支持向量机

Support Vector Machine

① 硬间隔 SVM

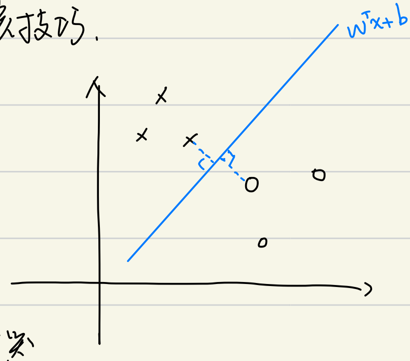
② 软间隔 SVM

③ 约束优化问题

硬间隔SVM (hard margin SVM)

SVM所涉及到的东西：间隔、对偶、核技巧.

$$\text{SVM}\begin{cases}\text{hard-margin SVM}\\ \text{soft-margin SVM}\\ \text{kernel SVM}\end{cases}$$



其中hard-margin SVM又叫作最大间隔分类器

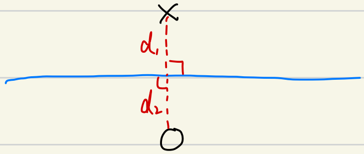
目的是寻找一个超平面，该超平面能够将不同类数据分开，且所有数据点到该超平面的最小距离最大

定义超平面为 $w^Tx+b=0$

则数据x_i到该超平面的距离为 $\dfrac{|w^Tx_i+b|}{||w||}$

则SVM目标：
$$\begin{cases}\max\limits_{w,b}\ \min\limits_{x_i,y_i}\ \dfrac{1}{||w||}|w^Tx_i+b|\\ s.t.\ \ y_i(w^Tx_i+b)>0\end{cases}$$
条件是超平面能够将数据分开.

稍微想一下就能知道，两类数据中，到满足上述目标的超平面的距离最小值是相等的. 见下图.



$d_1=d_2$.

若$d_1<d_2$，那么最小距离为d_1，但此时的d_1明显不是最大值.

化简 $\begin{cases} \max\limits_{w,b} \min\limits_{x_i,y_i} \frac{1}{\|w\|} |w^T x_i + b| & ① \\ s.t.\ \ y_i(w^T x_i + b) > 0. & ② \end{cases}$

对于 ②，可以写成 $\exists r > 0.\ \min\limits_{x_i,y_i} y_i(w^T x_i + b) = r$.

则 ① 也可以写成 $\max\limits_{w,b} \frac{1}{\|w\|} \min\limits_{x_i,y_i} y_i(w^T x_i + b) = \max\limits_{w,b} \frac{1}{\|w\|} \min\limits_{x_i,y_i} r$.

$\boxed{令\ r=1}$，则 ② = $s.t.\ y_i(w^T x_i + b) \geq 1$

① = $\max\limits_{w,b} \frac{1}{\|w\|} = \min\limits_{w,b} \boxed{\frac{1}{2}} w^T w$.　为了后面求导时约去系数

$w^T x_i + b = 0$ 与 $2w^T x_i + 2b = 0$ 代表的是同一个超平面，

所以不论 r 为多少，我们都可以通过放缩来使得 $r=1$.

则目标为 $\boxed{\begin{cases} \min\limits_{w,b} \underbrace{\frac{1}{2} w^T w}_{二次} & (1) \\ s.t.\ \underbrace{y_i(w^T x_i + b) \geq 1.}_{N个约束，线性.} & \forall i = 1,2,\cdots,N. \end{cases}}$　凸优化问题.

关于凸优化问题，见如下链接

https://blog.csdn.net/xu_fengyu/article/details/84727096

针对于 w, b 而言

很自然地，想到用拉格朗日乘子法，$\boxed{将有约束问题转化为无约束问题}$

$\mathcal{L}(w,b,\lambda) = \frac{1}{2} w^T w + \boxed{\sum\limits_{i=1}^{N}} \lambda_i [1 - y_i(w^T x_i + b)]$，$\lambda_i \geq 0$.

$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \end{pmatrix}$　因为上面有 N 个约束.

