SVM

感知机中，inductive bias 中认为二分任务中的边界是线性的
但若 if not linearly seperable 呢？
这里介绍 kernel method (当然方法不限于此)

Def: A kernel K is a legal def of dot-product, i.e., 有一个隐式映射 Φ, s.t. $K(x,y) = \Phi(x)\cdot\Phi(y)$.

Eg. $K(x,y) = (x\cdot y+1)^d$, Φ: n维 $\to$ n^d维

这个定义看起来没啥用；但 perceptron 中因 $x\cdot z$ 是跳不出 linear 的，而 $x\cdot z$ 换为 $K(x,z)$ 就能拥抱高维

Formal Definition: $K(\cdot,\cdot)$ is a kernel if it can be viewed as a legal definition of inner product

· $\exists\phi: X\to\mathbb{R}^N$ s.t. $K(x,z) = \phi(x)\cdot\phi(z)$, range of ϕ: ϕ-space

但 ϕ 是隐式的！这仅是为了方便 view kernel as inner product.

例：$K(x,z) = (x\cdot z)^d$ 对应：$\Phi(x) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)$ ($x=(x_1,x_2)$)

可见 $\Phi: \mathbb{R}^2\to\mathbb{R}^3$, $\Phi(x)\cdot\Phi(z) = (x_1^2, x_2^2, \sqrt{2}x_1x_2)\cdot(z_1^2, z_2^2, \sqrt{2}z_1z_2)$

$= (x_1z_1+x_2z_2)^2 = (x\cdot z)^2 = K(x,z)$

※：同一个 Kernel，Φ 也许并不唯一！ Eg: $(x_1,x_2)\to\Phi(x) = (x_1^2, x_2^2, x_1x_2, x_2x_1)$

应避免显式地扩到高维，i.e., Φ！因为 feature space 会长的飞快！！

因此可以尝试用 $K(x,z)$ 代替 $x\cdot z$。例如 Perceptron 中：

$w_t = a_{i1}x_{i1} + \cdots + a_{ik}x_{ik}$ (可视为未正确分类的 sample 的线性组合).

对于第 t 的错误： mistake on positive : $a_{it}\leftarrow 1$, store x_{it}

mistake on negative : $a_{it}\leftarrow -1$, store x_{it}

$w_t\cdot x = a_{i1}K(x_{i1},x)\cdots + a_{ik}K(x_{ik},x)$

No.

Date . .

这个方法在margin 恰是Φ维空间中 linear boundary 时表现很好！

可知：若margin 在Φ-space 中，则 Perceptron makes $\left(\frac{R}{\gamma}\right)^2$ mistakes

Kernel Example:

Linear: $K(x,z) = x \cdot z$

Polynomial: $K(x,z) = (x \cdot z)^d$ or $K(x,z) = (1 + x \cdot z)^d$

Gaussian: $K(x,z) = \exp\left[-\frac{\|x-z\|^2}{2\sigma}\right]$

Laplace: $K(x,z) = \exp\left[-\frac{\|x-z\|}{2\sigma}\right]$

Properties of Kernel: K is kernel iff: ① K is symmetric

② For any training points $x_1, \cdots, x_m$, and $\forall a_1, \cdots, a_m \in \mathbb{R}$:

$$\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$$

i.e., $K = (K(x_i, x_j))_{i,j=1,\cdots,n}$ 是半正定的：$a^T K a \geq 0$

若 $K_1(\cdot,\cdot)$，$K_2(\cdot,\cdot)$ 均为kernels，则 $\forall c_1, c_2 \geq 0$，$K(x,z) = c_1 K_1(x,z) + c_2 K_2(x,z)$ 也是kernel！且 $K_1(\cdot,\cdot)\, K_2(\cdot,\cdot)$ 也是！ $K_2(x,z)$ $K_1(x,z)$ $\phi_2(z)$

Proof: $\phi(x) = (\sqrt{c_1}\phi_1(x), \sqrt{c_2}\phi_2(x))$，$\Phi(x)\Phi(z) = c_1 \underline{\phi_1(x)\phi_1(z)} + c_2 \underline{\phi_2(x)}$

$\phi(x) = (\phi_{1,i}(x)\ \phi_{2,j}(x))\ i \in \{1, \cdots, n\}\ j \in \{1, \cdots, m\}$

$$\phi(x)\phi(z) = \sum_{i,j} \phi_{1,i}(x)\phi_{2,j}(x)\phi_{1,i}(z)\phi_{2,j}(z)$$

$$= \underbrace{\sum_i \phi_{1,i}(x)\phi_{1,i}(z)}_{K_1(x,z)} \left(\underbrace{\sum_j \phi_{2,j}(x)\phi_{2,j}(z)}_{K_2(x,z)}\right)$$

综合上述所有铺垫，SVM，Support Vector Machine 问世！

Def: Margin of example x w.r.t. linear seperation w is the distance from x to plane $w \cdot x = 0$

Def: Margin γ_w of a set S is min margin over $x \in S$, w.r.t. a linear separator w.

