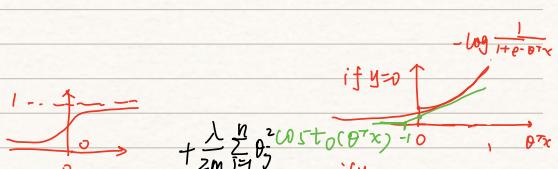


SVM | support Vector machines

optimization objective

回顾: logistic Regression: hypothesis: $h_{\theta}(x) = \frac{1}{1+e^{-\theta^T x}}$



$$\text{cost function: } -y \cdot \log\left(\frac{1}{1+e^{-\theta^T x}}\right) - (1-y) \log\left(\frac{1}{1+e^{-\theta^T x}}\right)$$

logistic 回归和 SVM 的区别:

1. logistic regression:

$$\min_{\theta} \left[\sum_{i=1}^m y^{(i)} \left(\log \frac{1}{1+e^{-\theta^T x^{(i)}}} \right) + (1-y^{(i)}) \left(\log \left[\frac{1}{1+e^{-\theta^T x^{(i)}}} \right] \right) + \frac{\lambda}{2} \sum_{j=1}^n \theta_j^2 \right]$$

① 去掉取平均的分母 (获得梯度) $\frac{1}{m}$ (值不变) $\rightarrow \text{cost}_1(\theta^T x^{(i)})$

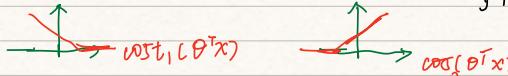
② 更改 cost function (参数化, 即激活函数)

③ 更改形式从 $A + \lambda B$ 转变到 $CA + B$ [C 以视为 $x^T \frac{1}{\lambda}$, $C = \frac{1}{\lambda}$].

hypothesis: $h_{\theta}(x) = \begin{cases} 1, & \theta^T x \geq 0 \\ 0, & \text{else.} \end{cases}$

2. SVM:

$$\text{cost: } \min_{\theta} \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

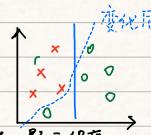


cost function 由 λ 控制
"margin" 由 λ 求

But: SVM 容易受离群值影响; eg:

对策: C 不要太小 (减少 cost_1 的影响)

: C 视为正则项, 正则模型受离群值的影响程度.

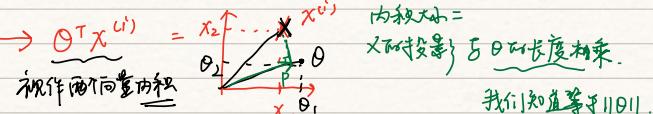


背后的数学原理:

SVM Decision Boundary

$$\min_{\theta} \frac{1}{2} \sum_{j=1}^n \theta_j^2 \rightarrow \frac{1}{2} (\theta_1^2 + \theta_2^2) = \frac{1}{2} (\sqrt{\theta_1^2 + \theta_2^2})^2 = \frac{1}{2} \|\theta\|^2 \quad (\|\theta\| = \sqrt{\theta_1^2 + \theta_2^2})$$

$$\text{s.t. } \theta^T x^{(i)} \geq 1 \quad \text{if } y^{(i)} = 1 \\ \theta^T x^{(i)} \leq -1 \quad \text{if } y^{(i)} = 0$$



一范数 距离该距离 $\|\theta\|$

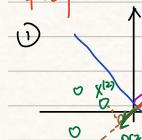
内积大小 =
x下投影与 d 的长度有关.

我们知道等于 $\|\theta\|$.

$$= \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} (x_1^{(i)} x_2^{(i)}) = p^{(i)} \|\theta\|$$

$$= \theta_1 x_1^{(i)} + \theta_2 x_2^{(i)}$$

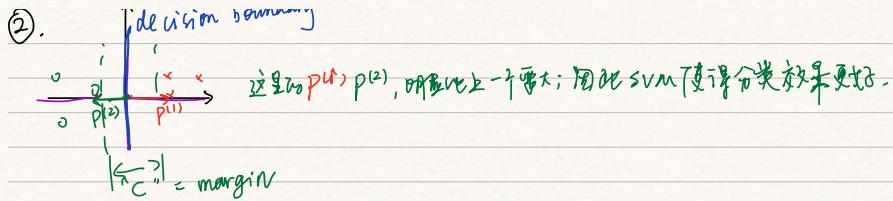
举例:



$$\boxed{y = \theta^T x = 0}$$

1. 求决策边界的法向量 = θ
2. 数据在 θ 方向上法向量即为 p
3. $p \cdot \|\theta\| = \theta^T x^{(i)}$ 要求 $\begin{cases} \geq 1 \\ \leq -1 \end{cases}$

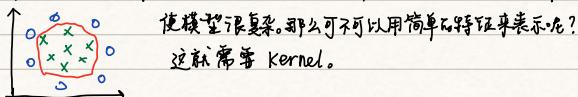
decision boundary.



Kernels

1. 为什么要有 kernel?

