SVM

假如分类函数 $g(x)$ 的类别 y 可取值为 $\{-1, 1\}$，则：

$$g(x) = \begin{cases} -1 & x<0 \\ 1 & x\geq 0 \end{cases}$$

则目标函数 $h_\theta(x) = g(\theta^T x)$，在假设：数据集是线性可分，下，目标函数变为：$h_{w,b}(x) = g(w^T x + b)$

而 w、b 可以唯一确定一个超平面.

一个点 $(x^{(i)}, y^{(i)})$ 到由 w.b 确定的超平面的函数间隔为：

$$\hat{\gamma}^{(i)} = y^{(i)}(w^T x^{(i)} + b)$$

可知：对于正确分类的数据点，函数间隔不小于0

超平面与整个数据集合的距离是：

$$\hat{\gamma} = \min_i \hat{\gamma}^{(i)}$$

函数间隔的问题在于只要成倍增大 w、b，就能使函数间隔变大，为了解决这个问题，就有了几何间隔的定义：

几何间隔：$\max_{w,b} \gamma \quad s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq \gamma$ 且 $\|w\| = 1$

即：在 $\|w\|=1$ 条件下函数间隔最小值.

离超平面最近点的距离定义为 $\frac{1}{\|w\|}$

最优间隔分类器，数学表示：

$$\max_{\gamma,w,b} \gamma \quad : s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq \gamma,\ i=1,2,3,\dots m$$
$$\|w\| = 1$$

因为约束是非凸性约束，所以该问题不易解，它的最优解容易达到局部最优

⇓ 转换为凸性问题，$\boxed{\text{几何间隔 } \gamma = \frac{\hat{\gamma}}{\|w\|} \text{ 函数间隔}}$

① $\max_{\gamma,w,b} \frac{\hat{\gamma}}{\|w\|} \Rightarrow$ ② 令 $\hat{\gamma}=1$（这与 $\|w\|=1$ 的原理一样，都是缩放 (w,b)），得到：$\max_{\gamma,w,b} \frac{1}{\|w\|}$

③ 前的极大值与 $\frac{1}{2}\|w\|^2$ 的极小值等同，得：$\boxed{\min_{w,b} \frac{1}{2}\|w\|^2 \quad s.t.\ y^{(i)}(w^T x^{(i)} + b) \geq 1,\ i=1,2,\dots m}$

得到了凸问题，

接下来，需要先将约束转为方程的一部分，即需要 $\boxed{\text{拉格朗日}}$ 方程.

拉格朗日对偶(Lagrange duality)

先抛开上面的问题，看看存在等式约束的极值问题解法，比如下面的问题：

A、 $\min_w f(w) \quad$; s.t. $h_i(w)=0$, $i=1,2,\dots l$

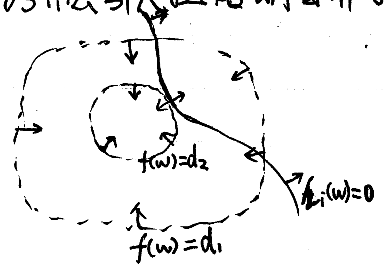
目标函数为 $f(w)$，后面为约束条件，通常解法是引入拉格朗日算子，得到对应的拉格朗日公式为：

$$\mathcal{L}(w,\beta)=f(w)+\sum_{i=1}^{l}\beta_i h_i(w)$$

其中：(β 为拉格朗日算子)
(l 为等式约束的个数)

然后分别对 w. β 求偏导，使偏导数为0，然后解出 w. β.

为什么引入拉格朗日算子可以求出极值？如下图辅助思考：



其中：虚线是目标函数 f 的等高线，实线为约束 h

箭头表示法线方向

我们很容易发现：在最优解处，f 和 h 的斜率平行。

接下来看下有不等式约束的极值问题解法：问题如下：

B. $\min_w f(w) \quad$ s.t. $g_i(w)\le 0 \quad i=1,2,\dots k$

$h_i(w)=0$, $i=1,2,\dots l$

定义一般化的拉格朗日公式：$\mathcal{L}(w,\alpha,\beta)=f(w)+\sum_{i=1}^{k}\alpha_i g_i(w)+\sum_{i=1}^{l}\beta_i h_i(w)$

