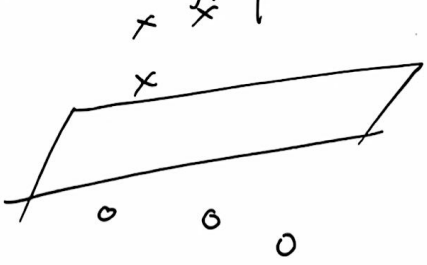
SVM 推导与总结：

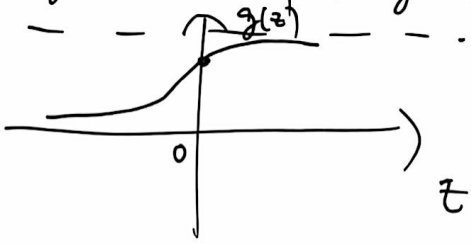
SVM support vector machine.

在 classifier 中，我们会确定一条 boundary，即 $h_\theta(x)$ 会与这个 boundary 密切相关。在高维空间中，这个 boundary 就是一个 hyperplane.



因此，我们可以将这个面作一个重要的决策依据。

在 Sigmoid 函数中，$g(z)$ 形如



z 由 $\theta^T x$ 拟合，

直观感受是，$|\theta^T x|$ 越大，即离 $z=0$ 越远，我们越有把握认为 $\hat{y} = g(\theta^T x)$ 处于其中某一状态，假如 $|\theta^T x|$ 很小，我们判别越模糊。

因此引入新的优化目标：

（我们此时 ~~$y \in \{0,1\}$~~ 而是 $y \in \{-1, 1\}$）

我们希望 ~~$y=0$~~ $y=-1$ 时 ~~尽可能~~ $\theta^T x$ 尽可能小（为负）

$y=1$ 时 $\theta^T x$ 尽可能大（为正）

因此提出 functional margin

$$\gamma = y^{(i)}(\theta^T x)$$

如果将 $x_0=1$ 项分离出来，

$$\gamma = y(w^T x + b)$$

但这会引出一个问题，我们知道

当 $\lambda \neq 0$ 时 $a(w^T x + b) = 0$ 是同一个超平面，但

$$\gamma = y(w^T x + b)$$
$$= \lambda y (w^T x + b)$$

我们可以依通过不断放缩 (w^T, b) 提高 γ，这能实现优化目标"（提高 γ）但达不到目的.（提高预测的精准度）

我们可以通过规范化 w^T ~~b~~ 克服上述的漏洞.

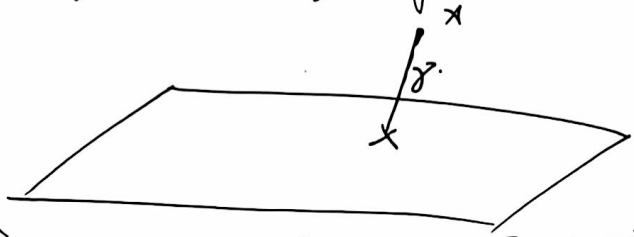
$$\hat{\gamma} = y\left(\frac{w^T}{\|w\|} x + \frac{b}{\|w\|}\right)$$

此时 w^T, b 缩放将不再影响 γ

同时，此时 $\frac{w^T x + b}{\|w\|}$ 就刚好是 geometric margin.

---

geometric ~~mg~~ margin 计算.



首先，~~w^T~~ $\vec{w}$ 表示与 $w^T x + b = 0$ 正交的.

假设 $w^T x + b$ 上有两点 x_1, x_2

$$\begin{cases} w^T x_1 + b = 0 \\ w^T x_2 + b = 0 \end{cases} \Rightarrow w^T(x_1 - x_2) = 0.$$

说明 $\vec{w}$ 与 $w^T x + b = 0$ 正交.

因此，假设 x 在 w^Tx+b 上投影的点是 p.

有 $\vec{x}-\vec{p}=\gamma\frac{\vec{w}}{\|w\|}$ ← 一个单位向量.

$\vec{p}=\vec{x}-\gamma\frac{\vec{w}}{\|w\|}$

$\vec{p}$ 满足 $w^Tx_p+b=0$.

$w^T\left(x-\gamma\frac{w}{\|w\|}\right)+b=0$.

$w^Tx-\gamma\frac{w^Tw}{\|w\|}+b=0$.

$\gamma=\frac{w^Tx+b}{\|w\|}$

然后，我们有 SVM 的核心算法.

optimal margin classifier.

我们在考虑一组样本 $(x^{(i)},y^{(i)})$ 时，我们希望找到一个 hyperplane，将 $\{x^{(i)}\mid y^{(i)}=1\}$ 和 $\{x^{(i)}\mid y^{(i)}=-1\}$ 分离开。此时，我们要选取一个合适的 $f(0)$，这样我们能更好地迭代优化.

在 SVM 中，我们不再关心样本数据所服从的某种分布，而是关心算法如何才能在 train set 上有更好的预测的把握。

更甚，我们更加关心那些离 hyperplane 更近的点，我们更希望在这种边缘上上有更好的准确率.

于是，我们想优化下面的问题：

$\gamma=\min\hat{\gamma}$

我们要通过调整 w 与 b 使 $\max\gamma$，

即 $\max_{w,b}\gamma$

s.t. $\forall i\ \frac{y^{(i)}(w^Tx^{(i)}+b)}{\|w\|}\geq\gamma$

即 $\max_{w,b,\gamma}\gamma$

s.t. $\forall i\ y^{(i)}(w^Tx^{(i)}+b)\geq\gamma$

$\|w\|=1$

$\|w\|=1$ 的约束条件并不好实现，

我们希望约束条件是一个 convex 的点集或者 affine 的超平面，这样可以很好地使用 Lagrange dual.