则目标转化为 $\begin{cases} \min\limits_{w,b} \max\limits_{\lambda} \mathcal{L}(w,b,\lambda) \\ s.t.\ \ \lambda_i \geq 0 \end{cases}$ (2)

新目标与原目标等价的探讨

由于 $\lambda_i \geq 0$. 则 当 $1-y_i(w^Tx_i+b) > 0$ 时. $\max\limits_{\lambda} \mathcal{L}(w,b,\lambda) = \infty$

当 $1-y_i(w^Tx_i+b) \leq 0$ 时, $\max\limits_{\lambda} \mathcal{L}(w,b,\lambda) = 0$.

$\therefore \min\limits_{w,b} \max\limits_{\lambda} \mathcal{L}(w,b,\lambda) = \min\limits_{w,b} (\infty, \frac{1}{2}w^Tw) = \min\limits_{w,b} \frac{1}{2}w^Tw$

隐式地, 此时 $1-y_i(w^Tx_i+b)$ 必定小于等于0

也就是说, 加入 $\max\limits_{\lambda}$ 和 $\lambda_i \geq 0$ 的条件是为了在求解过程中.
将 $1-y_i(w^Tx_i+b) > 0$ 这种情况给屏蔽掉

$\therefore$ 两者是等价的.

利用对偶性质, 得到对偶问题为

$\begin{cases} \max\limits_{\lambda} \min\limits_{w,b} \mathcal{L}(w,b,\lambda) \\ s.t.\ \ \lambda_i \geq 0. \end{cases}$ (3)

一般地, $\min \max \mathcal{L} \geq \max \min \mathcal{L}$ (弱对偶)

但是, 原问题目标函数是二次的, 约束是线性地 (凸优化二次问题)

则此时, $\min \max \mathcal{L} = \max \min \mathcal{L}$ (强对偶).

则对其进行求解：$\begin{cases} \max\limits_{\lambda} \min\limits_{w,b} \mathcal{L}(w,b,\lambda) \\ s.t. \quad \lambda_i \geq 0 \end{cases}$

先解 $\min\limits_{w,b} \mathcal{L}(w,b,\lambda)$，

$$\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial b} = \frac{\partial}{\partial b}\left[\sum_{i=1}^{N} \lambda_i (1 - y_i (w^T x_i + b))\right]$$

$$= \frac{\partial}{\partial b}\left[\sum_{i=1}^{N} \lambda_i y_i b\right]$$

$$= \sum_{i=1}^{N} \lambda_i y_i$$

令 $\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial b} = 0$，得：$\sum\limits_{i=1}^{N} \lambda_i y_i = 0$ 。(a)

将(a)代入 $\mathcal{L}(w,b,\lambda)$，得：

$$\mathcal{L}(w,b,\lambda) = \frac{1}{2} w^T w + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i w^T x_i$$

令 $\frac{\partial \mathcal{L}(w,b,\lambda)}{\partial w} = w - \sum\limits_{i=1}^{N} \lambda_i y_i x_i = 0$.

得 $w = \sum\limits_{i=1}^{N} \lambda_i y_i x_i$ (b)

将(b)代入 $\mathcal{L}(w,b,\lambda)$，得：

$$\min_{w,b} \mathcal{L}(w,b,\lambda) = \frac{1}{2}\left(\sum_{i=1}^{N} \lambda_i y_i x_i\right)^T \sum_{j=1}^{N} \lambda_j y_j x_j + \sum_{i=1}^{N} \lambda_i - \sum_{i=1}^{N} \lambda_i y_i \left(\sum_{j=1}^{N} \lambda_j y_j x_j\right)^T x_i$$

$$= \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j - \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_j^T x_i + \sum_{i=1}^{N} \lambda_i$$

$$= -\frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N} \lambda_i$$

则问题变为：
$$\begin{cases}\max\limits_{\lambda}\ -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j y_i y_j x_i^T x_j + \sum_{i=1}^{N}\lambda_i \\ s.t.\ \ \lambda_i \geq 0,\ \ \sum_{i=1}^{N}\lambda_i y_i = 0\end{cases} \qquad (4)$$

带约束

$$(1) \overset{\text{无约束}}{\Longleftrightarrow} (2) \overset{\text{强对偶}}{\Longleftrightarrow} (3) \Longleftrightarrow (4)$$

原问题与对偶问题是强对偶关系 $\Longleftrightarrow$ 满足KKT条件.

