

支持向量机

分布 p(θ)

支持向量机 (SVM) Support Vector Machines

① 硬边界 SVM: 即假设线性可分

如何寻找最优的超平面?

定义最小的 margin.

已经证明, margin 最大化能力最强.

因此要寻找 margin 最大的超平面.

定义 hyper plane: $\{x | w^T x + w_0 = 0\}$, w 代表法向量

再定义两条线, 使得 $|w^T x + w_0| = 1$. 且它们通过离超平面最近的点.

设这两个点分别为 x_1, x_2

则: $w^T x^{(1)} + w_0 = 1, w^T x^{(2)} + w_0 = -1$

$\therefore w^T (x^{(1)} - x^{(2)}) = 2 \therefore \gamma = \frac{1}{2} \cdot \frac{w^T (x^{(1)} - x^{(2)})}{\|w\|}$ 便是 margin

要最大化 margin, 就是 minimize $\|w\|$. 对于样本 $\{x^{(i)}, y^{(i)}\}$, 有

约束条件: $w^T x^{(i)} + w_0 \begin{cases} \geq 1, & \text{if } y^{(i)} = 1 \\ \leq -1, & \text{if } y^{(i)} = -1 \end{cases} \Rightarrow y^{(i)} (w^T x^{(i)} + w_0) \geq 1$

转为优化: minimize $\frac{1}{2} \|w\|^2$

s.t. $y^{(i)} (w^T x^{(i)} + w_0) \geq 1, \forall i$

这是一个二次优化问题 (QP problem), 通常转换为对偶问题

构造 Lagrange 函数:

$$L_p(w, w_0, \{\alpha_i\}) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^N \alpha_i [y^{(i)} (w^T x^{(i)} + w_0) - 1]$$

$$= \frac{1}{2} w^T w - w^T \sum_{i=1}^N \alpha_i y^{(i)} x^{(i)} - w_0 \sum_{i=1}^N \alpha_i y^{(i)} + \sum_{i=1}^N \alpha_i$$

$\frac{\partial L_p}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^N \alpha_i y^{(i)} x^{(i)}$

$\frac{\partial L_p}{\partial w_0} = 0 \Rightarrow \sum_{i=1}^N \alpha_i y^{(i)} = 0$

代入原式: 转为对偶问题

$$L_d(\{\alpha_i\}) = -\frac{1}{2} w^T w + \sum_{i=1}^N \alpha_i$$

$$= -\frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)} + \sum_{i=1}^N \alpha_i$$

全约束最大

μ₁ 2)

μ₂ 2)

μ₃ 2)

μ₄ 2)

μ₅ 2)

μ₆ 2)

μ₇ 2)

μ₈ 2)

μ₉ 2)

μ₁₀ 2)

μ₁₁ 2)

μ₁₂ 2)

μ₁₃ 2)

μ₁₄ 2)

μ₁₅ 2)

μ₁₆ 2)

μ₁₇ 2)

μ₁₈ 2)

μ₁₉ 2)

μ₂₀ 2)

μ₂₁ 2)

μ₂₂ 2)

μ₂₃ 2)

μ₂₄ 2)

μ₂₅ 2)

μ₂₆ 2)

μ₂₇ 2)

μ₂₈ 2)

μ₂₉ 2)

μ₃₀ 2)

μ₃₁ 2)

μ₃₂ 2)

μ₃₃ 2)

μ₃₄ 2)

μ₃₅ 2)

μ₃₆ 2)

μ₃₇ 2)

μ₃₈ 2)

μ₃₉ 2)

μ₄₀ 2)

μ₄₁ 2)

μ₄₂ 2)

μ₄₃ 2)

μ₄₄ 2)

μ₄₅ 2)

μ₄₆ 2)

μ₄₇ 2)

μ₄₈ 2)

μ₄₉ 2)

μ₅₀ 2)

μ₅₁ 2)

μ₅₂ 2)

μ₅₃ 2)

μ₅₄ 2)

μ₅₅ 2)

μ₅₆ 2)

μ₅₇ 2)

μ₅₈ 2)

