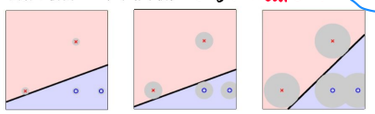
7.1 最大间隔分类面

对于感知器算法，无法区分分类优劣 $E_{out}(w) \le E_{in}(w) + \Omega(\mathcal{H})$



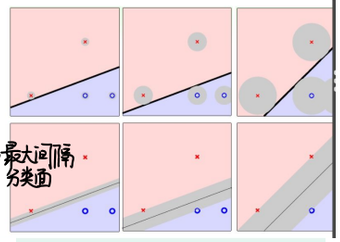
假设 $\mathbf{x} \approx \mathbf{x}_n + \Delta\mathbf{x}_n$，$\Delta\mathbf{x}_n \sim N(\mathbf{x}_n, \sigma_n)$

$\mathbf{x}_n$ 离分类面越远 ⇔ 容许σ_n越大，噪声的容忍度越大 ⇔ 不易出现过拟合

⇔ 更鲁棒分类面 ⇔ 对噪声的容忍度更大 ⇔ 离分类面最近的$\mathbf{x}_n$ 到分类面距离

最右边的最佳----因为离分类面最近的$\mathbf{x}_n$ 到分类面距离**最大**，对噪声最鲁棒

对噪声鲁棒的分类面 ⇕ 分类面到两边样本的距离要大 ⇕ “胖胖”的分类面 ⇔ 最大间隔分类面



算法的目的是如何找到“胖胖”的分类面

所有样本正确分类 ⇕ $y_n = \text{sign}(\mathbf{w}^T\mathbf{x}_n)$

$$\max_{\mathbf{w}} \ \text{margin}(\mathbf{w}) \quad \text{subject to} \ \ \mathbf{w} \text{ classifies every } (\mathbf{x}_n, y_n) \text{ correctly}$$

$$\text{margin}(\mathbf{w}) = \min_{n=1,\dots,N} \text{distance}(\mathbf{x}_n, \mathbf{w})$$

$y_n \mathbf{w}^T\mathbf{x}_n > 0$

算法目的：找到“最大间隔”的分类面

7.2 标准最大间隔问题

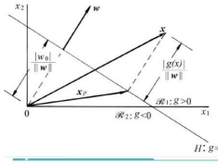
为便于计算分类器特性，支撑向量机的分析过程中样本向量不做增广 $x_0 = 1$

$h(x) = \text{sign}(\mathbf{w}^T\mathbf{x}) \Rightarrow h(x) = \text{sign}(\mathbf{w}^T\mathbf{x} + b)$

样本到分类面的距离

$$\max_{\mathbf{w}} \ \text{margin}(\mathbf{w}) \quad \text{Subject to} \ \ \text{every } y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0$$

$$\text{margin}(\mathbf{w}) = \min_{n=1,\dots,N} \text{distance}(\mathbf{x}_n, \mathbf{w})$$



$$g(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b = \mathbf{w}^T\left(\mathbf{x}_p + r\frac{\mathbf{w}}{\|\mathbf{w}\|}\right) + b = \mathbf{w}^T\mathbf{x}_p + b + r\frac{\mathbf{w}^T\mathbf{w}}{\|\mathbf{w}\|} = r\|\mathbf{w}\|$$

$$|r| = \frac{|g(\mathbf{x})|}{\|\mathbf{w}\|} \Rightarrow \text{distance}(\mathbf{x}_n, \mathbf{w}) = \frac{1}{\|\mathbf{w}\|}|\mathbf{w}^T\mathbf{x} + b| \Rightarrow \frac{1}{\|\mathbf{w}\|} y_n(\mathbf{w}^T\mathbf{x}_n + b)$$

同等

$$\max_{\mathbf{w}} \ \text{margin}(\mathbf{w}) \quad \text{Subject to} \ \ \text{every } y_n(\mathbf{w}^T\mathbf{x}_n + b) > 0 \quad \text{margin}(\mathbf{w}) = \min_{n=1,\dots,N} \frac{1}{\|\mathbf{w}\|} y_n(\mathbf{w}^T\mathbf{x}_n + b)$$

分类面的尺度缩放：

$\mathbf{w}^T\mathbf{x}_n + b = 0$

$3\mathbf{w}^T\mathbf{x}_n + 3b = 0$

$\therefore \ \min_{n=1,\dots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$

$$\text{margin}(\mathbf{w}) = \frac{1}{\|\mathbf{w}\|}$$

$$\max_{\mathbf{w}} \ \frac{1}{\|\mathbf{w}\|} \quad \text{Subject to} \ \ \min_{n=1,\dots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$$

