top \mathbf{G} \mathbf{Y} \mathbf{G} \alpha \right\}$$

仍然可以使用SMO算法进行求解。与之前相比，主要是加入了新的约束条件。

内容提要

分类问题

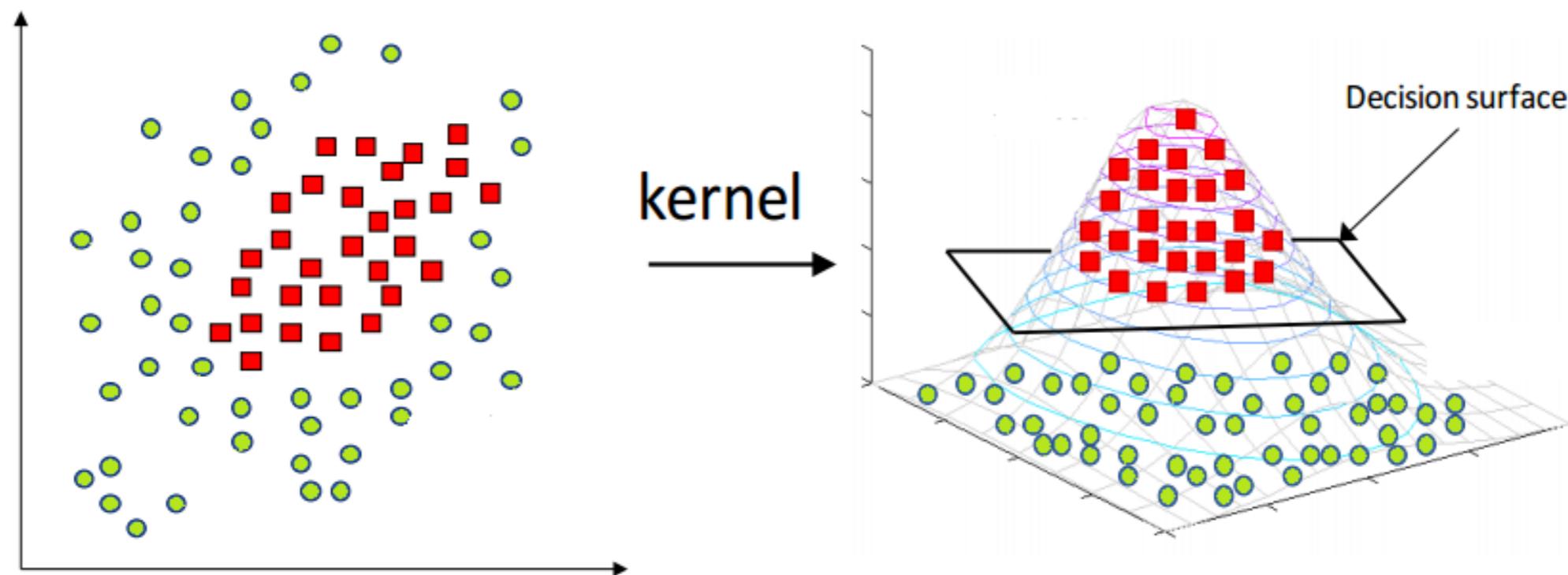
支持向量机

松弛变量

核函数

空间映射

虽然松弛变量能够允许一定的离群点，但是只适合线性不可分情况不严重的情况。如果线性不可分情况非常严重，需要进一步进行空间映射，将低维空间的线性不可分问题转为高维空间的线性可分问题。



图片来源：[https://medium.com/analytics-vidhya](https://medium.com.analytics-vidhya)

核函数

将所有样本点从原来的不可分空间装换到一个新的可分的特征空间，我们需要定义一个映射：

$$\Phi : \mathbf{x} \rightarrow \Phi(\mathbf{x})$$

核函数是一个函数，定义如下：

$$K(\mathbf{x}_1, \mathbf{x}_2) = \langle \Phi(\mathbf{x}_1), \Phi(\mathbf{x}_2) \rangle$$

其中， $\langle \cdot, \cdot \rangle$ 表示两个向量的内积。注意到支持向量机的预测其实是有支持向量的内积所决定的：

$$f(\mathbf{x}) = \text{sign}\left(\sum_{i=1}^N \hat{\alpha}_i y_i \mathbf{x}_i^\top \mathbf{x} \right) = \text{sign}\left(\sum_{i=1}^N \hat{\alpha}_i y_i \langle \mathbf{x}_i, \mathbf{x} \rangle \right)$$

核函数的优点

核函数与在映射后的特征空间计算内积是等价的。好处在于可以直接在低维空间计算内积，不需要显式地进行空间映射，计算效率更高。

$$\begin{aligned} f(\mathbf{x}) &= \text{sign}\left(\sum_{i=1}^N \hat{\alpha}_i y_i \langle \Phi(\mathbf{x}_i), \Phi(\mathbf{x}) \rangle \right) \\ &= \text{sign}\left(\sum_{i=1}^N \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) \right) \end{aligned}$$

常见的核函数：

- ① 线性核函数： $K(\mathbf{x}_1, \mathbf{x}_2) = \langle \mathbf{x}_1, \mathbf{x}_2 \rangle$
- ② 多项式核函数： $K(\mathbf{x}_1, \mathbf{x}_2) = (\langle \mathbf{x}_1, \mathbf{x}_2 \rangle + R)^d$
- ③ 高斯核函数： $K(\mathbf{x}_1, \mathbf{x}_2) = \exp\left(-\frac{||\mathbf{x}_1 - \mathbf{x}_2||^2}{2\sigma^2}\right)$

总结

- 分类是人工智能中的常见问题，是在特征空间中将属于不同类别的数据点进行分隔的任务。
- 支持向量机是一种最大化分类间隔的分类器。
- 由于满足KKT条件，支持向量机的优化问题通常可以转换为对偶问题求解。
- 通过引入松弛变量，我们可以训练得到软约束的支持向量机。
- 可以通过使用核函数让支持向量机处理线性不可分数据。

谢谢