μ₅₉ 2)

μ₆₀ 2)

μ₆₁ 2)

μ₆₂ 2)

μ₆₃ 2)

μ₆₄ 2)

μ₆₅ 2)

μ₆₆ 2)

μ₆₇ 2)

μ₆₈ 2)

μ₆₉ 2)

μ₇₀ 2)

μ₇₁ 2)

μ₇₂ 2)

μ₇₃ 2)

μ₇₄ 2)

μ₇₅ 2)

μ₇₆ 2)

μ₇₇ 2)

μ₇₈ 2)

μ₇₉ 2)

μ₈₀ 2)

μ₈₁ 2)

μ₈₂ 2)

μ₈₃ 2)

μ₈₄ 2)

μ₈₅ 2)

μ₈₆ 2)

μ₈₇ 2)

μ₈₈ 2)

μ₈₉ 2)

μ₉₀ 2)

μ₉₁ 2)

μ₉₂ 2)

μ₉₃ 2)

μ₉₄ 2)

μ₉₅ 2)

μ₉₆ 2)

μ₉₇ 2)

μ₉₈ 2)

μ₉₉ 2)

μ₁₀₀ 2)

转为对偶问题求化:

$$\text{Maximize } \sum \alpha_i - \frac{1}{2} \sum_i \sum_j \alpha_i \alpha_j y^{(i)} y^{(j)} (X^{(i)})^T X^{(j)}$$

$$\text{St. } \sum \alpha_i y^{(i)} = 0 \quad \& \quad \alpha_i \geq 0$$

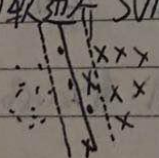
 位于虚线上的点称为支持向量. 上式 $\alpha_i y^{(i)}$ 只有支持向量
 处 α_i 才非 0
 设 SV 是支持向量集合, 则上面的

$$W = \sum_{x^{(i)} \in \text{SV}} \alpha_i y^{(i)} x^{(i)}$$

 有: $y^{(i)} (W^T x^{(i)} + w_0) = 1 \Rightarrow w_0 = y^{(i)} - W^T x^{(i)}$
 此时 w_0 依赖于一个样本, 不稳定, 我们将所有支持向量
 求得的 w_0 平均: $w_0 = \frac{1}{|\text{SV}|} \sum_{x^{(i)} \in \text{SV}} (y^{(i)} - W^T x^{(i)})$
 因此, SVM 的判别函数

$$g(x) = W^T x + w_0$$

$$= \left(\sum_{x^{(i)} \in \text{SV}} \alpha_i y^{(i)} x^{(i)} \right)^T x + \frac{1}{|\text{SV}|} \sum_{x^{(i)} \in \text{SV}} (y^{(i)} - W^T x^{(i)})$$

 choose $\begin{cases} c_1, & \text{if } g(x) > 0 \\ c_2, & \text{else} \end{cases}$
 若为 k 分类问题, 我们看作 k 个 2 类问题, 则需要 k 个
 判别函数 $g_k(x)$
 ② 软边界 SVM: 线性不可分情况

 前面要求所有样本均在虚线之外, 这是一个很
 强的条件, 由于种种原因不能由硬间隔分开
 因此我们允许少量样本位于虚线之内
 甚至分类错误, 并给于这些样本相应的惩罚
 此时的约束: $y^{(i)} (W^T x^{(i)} + w_0) \geq 1 - \xi_i$
 再引入松弛变量 $\xi_i, i=1, 2, \dots, N$

$$\Rightarrow y^{(i)} (W^T x^{(i)} + w_0) \geq 1 - \xi_i$$

可允许样本落入虚
 <1> $\xi_i = 0$ 时, x
 <2> $0 < \xi_i < 1$ 时,
 <3> $\xi_i > 1$ 时,
 定义 soft error
 于是: Minimize

$$\text{St.}$$

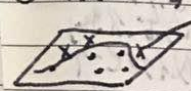
 C 是一个正则化
 构造 Lagrange

$$L_p = \frac{1}{2} \|W\|^2$$

 其对偶问题:

$$\text{Maximize } \sum$$

$$\text{St.}$$

 同样, 构造全
 ③ 核方法

 input space
 在 input space
 构造 $\phi(x)$
 寄希望于在
 核方法中不

$$\xi =$$

 则 g
 即 g

支持向量

线上
支持向量

$x^{(i)}$

k

是一个很
分开
线之内

可允许样本落入虚线之内
 $\langle 1 \rangle \xi_i = 0$ 时, $x^{(i)}$ 分对了, 无惩罚
 $\langle 2 \rangle 0 < \xi_i < 1$ 时, $x^{(i)}$ 分对了, 但在虚线内部, 给予小惩罚
 $\langle 3 \rangle \xi_i > 1$ 时, $x^{(i)}$ 分错了, 给予大惩罚
 定义 soft error: $\sum \xi_i$, 我们希望最小化 soft error
 于是:

$$\text{Minimize } \frac{1}{2} \|w\|^2 + C \sum \xi_i$$

$$\text{s.t. } y^{(i)}(w^T x^{(i)} + w_0) \geq 1 - \xi_i, \xi_i \geq 0, \forall i$$

C 是一个正则化参数, 在边界最大化和训练误差最小化之间折衷
 构造 Lagrange function:

$$L_p = \frac{1}{2} \|w\|^2 + C \sum \xi_i - \sum \alpha_i [y^{(i)}(w^T x^{(i)} + w_0) - 1 + \xi_i] - \sum \mu_i \xi_i$$

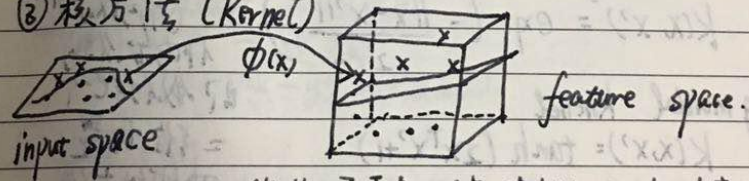
其对偶问题:

$$\text{Max/Minimize } \sum \alpha_i - \frac{1}{2} \sum \sum \alpha_i \alpha_j y^{(i)} y^{(j)} (x^{(i)})^T x^{(j)}$$

$$\text{s.t. } \sum \alpha_i y^{(i)} = 0 \text{ 且 } 0 \leq \alpha_i \leq C, \forall i$$

同样, 通过令偏导为 0 求得参数

② 核方子 (Kernel)



在 input space 线性不可分, 考虑将特征 x 换一个表示方法,
 通过 $\phi(x)$ 映射到 n -空间 (通常维度更高)
 寄希望于在新空间中 $\phi(x)$ 线性可分, 但是 ϕ 是很难确定的
 核方子不需确定 ϕ , 只需通过核函数取考虑内积来衡量