Def: Margin γ of set S is the maximum γ_w over all linear sep. w

因此 SVM目标为: Input: $S=\{(x_1,y_1)\cdots(x_m,y_m)\}$

Find: some w and largest $\boxed{\text{maximum } \gamma}$ where:

① $\|w\|^2=1$ ② for all i, $y_i w\cdot x_i \ge \gamma$

Output: Maximum marginal separator

等效为: $y_i\cdot\frac{w}{\gamma}\cdot x_i \ge 1$, 令 $w'=w/\gamma$. 可见是 $y_i w'\cdot x_i \ge 1$

满足前提下: minimize $\|w'\|^2$, i.e.,

$$\arg\min_w \|w\|^2, \quad \text{s.t.}, \quad \forall i, \ y_i w\cdot x_i \ge 1$$

Dual Problem*: (Equivalent to:)

$$\max_\alpha \sum \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i\alpha_j y_i y_j \phi(x_i)^T\phi(x_j)$$

$$= \max_\alpha \sum \alpha_i - \frac{1}{2}\sum_{i=1}^m \sum_{j=1}^m \alpha_i\alpha_j y_i y_j K(x_i,x_j)$$

$$\text{s.t.} \sum_{i=1}^m \alpha_i y_i = 0, \quad \alpha_i \ge 0$$

最终: classifier 为: $w=\sum_i \alpha_i y_i x_i$

*: 这一坨式子咋得到的?

在最大化间隔同时, 允许部分样本不满足约束(尽可能少)

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m l_{0/1}(y_i(w^Tx_i+b)-1), \quad C: \text{constant}$$

$$l_{0/1}(z)=\begin{cases}1 & \text{if } z<0\\ 0 & \text{otherwise}\end{cases}$$, 取 hinge loss: $l_{hinge}(z)=\max(0,1-z)$

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \max(0, 1-y_i(w^Tx_i+b))$$, 引入slack变量 $\xi_i \ge 0$:

$$= \min_{w,b,\xi_i} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^m \xi_i \quad ①$$

三个约束条件

$$\text{s.t. } y_i(w^Tx_i+b) \ge 1-\xi_i \ ②, \quad \xi_i \ge 0 \ ③$$

No.

Date

则构造 Lagrange: $L(w,b,\alpha,\xi,\mu)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\xi_i$
$+\sum_{i=1}^{m}\alpha_i(1-\xi_i-y_i(w^Tx_i+b))-\sum_{i=1}^{m}\mu_i\xi_i$

$\frac{\partial L}{\partial w}=0 \Rightarrow w=\sum_{i=1}^{m}\alpha_i y_i x_i$ ①

$0=\sum_{i=1}^{m}\alpha_i y_i$ ② $C=\alpha_i+\mu_i$ ③ 代回去: *

Dual Problem: $\max_{\alpha}\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$ $=K(x_i,x_j)$

s.t. $\sum_{i=1}^{m}\alpha_i y_i=0$, $0\le\alpha_i\le C$, $i=1,2,\cdots,m$

这是 $K(x,z)=x\cdot z$ 时; 若为 $\Phi(x)$, 则 $\Phi(x_i)^T\Phi(x_j)=\Phi(x_i)\cdot\Phi(x_j)$

△: 附: 使用拉格朗日乘子法得到的便是"对偶问题"

*: $\inf_{w,b} L(w,b,\alpha)=\frac{1}{2}w^Tw+\sum_{i=1}^{m}\alpha_i-\sum_{i=1}^{m}\alpha_i y_i w^T x_i-\sum_{i=1}^{m}\alpha_i y_i b$

$=\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y_i x_i-w^T\sum_{i=1}^{m}\alpha_i y_i x_i+\sum_{i=1}^{m}\alpha_i-b\sum_{i=1}^{m}\alpha_i y_i$

$=-\frac{1}{2}w^T\sum_{i=1}^{m}\alpha_i y_i x_i+\sum_{i=1}^{m}\alpha_i$

$=-\frac{1}{2}\sum_{i=1}^{m}\alpha_i y_i x_i^T\sum_{i=1}^{m}\alpha_i y_i x_i+\sum_{i=1}^{m}\alpha_i$

$=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y_i y_j x_i^T x_j$

Campus

扫描全能王 创建