则该问题的KKT条件为.

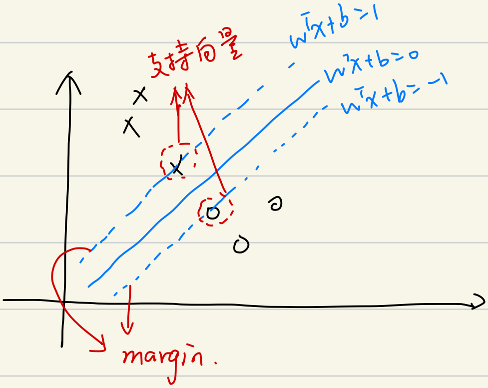
$$\begin{cases}\frac{\partial L}{\partial w}=0 \\ \frac{\partial L}{\partial b}=0 \\ \lambda_i\left[1-y_i(w^Tx_i+b)\right]=0 \\ \lambda_i \geq 0 \\ 1-y_i(w^Tx_i+b)\leq 0\end{cases}$$

松弛互补条件

当 $1-y_i(w^Tx_i+b)<0$ 时、 $\lambda_i=0$

当 $1-y_i(w^Tx_i+b)=0$ 时， $\lambda_i<0$.

由上述计算，已经得出 $\boxed{W^*=\sum_{i=1}^{N}\lambda_i y_i x_i}$，下面需要求解 b^*.



由图可知，将支持向量代入 w^Tx+b，结果为1

即当 x_k 为支持向量时.

$$w^Tx_k+b=1 \Rightarrow \boxed{\begin{aligned} b^* &= 1-w^Tx_k \\ &= 1-\sum_{i=1}^{N}\lambda_i y_i x_i^T x_k\end{aligned}}$$

从KKT条件中还可以看出，对于不在margin上的数据，其对应的 $\lambda_i=0$.

对(4)进行求解可以得到 λ，这里不进行求解，可以用SMO算法.

或者用现成的求解QP问题的工具就可以得到

支持向量机就是一个QP问题（Quadratic Programming，二次规划）

软间隔SVM (Soft margin SVM)

软间隔SVM允许SVM出现错误，通过在优化条件中加入损失项来达到该目的。即 $\min \frac{1}{2}w^Tw + loss$

错误 → $y_i(w^Tx_i+b)<1$

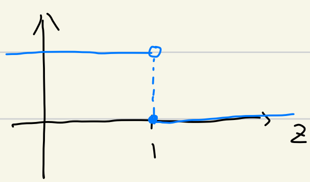
① 一般地，会用0/1损失来统计出错的个数.

$$loss = \sum_{i=1}^{N} I\{y_i(w^Tx_i+b)<1\}.$$

$I\{z\}$为指示函数，z为真时为1，z为假时为0.

令 $z = y_i(w^Tx_i+b)$

图像如下所示.



可见，其不连续，不利于求解.

② 当统计个数行不通时，使用距离来代替.

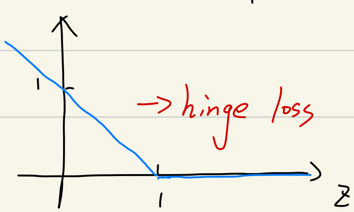
当 $y_i(w^Tx_i+b)\geq 1$ 时，$loss = 0$

当 $y_i(w^Tx_i+b)<1$ 时，$loss = 1-y_i(w^Tx_i+b)$.

$\frac{1}{\|w\|}y_i(w^Tx_i+b)$ 是距离，但 $\frac{1}{\|w\|}$ 对所有x_i都一样，所以我们只考虑 $y_i(w^Tx_i+b)$，$1-y_i(w^Tx_i+b)$ 就等价于当前点与margin之间的距离.