标准的最大间隔问题：

条件松弛后的值域 $y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1, for\ all\ n$

原始约束条件下的值域 $\min_{n=1,\dots,N} y_n(\mathbf{w}^T\mathbf{x}_n + b) = 1$

条件松弛后的解仍然在紫色值域

如果$(\mathbf{w}^*, b^*)$ 位于蓝色值域，即对所有样本：

$y_n(\mathbf{w}^{*T}\mathbf{x}_n + b^*) \ge 1.6$

根据分类面$(\mathbf{w}^*, b^*)$ 取值的尺度不变：

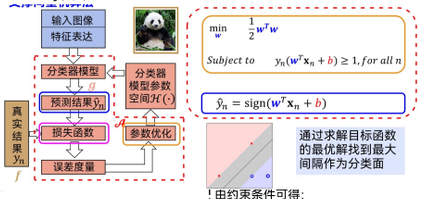
$y_n\left(\frac{\mathbf{w}^{*T}}{1.6}\mathbf{x}_n + \frac{b^*}{1.6}\right) \ge 1$

$\left(\frac{\mathbf{w}^{*T}}{1.6}, \frac{b^*}{1.6}\right)$为最佳解 ⇒ 矛盾！！！

$$\max_{\mathbf{w}} \ \frac{1}{\|\mathbf{w}\|} \Rightarrow \min_{\mathbf{w}} \|\mathbf{w}\| \Rightarrow \min_{\mathbf{w}} \sqrt{\mathbf{w}^T\mathbf{w}} \Rightarrow \min_{\mathbf{w}} \frac{1}{2}\mathbf{w}^T\mathbf{w}$$

$$\text{Subject to} \quad y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1, for\ all\ n$$

7.3 SVM

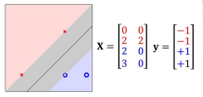


$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad \text{Subject to} \ \ y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1, for\ all\ n$$

$\hat{y}_n = \text{sign}(\mathbf{w}^T\mathbf{x}_n + b)$

通过求解目标函数的最优解找到最大间隔作为分类面

最大间隔面的求解示例：



$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

由约束条件可得：

$$\begin{cases} -b \ge 1 & (1) \\ -2w_1 - 2w_2 - b \ge 1 & (2) \\ 2w_1 + b \ge 1 & (3) \\ 3w_1 + b \ge 1 & (4) \end{cases}$$

对约束条件计算可得：

$(1) + (3) \Rightarrow w_1 \ge +1$

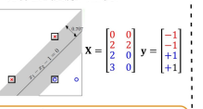
$(2) + (3) \Rightarrow w_2 \le -1$

由目标函数$\frac{1}{2}\mathbf{w}^T\mathbf{w}$取最小值可得：$w_1 = 1, \ w_2 = -1$

代入约束条件：$b = -1$

$\hat{y} = g(\mathbf{x}) = \text{sign}(x_1 - x_2 - 1)$

为什么叫支撑向量机？



$$\mathbf{X} = \begin{bmatrix} 0 & 0 \\ 2 & 2 \\ 2 & 0 \\ 3 & 0 \end{bmatrix} \quad \mathbf{y} = \begin{bmatrix} -1 \\ -1 \\ +1 \\ +1 \end{bmatrix}$$

$$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w} \quad \text{Subject to} \ \ y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1, for\ all\ n$$

优化得到的解为：

$w_1 = 1, \ w_2 = -1, \ b = -1$

$g(\mathbf{x}) = \text{sign}(x_1 - x_2 - 1)$

$\text{margin}(\mathbf{w}) = \frac{1}{\|\mathbf{w}\|} = \frac{1}{\sqrt{2}}$

➢分类面由边界上的样本确定，其他样本不起作用

➢边界上的样本被称为支撑向量（候选）

支撑向量机(SVM)—Support Vector Machine

----借助支撑向量学到间隔最大分类面

$y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1$

$y_n s_n \ge 1$

$1 - y_n s_n \le 0$

支撑向量机算法的损失函数

$L_{SVM} = \max(0, 1 - ys)$ Hinge Loss

$\frac{\partial L_{SVM}(\mathbf{w})}{\partial \mathbf{w}} = [1 - y_n(\mathbf{w}^T\mathbf{x}_n) \ge 0](-y_n\mathbf{x}_n)$ (增广后)

