

支持向量机，被称为最好的监督学习算法

逻辑回归问题

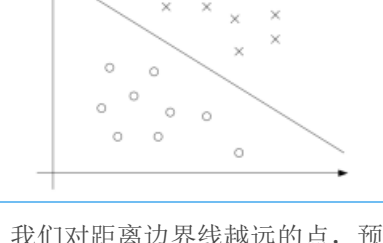
probability $p(y = 1|x, \theta)$ is modeled by $h^\theta(x) = g(\theta^T x)$.

当且仅当 $h^\theta(x) \geq 0.5$ ，也就是 $\theta^T x \geq 0$ 的时候，我们才会预测出“1”。

$\theta^T x$ 越大， $h^\theta(x)$ 也就越大，我们对预测 $h^\theta(x)=1$ 的信心也就越强

总结：如果对一个给定的训练集，我们能够找到一个 θ ，满足当 $y=1$ 时 $\theta^T x \gg 0$, $y=0$ 时 $\theta^T x \ll 0$ ，我们就说这个模型对训练数据拟合的很好

图示



我们对距离边界越远的点，预测正确的信心越大

该线叫做 **分类超平面 (separating hyperplane)**

记号约定

与以前不同，这次我们采用新的记号约定，用 $y \in \{-1, 1\}$ 代替 $y \in \{0, 1\}$ ，用 w, b 代替 θ (b 是截距项，相当于原来的 θ_0 ， w 相当于 $[\theta_1, \dots, \theta_n]^T$)

$$h_{w,b}(x) = g(w^T x + b).$$
$$g(z) = 1 \text{ if } z \geq 0, \text{ and } g(z) = -1 \text{ otherwise.}$$

函数边界和几何边界

(Functional and geometric margins)

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

注：若 $y=1$ ，为了让函数边界很大（也就是预测很准确），那么就需要 $(w \cdot b)$ 是很大的正数
若 $y=0$ ，为了让函数边界很大（也就是预测很准确），那么就需要 $(w \cdot b)$ 是绝对值很大的负数

函数边界代表了确信程度

因为若 $z > 0$ ，则 $g(z)=1$ ；若 $z < 0$ ，则 $g(z)=-1$ ，
所以 $g(w \cdot b) = g(2w \cdot b)$ ，
也就是说让 $w \cdot b$ 扩大一倍，对预测的结果没有任何影响，即预测结果只与式子的正负有关，
但是这样就会改变函数间隔，使其扩大一倍

所以我们会求出无穷组 w, b ，但我们的求解目标是只求唯一组，所以要对 w, b 添加归一化条件，使其解唯一。

最后我们将函数边界定义为所有训练样本中函数边界的最小值

$$\hat{\gamma} = \min_{i=1, \dots, m} \hat{\gamma}^{(i)}.$$

几何边界 (几何间隔)

图中给出了对应 (w, b) 的分类边界，其倾斜方向（即法线方向）为向量 w 的方向。这里的向量 w 是与分类超平面垂直的

图中点A到分类边界的距离就是线段AB的长度，我们记作 $\gamma^{(i)}$

$\gamma/||w||$ 是一个单位长度的向量，指向与 w 相同的方向。

点A表示 $x^{(i)}$ ，可以在分类边界线上找到点B，位置为 $x^{(i)} - \gamma^{(i)} \cdot w/||w||$

因为分类边界线上的所有点都满足 $w^T x + b = 0$

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

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

这只是针对边界线的正向一侧

更广泛的，有 $\gamma^{(i)} = y^{(i)} \left(\left(\frac{w}{||w||} \right)^T x^{(i)} + \frac{b}{||w||} \right)$

最后我们将几何边界定义为所有训练样本中几何边界的最小值

$$\gamma = \min_{i=1, \dots, m} \gamma^{(i)}$$

注意1：如果 $||w|| = 1$ ，那么函数边界 (functional margin) 就等于几何边界 (geometric margin)

注意2：几何边界不受参数缩放的影响，即同时把 w, b 缩小或者放大，几何边界不变（由几何边界的公式也能看出来）
所以我们在对 w, b 进行拟合的时候，为了方便我们可以对 w, b 进行约束，而不会改变什么重要项，比如可以设置 $||w|| = 1$ ，或者 $||w||=5$ ，或者 $||w||=||w_0||=2$ 等等。

replace (w, b) with $(w/||w||_2, b/||w||_2)$

最优边界分类问题

the optimal margin classifier

到目前为止，我们都是假定训练集是线性可分的 (linearly separable)，也就是说能够用分类超平面将样本一分为二

我们的目的是找一个边界，能够最大的把正向和负向样本分隔开，中间留有一段空白区，这个边界就是几何边界

怎么找最大几何边界所对应的那一组 (w, b) 呢，为此提出优化问题

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

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

说明：让 γ 取最大值，也就是使得每一个训练样本的函数边界都至少为 γ （因为 $\min_{i=1, \dots, m} y^{(i)}(w^T x^{(i)} + b) = \gamma$ ，这个约束条件还能保证函数边界与几何边界相等，所以我们就还能够保证所有的几何边界都至少为 γ ）