其中的 α. β 都是拉格朗日算子，如果按上述公式求解会出问题，因为我们要求的是最小值，而这里的 $g(w)$ 已经不是0了，我们将 α 调成很大的正值，来使最后的函数结果为负无穷。为了排除这种情况，我们定义了下面的函数：

$$\theta_p(w)=\max_{\alpha,\beta;\alpha_i\ge 0}\mathcal{L}(w,\alpha,\beta)$$

其中 P 代表 primal

精妙处在于：$\alpha\ge 0$，求极大值

假设：$g_i(w)>0$ 或 $h_i(w)\ne 0$，那么总可以调整 α_i. β_i 使得 $\theta_p(w)$ 有最大值为正无穷，只有当 g. h 满足约束时 $\theta_p(w)$ 为 $f(w)$。因此有：

$$\boxed{\theta_p(w)=\begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise}\end{cases}}$$

C. 即有： $\min_w f(w)=\min_w \theta_p(w)=\min_w \max_{\alpha,\beta;\alpha\ge 0}\mathcal{L}(w,\alpha,\beta)$

如果直接对上式求解，首先要面对两个参数，而 α 也是不等式约束，再在 w 上求最小值，这个过程不容易，所以引入 $\boxed{\text{对偶问题}}$

No.

Date

对偶问题：

假设：$\theta_D(\alpha,\beta)=\min_{w} l(w,\alpha,\beta)$，其中D表示对偶。

它将问题转化为，将α,β看作固定值，去求解拉格朗日关于w的最小值，之后，再对$\theta_D(\alpha,\beta)$求最大值的话：

$$\boxed{\max_{\alpha,\beta;\alpha_i\geq 0}\theta_D(\alpha,\beta)=\max_{\alpha,\beta;\alpha_i\geq 0}\min_{w} l(w,\alpha,\beta)}$$

这就是原问题的对偶问题。

一般的有：$\max\min(x)\leq\min\max(x)$，

所以有：$$\boxed{d^*=\max_{\alpha,\beta;\alpha_i\geq 0}\min_{w} l(w,\alpha,\beta)\leq\min_{w}\max_{\alpha,\beta;\alpha_i\geq 0} l(w,\alpha,\beta)=p^*}$$

在假设：

1. f是凸函数
2. 约束不等式g是凸函数（线性函数都是凸函数）
3. 约束等式h是仿函数（仿函数定义：$h(w)=w^Tx+b$，几乎和线性等价，只不过允许截距b存在）

条件下，一定存在：w^*,α^*,β^*，使得w^*是原始问题的解，α^*,β^*是对偶问题的解，且$p^*=d^*=l(w^*,\alpha^*,\beta^*)$。这样的$w^*,\beta^*,\alpha^*$需要满足KKT条件：

$$\boxed{\begin{aligned}
&\frac{\partial}{\partial w_i}l(w^*,\alpha^*,\beta^*)=0, && i=1,2,\cdots n \quad ①\\
&\frac{\partial}{\partial \beta_i}l(w^*,\alpha^*,\beta^*)=0, && i=1,2,\cdots l \quad ②\\
&\alpha_i^* g_i(w^*)=0, && i=1,2,\cdots K \quad ③\\
&g_i(w^*)\leq 0, && i=1,2,\cdots K \quad ④\\
&\alpha^*\geq 0, && i=1,2,\cdots K \quad ⑤
\end{aligned}}$$

$\left\{\begin{array}{l}\alpha\text{是不等式约束的拉格朗日算子}\\ \beta\text{是等式约束的拉格朗日算子}\end{array}\right.$

再次审视公式③，它被称为KKT dual complementarity条件，隐含了：

如果$\alpha^*>0$，那么$g_i(w^*)=0$，就是说，$g_i(w^*)=0$时，w处于可行域的边界上，这时才是起作用的约束。

而其它位于可行域内（$g_i(w^*)<0$的）点都是不起作用的约束，其$\alpha^*=0$。

这个KKT条件会用来解释支持向量和SMO的收敛测试。

No.

Date

最优间隔分类器：

回到原始问题：

$$\min_{w,b} \frac{1}{2}\|w\|^2 \qquad s.t.\ g_i(w) = -y^{(i)}(w^T x^{(i)}+b)+1 \le 0,\ i=1,2\ldots m$$

对应的拉格朗日方程为：

$$l(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}(w^T x^{(i)}+b)-1\right] \qquad ⑥$$

其后，按对偶问题的求解方法：（该问题符合 $p^*=d^*$ 的假设）⑦

$$d^* = \max_{\alpha,\beta;\alpha_i\ge 0}\ \min_{w} l(w,\alpha,\beta)$$ ，因为原始问题没有等式约束，所以拉格朗日乘子[illegible]