$\hat{y}_n = \text{sign}(\mathbf{w}^T\mathbf{x}_n + b)$

$s = \mathbf{w}^T\mathbf{x}_n (\mathbf{w}^T\mathbf{x}_n + b)$

$\begin{cases} 1 - y_n\mathbf{w}^T\mathbf{x}_n \ge 0 & \text{取 } 1 \quad (-y_n\mathbf{x}_n) \\ 1 - y_n\mathbf{w}^T\mathbf{x}_n < 0 & \text{取 } 0 \quad 0 \end{cases}$

- 初始化权向量$\mathbf{w}_0$ **Stochastic Gradient Descent(SGD)**
- ***for*** $t = 0, 1, 2, \dots$ (t 代表迭代次数)
  - ① 计算梯度：$\nabla L_{SVM}(\mathbf{w}_t) = \frac{1}{B}\sum_{n=1}^{B}[\![1 - y_n(\mathbf{w}^T\mathbf{x}_n) \ge 0]\!](-y_n\mathbf{x}_n)$
  - ② 对权向量$\mathbf{w}_t$进行更新：$\mathbf{w}_{t+1} \leftarrow \mathbf{w}_t - \eta\nabla L_{SVM}(\mathbf{w}_t)$

…直到对任意$\mathbf{x}_n$满足$1 - y_n(\mathbf{w}_{t+1}^T\mathbf{x}_n) < 0$，或者迭代足够多次数

返回最终的$\mathbf{w}_{t+1}$作为学到的g

$L_{0/1} = [\![\hat{y} \ne y]\!]$

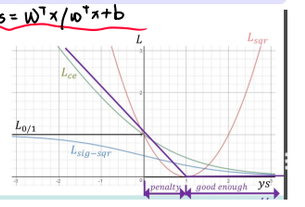
$L_{sqr} = (ys - 1)^2$

$L_{sig-sqr} = (\theta(ys) - 1)^2$

$L_{ce} = \ln(1 + \exp(-ys))$

$L_{SVM} = \max(0, 1 - ys)$

$s = \mathbf{w}^T\mathbf{x}/\mathbf{w}^T\mathbf{x} + b$



SVM一般求解

$\min_{\mathbf{w}} \ \frac{1}{2}\mathbf{w}^T\mathbf{w}$

Subject to $y_n(\mathbf{w}^T\mathbf{x}_n + b) \ge 1$ for all n （寻工求解不易 / 梯度下降无法处理约束条件）

SVM求解模型特点：

$\begin{cases} (\mathbf{w}, b)$目标函数为二次函数(凸函数) \\ $(\mathbf{w}, b)$约束条件线性 \end{cases}$

⇒二次规划问题，有成熟求解算法

二次规划(QP)的求解：

最佳的$\mathbf{u} \leftarrow \text{QP}(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$

$\min_{\mathbf{u}} \ \frac{1}{2}\mathbf{u}^T\mathbf{Q}\mathbf{u} + \mathbf{p}^T\mathbf{u}$

Subject to $\mathbf{a}_m^T\mathbf{u} \ge c_m$ for $m = 1, 2, \dots M$

$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}, \ \mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}, \ \mathbf{p} = \mathbf{0}_{d+1}, \ \mathbf{a}_n^T = y_n[1 \ \ \mathbf{x}_n^T], \ c_n = 1, \ M = N$

$\mathbf{u} = \begin{bmatrix} b \\ \mathbf{w} \end{bmatrix}$ $\mathbf{Q} = \begin{bmatrix} 0 & 0 \\ 0 & I_d \end{bmatrix}$ $\mathbf{p}^T = 0$ $c_m = 1$

利用二次规划(QP)实现支撑向量机

① $\mathbf{Q} = \begin{bmatrix} 0 & \mathbf{0}_d^T \\ \mathbf{0}_d & \mathbf{I}_d \end{bmatrix}$, $\mathbf{p} = \mathbf{0}_{d+1}$, $\mathbf{a}_n^T = y_n[1 \ \ \mathbf{x}_n^T]$, $c_n = 1$,

② $\begin{bmatrix} b \\ \mathbf{w} \end{bmatrix} \leftarrow \text{QP}(\mathbf{Q}, \mathbf{p}, \mathbf{A}, \mathbf{c})$

③ 返回最终的b和$\mathbf{w}$作为学到的g_{SVM}

线性硬间隔SVM算法(*Linear Hard-Margin SVM Algorithm*)

➢ *Hard-Margin*：没有任何样本会落入到“胖胖的”间隔区域里！

➢ *Linear*：样本$\mathbf{x}_n$是线性可分的！*如果*不是线性可分？ $\mathbf{z}_n = \Phi(\mathbf{x}_n)$