我们对该问题进行求解，就得到了想要的 (w, b)

然而，“ $||w|| = 1$ ”这个约束条件很讨厌，是非凸的 (non-convex)，而且这个优化问题也明显不是那种我们能够便利给某些标准优化软件 (standard optimization software) 就能解决的，
所以需要改善该问题，例如

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

因为函数边界和几何边界有关系 $\gamma = \hat{\gamma}/||w||$

所以我们还是求的几何边界的最大值，只不过此时的 w 不受 $||w|| = 1$ 的约束了。

然而，此时目标函数仍然不是凸函数，仍然不能直接代入优化软件进行计算，所以继续改写

前面说过，可以缩放 w, b 而不引起实质性改变，所以我们设置 w, b 使得函数边界为1，也就是 $\hat{\gamma} = 1$ 。

将全局的函数边界设置为1，根据函数边界和几何边界的关系，几何边界为 $\hat{\gamma}/||w|| = 1/||w||$

由于求 maximizing $\gamma/||w|| = 1/||w||$ is the same thing as minimizing $\text{pow}(||w||, 2)$

所以得到优化问题

$$\min_{w,b} \frac{1}{2} ||w||^2 \quad \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m$$

这个问题是一个凸二次对象 (a convex quadratic objective)，且仅受线性约束 (only linear constraints)，
对这个问题进行求解，我们就能得到最优边界分类器 (optimal margin classifier)，
这个优化问题的求解可以使用商业二次规划 (commercial quadratic programming，缩写QP)

Support Vector Machines (Part 1)

拉格朗日对偶性

Lagrange duality

先把上述的SVMs和最优边界分类器问题都放到一边，仅谈一谈约束优化问题的求解

给出问题

$$\min_w \quad f(w) \quad \text{s.t.} \quad h_i(w) = 0, \quad i = 1, \dots, l.$$

问题1 (只包含等式约束)

定义拉格朗日乘数法 (Lagrangian)

$$\mathcal{L}(w, \beta) = f(w) + \sum_{i=1}^l \beta_i h_i(w) \quad \text{the } \beta_i \text{'s are the Lagrange multipliers.}$$

求偏导

$$\frac{\partial \mathcal{L}}{\partial w_i} = 0; \quad \frac{\partial \mathcal{L}}{\partial \beta_i} = 0,$$

可以解出 w 和 β

原始优化问题 (primal optimization problem)

$$\min_w \quad f(w) \quad \text{s.t.} \quad g_i(w) \leq 0, \quad i = 1, \dots, k, \quad h_i(w) = 0, \quad i = 1, \dots, l.$$

定义广义拉格朗日函数 (generalized Lagrangian)

$$\mathcal{L}(w, \alpha, \beta) = f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w) \quad \text{the } \alpha_i \text{'s and } \beta_i \text{'s are the Lagrange multipliers.}$$

设有 $\theta_P(w) = \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$ the "P" subscript stands for "primal."

那么， $\theta_P(w) = \begin{cases} \min_{\alpha, \beta: \alpha_i \geq 0} f(w) + \sum_{i=1}^k \alpha_i g_i(w) + \sum_{i=1}^l \beta_i h_i(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$

因此， $\theta_P(w) = \begin{cases} f(w) & \text{if } w \text{ satisfies primal constraints} \\ \infty & \text{otherwise.} \end{cases}$

引出最小化问题 $\min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta)$ 由第四步得知，若满足约束，则此式等同于原始优化问题，也就是有同样的解

我们把这个解定义为 $p^* = \min_w \theta_P(w)$

接下来再看一个不一样的式子

$$\theta_D(\alpha, \beta) = \min_w \mathcal{L}(w, \alpha, \beta) \quad \text{the "D" subscript stands for "dual."}$$

注意：在对 θ_P 的定义中，是找最大值对应的 α, β 取值，这里则是找最小值对应的 w 值。

给出对偶优化问题 (dual optimization problem)

$$\max_{\alpha, \beta: \alpha_i \geq 0} \theta_D(\alpha, \beta) = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \quad \text{该问题与第五步的问题相比，只有min,max的顺序有变动}$$