首先：先固定 α，以 w,b 为变量，最小化 l；最小化 l 时，求解 l 对 w,b 的偏导，并将导数设为 0，可以得到：

$$\nabla_w l(w,b,\alpha) = w-\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}=0 \Rightarrow w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)} \qquad Ⓐ$$

$$\frac{\partial}{\partial b} l(w,b,\alpha) = \sum_{i=1}^{m}\alpha_i y^{(i)}=0 \qquad Ⓑ$$

将ⒶⒷ代入⑥可得：

$$l(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}(w^T x^{(i)}+b)-1\right]$$

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

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

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

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

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

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

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

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

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

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

得到：$l(w,b,\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j(x^{(i)})^T x^{(j)}$ ⟵ 最小化 $l(w,b,\alpha)$ 得到

接下来是 $d^*=\max_{\alpha,\beta;\alpha_i\ge 0}\min_{w} l(w,b,\alpha)$ 的极大化过程：

$(x^{(i)})^T x^{(j)}$ 可以表示为 $\langle x^{(i)},x^{(j)}\rangle$ 向量内积。

$$\boxed{\begin{aligned}\max_\alpha\ & W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}y^{(i)}y^{(j)}\alpha_i\alpha_j\langle x^{(i)},x^{(j)}\rangle\\ s.t.\ & \alpha_i\ge 0,\ i=1,2\ldots m\\ & \sum_{i=1}^{m}\alpha_i y^{(i)}=0\end{aligned}}$$

Ⓒ ⟵ $\boxed{即\ d^*}$

No.

Date

由⑦可知,($p^* = d^*$),一定存在 w^*, α^* 使得 w^* 是原问题的解,α^* 是对偶问题的解。

对于公式ⓒ来说,求 α_i,就是求解 α^*.

如果求出了 α_i,则根据Ⓐ即 $w = \sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}$ 即可求解出 w(也即 w^*,原始问题的解)

然后,可得:$b^* = -\frac{1}{2}\left(\max_{i:y^{(i)}=-1} w^{*T} x^{(i)} + \min_{i:y^{(i)}=1} w^{*T} x^{(i)}\right)$ ⑧

回到最初的超平面问题:

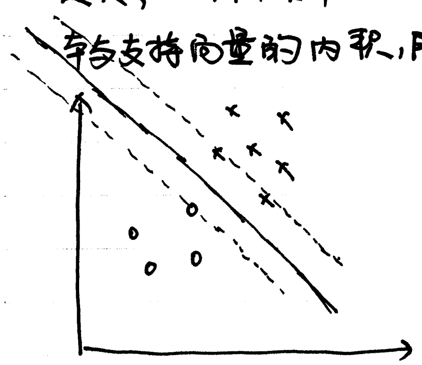
$$\boxed{\begin{aligned} w^T x + b &= \left(\sum_{i=1}^{m} \alpha_i y^{(i)} x^{(i)}\right)^T x + b \\ &= \sum_{i=1}^{m} \alpha_i y^{(i)} \langle x^{(i)}, x \rangle + b \end{aligned}}$$

也就是说,以前新来的样本要分类,首先根据 w 和 b 做一次线性运算,然后再看正负。现在有了 α_i,我们不需要求解 w,只需将新来的样本和训练数据中的所有样本做内积即可。

问题来了、与前面所有的样本都做运算不是很耗时吗?

其实,从KKT条件可知,只有支持向量的 $\alpha_i > 0$ 其它情况 $\alpha_i = 0$、因此只需要新来样本与支持向量的内积,然后运算即可。

⇕



虚线即为 $g(w)$,虚线上的点 $g_i(w)$ 为 0,则 $\alpha_i > 0$

⇒ 虚线以外的点,$g(w) \neq 0$,则由KKT可知 $\alpha_i = 0$ 即此点在计算中不作贡献

核技法

一般来说，将低维空间上的数据映射到高维空间，可以使数据的线性可分概率变大。而核技法就是映射到高维空间的一种技巧。

定义函数 $\phi(x)$ 为向量之间的映射，一般是从低维到高维。回到上面讲到的简化的最优问题：

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

$$\text{s.t.}\ \alpha_i \geq 0,\ i=1,2\ldots m$$

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

我们将 $\langle x^{(i)}, x^{(j)}\rangle$ 替换为 $\langle \phi(x^{(i)}), \phi(x^{(j)})\rangle$，这样就可以将低维空间上的数据映射到高维空间。