在非线性判决边界中，若采用线性回归，则会需要许多高阶特征组合来表示，如： $x_1^2, x_1x_2, x_2^2, x_3^2, \dots$



(注：1D然是“线性可分”)

2. Kernel：引入一些 landmark，其与 x 的相似度从更高维度上提取数据。样本靠近哪个 landmark，那个 landmark 的值就是 1，否则为 0。
然后再从这些 landmark 出发来对样本分类。

Given x : $f_i = \text{similarity}(x, l^{(i)}) = \exp\left(-\frac{\|x - l^{(i)}\|^2}{2\sigma^2}\right)$ 从一些书中找到一些 privilege.

$$\begin{cases} f_1 = \text{similarity}(x, l^{(1)}) = \exp\left(-\frac{\|x - l^{(1)}\|^2}{2\sigma^2}\right) \\ f_2 = \text{similarity}(x, l^{(2)}) = \exp\left(-\frac{\|x - l^{(2)}\|^2}{2\sigma^2}\right) \\ f_3 = \text{similarity}(x, l^{(3)}) = \exp\left(-\frac{\|x - l^{(3)}\|^2}{2\sigma^2}\right) \end{cases}$$

注： $\|x - l^{(i)}\|^2 = \sum_{j=1}^m (x_j - l_j^{(i)})^2$
(landmark也是矩阵)

3. similarity 的意义：就是从人民中选出人民代表！企图用几个 f [包含所有 x]！

若 $x \approx l^{(1)}$, $f_1 \approx \exp(0) = 1$
若 x 离 $l^{(1)}$ 很远： $f_1 \approx \exp(-\infty) = 0$.



4. σ^2 的意义：控制了 [landmark] 影响范围。
① if $\sigma^2=1$, $f = \exp(\frac{\|x - l^{(1)}\|}{2})$ ② if $\sigma^2=4$, 影响减小. ③ if $\sigma^2=0.5$, 更窄.



计算：对每一个样本 $(x^{(i)}, y^{(i)})$ ，计算与所有 landmark 的相似度：

$$X^{(i)} \rightarrow \begin{bmatrix} f_1^{(i)} \\ f_2^{(i)} \\ f_3^{(i)} \\ \vdots \\ f_m^{(i)} \end{bmatrix} = \begin{bmatrix} \text{sim}(x^{(i)}, l^{(1)}) \\ \text{sim}(x^{(i)}, l^{(2)}) \\ \text{sim}(x^{(i)}, l^{(3)}) \\ \vdots \\ \text{sim}(x^{(i)}, l^{(m)}) \end{bmatrix}$$

$$f = \begin{bmatrix} f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_m \end{bmatrix} \in \mathbb{R}^{m+1}, f_0 = 1 \quad (m \text{ 个样本})$$

$X^{(i)} \in \mathbb{R}^{n+1} \text{ or } \mathbb{R}^m \quad (n \text{ 维})$

映射到新空间

SVM with Kernels

1. 原来的 SVM：

hypothesis: $h_0(x) = \begin{cases} 1, & \theta^T x \geq 0 \\ 0, & \text{else.} \end{cases}$

$\theta^T x$ $\in [-\infty, +\infty]$, $h_0(x) = 0/1.$

SVM:

$$\text{cost} = \min_{\theta} C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T x^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T x^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

带 Kernel to SVM:

SVM hypothesis: $h_\theta(x) = \begin{cases} 1, & \theta^T f \geq 0 \\ 0, & \theta^T f < 0 \end{cases}$ $f \in \mathbb{R}^{m+1}$, $\theta \in \mathbb{R}^{m+1}$, $\theta^T f = \theta_0 f_0 + \theta_1 f_1 + \dots + \theta_m f_m$

$$\text{cost} = \min_{\theta} C \sum_{i=1}^m y^{(i)} \text{cost}_1(\theta^T f^{(i)}) + (1-y^{(i)}) \text{cost}_0(\theta^T f^{(i)}) + \frac{1}{2} \sum_{j=1}^n \theta_j^2$$

改动:
① 把原单特征 X 变为 f .
② 将参数的维度从原单特征 X 的 $m+1$ 维, 变成了 f 的 $m+1$ 维.

"Kernel" 的思想, 不应用于其他如逻辑回归中, 是因为计算量太大, 而且由 SVM 应用于 SVM.

正则项 $\sum \theta_j^2 = \theta^T \theta \dots$
可有多种形式.

SVM in parameters

① $C (= \frac{1}{\lambda})$ C 较大 \rightarrow 对前半部分权重大 \rightarrow 低 bias, 高 variance | C 可以理解为高逆 $\frac{1}{\lambda}$
 C 较小 \rightarrow 对前半部分权重小 \rightarrow 低 variance, 高 bias

② σ^2 σ^2 较大, 特征 f_i 变化平滑, f_i 漫温和和: 高 bias, 低 variance: 过拟合时, 增大 σ^2

(similarity $= \frac{\|f_i - f_j\|}{\sigma^2}$) σ^2 较小, f_i 变化陡峭: 低 bias, 高 variance: 欠拟合时, 减少 σ^2 (不用管 $\exp(-)$, 只记住 σ^2 越大, 结果越平滑.) $x > 0$ 单调性一致
 $\exp(x)$ $\exp(-x)$ $\exp(-\frac{x}{\sigma^2})$

SVM 和 Logistic regression 选择

n 代表特征维度, $x \in \mathbb{R}^{n+1}$; m 代表训练样本数.

① 若 $n \gg m$ 或 $n \approx m$ (如: $n=10000$, $m=10, \dots, 1000$) [特征多, 样本少]

\rightarrow 使用逻辑回归, 或不带 Kernel 的 SVM

原因: 特征多, 即使是线性回归也能学的好.

② 若 n 很小, m 中等 (如: $n=1-1000$, $m=10-10000$)

\rightarrow 使用带 Gaussian Kernel 的 SVM (原因: 因为 $n \ll m$, 所以用 kernel 将特征转化为 m 维)

③ 若 n 很大, m 很大 (如: $n=1-1000$, $m=50000+$)

\rightarrow 创造更多的特征 / 用不带 Kernel 的 SVM (否则太高维了!)