$$\vec{z} = \phi(\vec{x}), \text{ where } z_i = \phi(x_i), i=1, 2, \dots, M$$

$$\text{则 } g(\vec{z}) = \vec{w}^T \vec{z}$$

$$\text{即 } g(\vec{x}) = \vec{w}^T \phi(\vec{x}) = \sum_{j=1}^M w_j \phi_j(x)$$

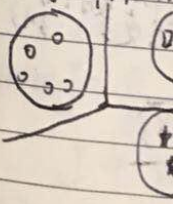
$\frac{1}{2} K(x^{(1)}, x^{(1)})$
 1. $W^{(1)} = X^T K$
 Maximize: $\sum \alpha_i - \frac{1}{2} \sum \alpha_i \alpha_j y_i y_j \phi(x^{(1)})^T \phi(x^{(j)})$
 s.t. $\sum \alpha_i y_i = 0$ & $0 \leq \alpha_i \leq C, \forall i$
 我们无需知道 ϕ 的形式, 只需要核函数 K , 反映 ϕ 的内积
 此时 feature space 称为 "核引导的特征空间"
 核矩阵 $K = [K(x^{(i)}, x^{(j)})]_{i,j=1}^N$, 其对称且正定
 (Kernel/Gram Matrix) 这时, $W = \sum \alpha_i y_i \phi(x^{(i)}) = \sum_{x^{(i)} \in SV} \alpha_i y_i \phi(x^{(i)})$
 于是, $g(x) = W^T \phi(x) = \sum \alpha_i y_i K(x^{(i)}, x) \Rightarrow$ 判别函数
 常用的核函数: $\phi > \text{polynomial}$: $K(x, x') = (x^T x' + 1)^2$
 例如: 二维向量, $\phi=2$ 时, 对于向量 $(x_1, x_2), (x'_1, x'_2)$, 核函数
 $K(x, x') = (x^T x' + 1)^2 = (x_1 x'_1 + x_2 x'_2 + 1)^2$
 $= 1 + 2x_1 x'_1 + 2x_2 x'_2 + 2x_1 x'_1 x_2 + 2x_1 x'_1 x'_2 + (x_1)^2 (x'_1)^2 + (x_2)^2 (x'_2)^2$ 这相当于 $\phi(x)$
 $= (1, \sqrt{2}x_1, \sqrt{2}x_2, x_1^2, x_2^2, \sqrt{2}x_1 x_2)^T$
 即我们提升到了 6 维
 2. $\phi > \text{Radical basis function (RBF) Kernel}$
 $K(x, x') = \exp(-\frac{\|x - x'\|^2}{2\sigma^2})$
 3. $\phi > \text{Sigmoidal Kernel}$
 $K(x, x') = \tanh(2x^T x' + 1)$
 4. $\phi > \text{Structured SVMs}$
 解决多类输出问题
 (1) Multi-Class SVM
 例如, 给出训练集 $\{(x_i, y_i), \dots, (x_N, y_N)\}, y_i \in \{1, 2, \dots, K\}$
 则优化目标: $\min_{w_1, \dots, w_K, \xi} \sum_{k=1}^K \|w_k\|^2 + C \sum_{i=1}^N \xi_i$
 s.t. $\forall j \neq y_i: w_{y_i}^T x_i \geq w_j^T x_i + 1 - \xi_i$
 $\forall j = y_i: w_{y_i}^T x_i \geq w_j^T x_i + 1 - \xi_i$

Joint feature map

<2> Structural
 硬边界: \min_w
 s.t.

软边界: \min
 s.t. \forall

<3> Cutting plane
 由于输出集合
 每次仅包含
 主无监督学习
 监督学习: 给定 $\{x\}$
 无监督学习: 给定 $\{x, y\}$
 聚类
 维度

① 聚类分析
 寻找样本
 词样本


y 不再有意义
 输出, 有歧义

Joint feature map $\phi(x, y)$ 描述了 x 与 y 的关系

No.
Date

<2> Structural SVM

硬边界: $\min_w \frac{1}{2} w^T w$

s.t. $\forall y \in Y | y_1 : w^T \phi(x_1, y_1) \geq w^T \phi(x_1, y) + 1$

\vdots
 $\forall y \in Y | y_k : w^T \phi(x_k, y_k) \geq w^T \phi(x_k, y) + 1$

软边界: $\min_w \frac{1}{2} w^T w + C \sum_{i=1}^N \xi_i$

s.t. $\forall y \in Y | y_1 : w^T \phi(x_1, y_1) \geq w^T \phi(x_1, y) + \Delta(y_1, y) - \xi_1$

\dots

$\forall y \in Y | y_N : w^T \phi(x_N, y_N) \geq w^T \phi(x_N, y) + \Delta(y_N, y) - \xi_N$

<3> Cutting plane Algorithm

由于输出集合 Y 可能非常大, 我们可以定义一个工作集。

每次仅包含部分约束条件。利用工作集去优化

无监督学习之聚类

监督学习: 给定 $\{x^i, y^i\}_{i=1}^N$, 学习 $\hat{y} = f(x; w)$

半监督学习: 给定 $\{x^i\}_{i=1}^N$, 学习 $\hat{y} = f(x; w)$

y { 分类 - y 是类别
回归 - y 是连续值
排序 - y 是序数