则 $loss = \sum_{i=1}^{N} \max(0,\ 1-y_i(w^Tx_i+b))$

图像如下图所示

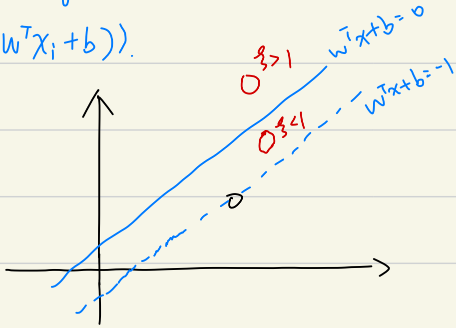


引入松弛变量 $\xi_i \geq 0$

$\xi_i = 1 - y_i(w^T x_i + b)$，由于 $\xi_i \geq 0$，故当 $y_i(w^T x_i + b) \geq 1$ 时，$\xi_i = 0$。
或者可以直接看作 $\xi_i = \max(0, 1 - y_i(w^T x_i + b))$。

此时软间隔的目标问题为

$$\begin{cases} \min\limits_{w,b} \frac{1}{2} w^T w + C \sum\limits_{i=1}^{N} \xi_i \\ s.t. \quad y_i(w^T x_i + b) \geq 1 - \xi_i \end{cases}$$



显然，对于每一个 x_i 都有一个对应的 ξ_i，来反映其不满足约束的程度。
C 是常数，当 $C \to \infty$ 时，会迫使所有 $\xi_i = 0$，则约束条件又会变为硬间隔的约束条件；当 C 取有限常数时，允许有一定的误差。

约束优化问题

(Constraint Optimization)

弱对偶性证明

约束优化问题(原问题, primal problem)

一般表达 $\begin{cases} \min\limits_x f(x) \\ s.t. \quad m_i(x) \le 0 \quad, i=1,2,\cdots,M \\ \qquad n_j(x) = 0 \quad, j=1,2,\cdots,N \end{cases}$

原问题是关于x的函数.

拉格朗日函数 (Lagrange)

$$\mathcal{L}(x,\lambda,\eta) = \underbrace{f(x)}_{\text{target function}} + \sum_{i=1}^{M} \underbrace{\lambda_i}_{\lambda_i \ge 0} m_i(x) + \sum_{j=1}^{N} \underbrace{\eta_j}_{\text{无要求}} n_j(x)$$

则使用拉格朗日乘子法将约束优化问题转化为(依旧是原问题)

$$\begin{cases} \min\limits_x \max\limits_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta) \\ s.t. \ \lambda_i \ge 0 \end{cases}$$

原问题对x的无约束的形式
仍然是关于x的函数.

在逻辑上考察两者的等价性:

$\begin{cases} \text{对于 } m_i(x)\text{, 若 } m_i(x) > 0\text{, 则 } \max\limits_{\lambda,\eta} \mathcal{L} = \infty\text{. 若 } m_i(x) \le 0\text{, 则 } \max\limits_{\lambda,\eta} \mathcal{L} \text{ 存在} \\ \text{对于 } n_j(x)\text{, 若 } n_j(x) \ne 0\text{, 则 } \max\limits_{\lambda,\eta} \mathcal{L} = \infty\text{, 若 } n_j(x) = 0\text{, 则 } \max\limits_{\lambda,\eta} \mathcal{L} \text{ 存在.} \end{cases}$

即 $\min\limits_x \max\limits_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta) = \min\limits_x (\infty, \infty, \max\limits_{\lambda,\eta} \mathcal{L}) = \min\limits_x \max\limits_{\lambda,\eta} \mathcal{L}.$

则只在约束优化的条件下, 才成立, 故两者等价

则对偶问题(dual problem)为

$$\begin{cases} \max\limits_{\lambda,\eta} \min\limits_{x} \mathcal{L}(x,\lambda,\eta) \\ s.t.\ \lambda_i \geq 0. \end{cases}$$

对偶问题为关于 λ, η 的函数

弱对偶性:对偶问题 ≤ 原问题.

$$\max_{\lambda,\eta} \min_{x} \mathcal{L}(x,\lambda,\eta) \leq \min_{x} \max_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta).$$

证明: $\underbrace{\min\limits_{x} \mathcal{L}(x,\lambda,\eta)}_{A(\lambda,\eta)} \leq \mathcal{L}(x,\lambda,\eta) \leq \underbrace{\max\limits_{\lambda,\eta} \mathcal{L}(x,\lambda,\eta)}_{B(x).}$

由于 $A(\lambda,\eta) \leq B(x)$

$\therefore \max A(\lambda,\eta) \leq \min B(x)$,即对偶问题 ≤ 原问题

弱对偶性的几何解释

原问题 $\begin{cases} \min\limits_{x\in\mathbb{R}^p} f(x) \\ s.t.\ m_1(x)\le 0 \end{cases}$ 忽略等式约束.