但有些时候，经过 ϕ 映射后的向量的维度过高，导致映射后向量内积的计算复杂度过高。为了解决这个问题，我们引入核函数：$K(x,z) = \langle \phi(x), \phi(z)\rangle$

核函数的意义就在于，定义核函数之后，可以不用明确出映射函数 ϕ，就可以计算两个向量在高维空间中的内积，而且时间复杂度低。

那么，什么样的函数才是正确的核函数呢？核函数是由映射函数乘积得到的，如所以，如果核函数合法，那么必然可以写成两个映射函数乘积的形式。

为了解决这个问题，定义核矩阵。对于一个数据集 $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$，定义一个 $m*m$ 的矩阵 K（K 可代表核函数也代表核矩阵），K 中的每个元素定义为 $K_{ij} = K(x^{(i)}, x^{(j)})$

首先，$K_{ij} = K_{ji}$，核矩阵是对称矩阵，其次，对于任意的 m 维向量 z，可以得到：

$$z^T K z = \sum_i\sum_j z_i K_{ij} z_j = \sum_i\sum_j z_i \phi(x^{(i)})^T \phi(x^{(j)}) z_j = \sum_i\sum_j z_i \Big(\sum_k \phi_k(x^{(i)})\phi_k(x^{(j)})\Big) z_j$$

$$= \sum_k\sum_i\sum_j z_i \phi_k(x^{(i)})\phi_k(x^{(j)}) z_j = \sum_k \Big(\sum_i z_i \phi_k(x^{(i)})\Big)^2 \geq 0$$

因为 z 是任意向量，所以 K 是半正定矩阵。事实上，这是核函数的充分必要条件。

Mercer定理：给定一个 K，$\mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$，那么 K 是合法核的充分必要条件是，对于任意一个有限数据集合，对应的核矩阵是对称半正定矩阵。

常用核函数：高斯核（高斯核对应的映射函数是映射到无限维的）

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

软间隔分类器.

之前讲述的最优间隔分类器,一直强调的是线性可分,但是,当数据是不可线性分割或映射到高维时依然不是线性可分,再或者是线性可分但实际应用中不可避免出现噪声时该怎么办?这里有个比较通用解法.

首先,对原始问题进行变形:

$$\min_{w,b}\ \frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i$$
$$\text{s.t.}\quad y^{(i)}(w^Tx^{(i)}+b)\geqslant 1-\varepsilon_i,\ i=1,2,\cdots m$$
$$\varepsilon_i\geqslant 0,\ i=1,2\cdots m$$

有些数据点可以被允许拥有比小1的几何间距,但要受惩罚,C是惩罚因子是个预设参数.

对其写出对应的拉格朗日方程:

$$\mathcal{L}(w,b,\varepsilon,\alpha,r)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{m}\varepsilon_i-\sum_{i=1}^{m}\alpha_i\left[y^{(i)}(w^Tx^{(i)}+b)-1+\varepsilon_i\right]-\sum_{i=1}^{m}r_i\varepsilon_i$$

按照上文中的对偶问题的推导,先针对 w,b 最小化,然后再针对 α 最大化,得到新的对偶问题,即:

#2

$$\max_\alpha\ W(\alpha)=\sum_{i=1}^{m}\alpha_i-\frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m}\alpha_i\alpha_j y^{(i)}y^{(j)}\langle x^{(i)},x^{(j)}\rangle$$
$$\text{s.t.}\quad 0\leq\alpha_i\leq C,\ i=1,2,\cdots m$$
$$\sum_{i=1}^{m}\alpha_i y^{(i)}=0$$

与#1对比,发现本问题只是对 α_i 做了进一步的约束.求解得到 α 后,w 可以按 $w=\sum_{i=1}^{m}\alpha_i y^{(i)}x^{(i)}$ 给出,但截距b的得到方式要变化.

另一个发生变化的地方在于KKT中的互补条件,现变为:

$$\left.\begin{aligned}\alpha_i=0 &\Rightarrow y^{(i)}(w^Tx^{(i)}+b)\geqslant 1\\ \alpha_i=C &\Rightarrow y^{(i)}(w^Tx^{(i)}+b)\leqslant 1\\ 0<\alpha_i<C &\Rightarrow y^{(i)}(w^Tx^{(i)}+b)=1\end{aligned}\right\}$$ 将用于SMO算法是否收敛的判断.

因为#1是假设数据集线性可分,不符合实际,因此我们主要对#2来求解.

对#2求解的方法:SMO算法.

No.

Date