即 $\max_{w,b,\gamma}\ \frac{\hat{\gamma}}{\|w\|}$ geometric margin.

$\forall i\ y^{(i)}(w^Tx^{(i)}+b)\geq\hat{\gamma}$

functional margin.

因为 $\hat{\gamma}$ 是与 $\|w\|$ 密切相关的，我们总能通过 scale $\|w\|$ 找到合适的 (w,b) 满足 $\hat{\gamma}=1$，且这不影响 $\frac{\hat{\gamma}}{\|w\|}$ (geometric margin)

于是有

$\max_{w,b,\gamma}\ \frac{1}{\|w\|}$

s.t. $\forall i\ y^{(i)}(w^Tx^{(i)}+b)\geq 1$

$\Leftrightarrow\ \min_{w,b,\gamma}\ \frac{1}{2}\|w\|_2^2$

s.t. $\forall i\ y^{(i)}(w^Tx^{(i)}+b)\geq 1$

注意，此时，我们已经达到目标：

优化的目标函数 $\frac{1}{2}\|w\|_2^2$ 是 convex 的，

限制条件是 affine 的面

于是，我们有

$$\mathcal{L}(w,b,\alpha) = \frac{1}{2}\|w\|_2^2 + \sum_{i=1}^{m}\alpha_i(-y^{(i)}(w^T x^{(i)}+b)+1)$$

where $\forall i \quad \alpha_i \geq 0$.

primal：

$$\min_{w,b}\max_{\alpha}\mathcal{L}(w,b,\alpha)$$

dual：

$$\max_{\alpha}\min_{w,b}\mathcal{L}(w,b,\alpha).$$

我们先解决dual问题（后面再证明这个问题是符合KTT条件的 $\Leftrightarrow p^* = d^*$）

$$\nabla_w \mathcal{L}(w,b,\alpha)$$

$$= w + \sum_{i=1}^{m}\alpha_i(-y^{(i)}x^{(i)})$$

$$= w - \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)} = 0 \quad ①$$

$$\nabla_b \mathcal{L}(w,b,\alpha)$$

$$= \sum_{i=1}^{m}\alpha_i(-y^{(i)}) = 0 \quad ②$$

由①得 $w^* = \sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}$ ③

将①③代入 $\mathcal{L}(w,b,\alpha)$

~~$\mathcal{L}(w,b,\alpha)$~~

$$\min_{w,b}\mathcal{L}(w,b,\alpha)$$

$$= \frac{1}{2}\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)T}\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)}$$

$$+ \sum_{i=1}^{m}\alpha_i(-y^{(i)})\sum_{i=1}^{m}\alpha_i y^{(i)} x^{(i)T} x^{(i)} + \sum_{i=1}^{m}\alpha_i$$

这里又混了，重写一下

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

整理下

$$\min_{w,b}\mathcal{L}(w,b,\alpha)$$

$$= \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)}\rangle.$$

剩下的问题是

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} y^{(i)} y^{(j)} \alpha_i \alpha_j \langle x^{(i)}, x^{(j)}\rangle$$

$$\text{s.t} \quad \forall i \ \alpha_i \geq 0.$$

$$\sum_{i=1}^{m}\alpha_i y^{(i)} = 0.$$

以上的算法是对那些离 boundary 很近的点非常敏感的（那些点就是 support vector）

这会带来一些问题

① 在拉格朗日对偶问题中，$\max_{\alpha}\min_{w,b}\mathcal{L}(w,b,\alpha)$ 在限制被违反时，会得到无穷

② 该算法受support vector的影响非常大，一个表现奇异的点，会使边界产生很大的变动。

所以会产生一定的想法，容忍一些点有少量的违背条件：

$$\min_{w,b,\gamma,\xi} \frac{1}{2}\|w\|_2^2 + C\sum \xi_i$$

$$\text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1 - \xi_i \quad \text{for } i = 1, 2, \ldots, m$$

且 $\xi_i \geq 0$ for $i = 1, 2, \ldots, m$

这里的C是个hyperparameter，做为惩罚。

得到 Lagrangian.

$$\mathcal{L}(w, b, \xi, \alpha, \beta)$$

$$= \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i(-y^{(i)}(w^T x^{(i)} + b) + (-\xi_i)) + \sum_{i=1}^{m}\beta_i(-\xi_i)$$

$$\nabla_w \mathcal{L} = w + \sum_{i=1}^{m}\alpha_i(-y^{(i)})x^{(i)} = 0 \quad ①$$

$$\nabla_b \mathcal{L} = \sum_{i=1}^{m}\alpha_i(-y^{(i)}) = 0 \quad ②$$

$$\frac{\partial}{\partial \xi_i}\mathcal{L} = C + \alpha_i(-1) + \beta_i(-1) = 0. \quad ③$$

将 ① ② ③ 代入

$$W(\alpha) = \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$

转化后的问题

$$\max_{\alpha} \sum_{i=1}^{m}\alpha_i - \frac{1}{2}\sum_{i,j}^{m} y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)}, x^{(j)}\rangle$$

s.t. $0 \le \alpha_i \le C$

$$\sum_{i=1}^{n}\alpha_i y^{(i)} = 0.$$