定义域：$D = \mathrm{dom} f \cap \mathrm{dom}\, m_1$

拉格朗日函数：$\mathcal{L}(x,\lambda) = f(x) + \lambda m_1(x)$，$\lambda \ge 0$.

对偶问题 $\begin{cases} \max\limits_{\lambda} \min\limits_{x} \mathcal{L}(x,\lambda) \\ s.t.\ \lambda \ge 0 \end{cases}$.

则 $p^* = \min\limits_{x} f(x)$ 原问题的最优解

$d^* = \max\limits_{\lambda} \min\limits_{x} \mathcal{L}(x,\lambda)$ 对偶问题的最优解

假设一个集合 $G = \{(m_1(x), f(x)) \mid x \in D\}$.

令 $u = m_1(x)$，$t = f(x) \Rightarrow G = \{(u,t) \mid x \in D\}$.

则 $p^* = \underline{\inf} \{t \mid (u,t)\in G,\ u\le 0\}$

集合的下确界

$d^* = \max\limits_{\lambda} \min\limits_{x} \mathcal{L}(x,\lambda)$

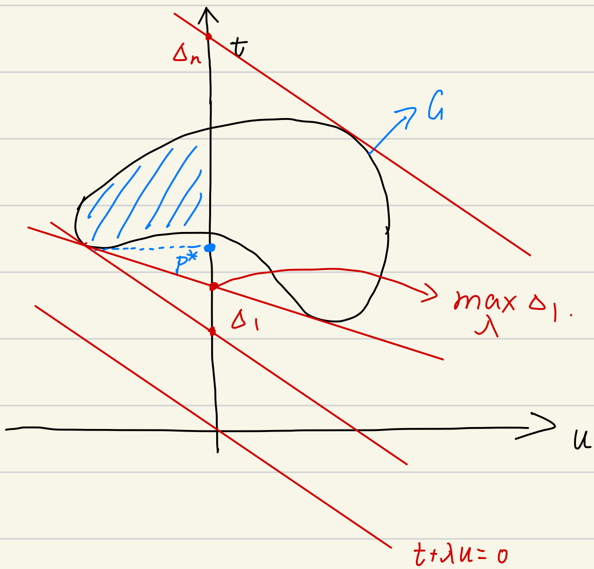
$= \max\limits_{\lambda} \min\limits_{x} (t + \lambda u)$

令 $g(\lambda) = \min\limits_{x} (t+\lambda u)$

$d^* = \max\limits_{\lambda} (t+\lambda u)$

$g(\lambda) = \inf \{t + \lambda u \mid (u,t) \in G\}$.

对上一页的定义进行可视化.（G定义为非凸集，更具一般性）.



由左图可知：$\{t+\lambda u \mid (u,t)\in G\} = \{\Delta_1, \cdots, \Delta_n\}$.

$g(\lambda) = \inf \{t+\lambda u \mid (u,t)\in G\} = \Delta_1$

则 $d^* = \max\limits_{\lambda} \Delta_1$

如何取λ，使得Δ_1最大？就等同于，通过改变λ，使得直线进行旋转，直到与G同时出现两个切点，（这是针对于所画的图像而言，不同集合使得Δ_1最大的条件会有些许不同）.

但此时，$d^* \le p^*$.

何时 $d^* = p^*$，通过图解释的话，就是当集合下界为凸时（右图）

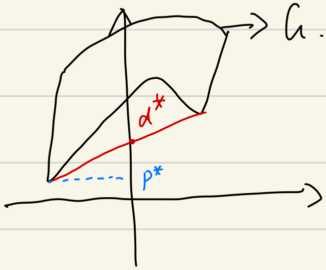
也就是说，在坐标轴的交点处存在集合的切线.

实际上，凸优化 + slater条件 $\Longrightarrow d^* = p^*$

但 slater 条件只是充分不必要条件，有些强对偶性不满足slater条件

svm 是一个二次规划问题，满足上述强对偶性成立的条件

但是我有一个问题，如果集合如下所示，是不是就错了？



Slater 条件

relative interior 意思是不考虑集合的边界

定义：$\exists\, \hat{x} \in \text{relint}\, D$，

s.t. $\forall\, i=1,\dots,M$，$m_i(\hat{x})<0$.

⇒ 几何意义：对于上一页的图解，slater就定了必定有一个点 (u,t) 在 t 轴的左边.

① 对于大多数凸优化问题，slater条件成立

② 放松的 slater：若 m 中有 k 个仿射函数，则只需校验另外 M-k 个函数是否满足条件。

最高次数为1的多项式函数.

则 SVM 的目标函数 $f(x)$ 是凸的，不等式约束是仿射函数，则 SVM 与其对偶函数是强对偶关系.

满足放松的 slater

$\boxed{\text{KKT条件}}$

原问题 $\begin{cases} \min\limits_{x} f(x) \\ s.t.\ \ m_i(x) \le 0,\ i=1,\cdots,M \\ \qquad n_j(x)=0,\ j=1,\cdots,N \end{cases}$

拉格朗日：$\mathcal{L}(x,\lambda,\eta) = f(x) + \sum_{i=1}^{M}\lambda_i m_i(x) + \sum_{j=1}^{N}\eta_j n_j(x)$，$\lambda_i \ge 0$.

最优解：$p^* \to x^*$ $\quad d^* \to \lambda^*, \eta^*$

Convex + slater $\Longrightarrow$ $\underbrace{\text{strong duality}}_{(d^*=p^*)}$ $\Longleftrightarrow$ KKT条件.

KKT：$\begin{cases} \text{① 满足可行条件} \\ \text{② 互补松弛} \\ \text{③ 梯度为0.} \end{cases}$

① 可行条件：$m_i(x^*) \le 0$. $n_j(x^*)=0$, $\lambda^* \ge 0$

② 互补松弛：
$$\begin{aligned} d^* = \max_{\lambda,\eta} g(\lambda,\eta) = g(\lambda^*,\eta^*) &= \min_{x} \mathcal{L}(x,\lambda^*,\eta^*) \\ &\le \mathcal{L}(x^*,\lambda^*,\eta^*) \\ &= f(x^*) + \sum_{i=1}^{M}\underbrace{\lambda_i^* m_i(x^*)}_{\le 0} + \sum_{j=1}^{N}\underbrace{\eta_j n_j(x^*)}_{=0} \\ &\le \underbrace{f(x^*)}_{p^*} \end{aligned}$$

又∵ $d^*=p^*$.

∴上述推导中的两个"≤"号都需要取"="

对于第二个"≤"号，如果其为"="，那么 $\boxed{\lambda_i^* m_i(x^*)=0}$

即为互补松弛的条件

③ 梯度为0. 还是看上述②中的推导.

当第一个"≤"号取"="时.

$$\min_x \mathcal{L}(x,\lambda^*,\eta^*) = \mathcal{L}(x^*,\lambda^*,\eta^*)$$

且 $m_i(x)$. $n_j(x)$ 都可微 (声明原问题时忘说了)

则 $$\boxed{\left.\frac{\partial \mathcal{L}(x,\lambda^*,\eta^*)}{\partial x}\right|_{x=x^*} = 0}$$

即为梯度为0的条件.

综上: KKT:

$$\begin{cases} m_i(x^*) \le 0 \\ n_j(x^*) = 0 \\ \lambda \ge 0 \\ \lambda_i^* m_i(x^*) = 0 \\ \left.\dfrac{\partial \mathcal{L}(x,\lambda^*,\eta^*)}{\partial x}\right|_{x=x^*} = 0 \end{cases}$$