我们把这个解定义为 $d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \theta_D(w)$

得到 $d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta) \leq \min_w \max_{\alpha, \beta: \alpha_i \geq 0} \mathcal{L}(w, \alpha, \beta) = p^*$ 因为由定理：一个函数的最大的最小值 "max min" 总是小于等于最小的最大值 "min max"

若我们找出 $d=p$ 的条件，就可以通过解该对偶优化问题，也就解出第五步的问题，也就解出原始优化问题

假设 f 和 g 都是凸函数， h 是仿射的 (affine)，并且存在 w^* 使得对于所有的 i , $g_i(w^*) < 0$ ，
在这种假设下，一定存在 α^*, β^* 使得 w^* 是原问题的解， α^*, β^* 是对偶问题的解，并且 $p^* = d^* = \mathcal{L}(w^*, \alpha^*, \beta^*)$

此外， w^*, α^*, β^* 这三个还会满足卡鲁希-库恩-塔克条件 (Karush-Kuhn-Tucker conditions, KKT)

注：等式又叫做KKT对偶互补条件 (dual complementarity condition)，这个等式暗示：当 $\alpha_i > 0$ 的时候，则有 $g_i(w^*) = 0$ 。（也就是说， $g_i(w^*) \leq 0$ 这个约束条件存在的话，则应该是相等关系，而不是不等关系。）

反过来，如果 w^*, α^*, β^* 满足KKT条件，则它们一定是原始优化问题及对偶优化问题的解

接下来我们来看使得 $d=p$ 的条件是什么

最优边界分类器

Optimal margin classifiers

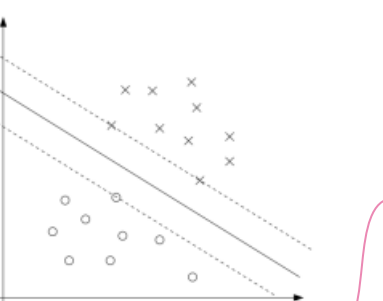
接前面的最优边界分类问题，优化问题3得到公式

$$\min_{w,b} \frac{1}{2} ||w||^2 \quad \text{s.t.} \quad y^{(i)}(w^T x^{(i)} + b) \geq 1, \quad i = 1, \dots, m$$

约束可以写成 $g_i(w) = -y^{(i)}(w^T x^{(i)} + b) + 1 \leq 0$

通过 KKT 对偶互补条件可知，只有训练样本的函数边界确定为 1 的情况下，才有 $\alpha_i > 0$
(这些样本对应的约束条件关系都是等式关系，也就是对应的 $g_i(w) = 0$)

如图



虚线上的点到分类超平面的距离为1，也就是函数边界为1，
 $\alpha_i > 0, g_i(w) = 0$ ，其他点对应的 $\alpha_i = 0$

$\alpha_i > 0$ 的这三个点，叫做支持向量

inner product $\langle x^{(i)}, x^{(j)} \rangle$ (think of this as $(x^{(i)})^T x^{(j)}$)

构造拉格朗日函数

$$\mathcal{L}(w, b, \alpha) = \frac{1}{2} ||w||^2 - \sum_{i=1}^m \alpha_i [y^{(i)}(w^T x^{(i)} + b) - 1] \quad \text{注意：约束中只有 } \alpha_i \text{ 而没有 } \beta_i, \text{ 因为原问题中只有不等式约束，没有等式约束}$$

首先就要使 $\mathcal{L}(w, b, \alpha)$ 取最小值，调整 w 和 b ，而使 α 固定，这样就能得到 0，具体方法也就是令 \mathcal{L} 关于 w 和 b 的导数 (derivatives) 为零。

求关于 w 的导数

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

求关于 b 的导数

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

将 w 的求导结果代入原式得到

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

再将 b 的求导结果代入得到

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

公式只包含了变量 α ，我们求出了 α 就能得到此时的 w 和 b 。

此时的拉格朗日函数是固定 w, b 的条件下函数最小值，也就

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

再求关于 α 的函数最大值，即

$$d^* = \max_{\alpha, \beta: \alpha_i \geq 0} \min_w \mathcal{L}(w, \alpha, \beta)$$

得到原始优化问题的对偶形式

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^m \alpha_i - \frac{1}{2} \sum_{i,j=1}^m y^{(i)} y^{(j)} \alpha_i \alpha_j (x^{(i)})^T x^{(j)}, \quad \text{s.t.} \quad \alpha_i \geq 0, \quad i = 1, \dots, m, \quad \sum_{i=1}^m \alpha_i y^{(i)} = 0.$$

可以证明，这个优化问题的条件是满足 $p^* = d^*$ 和 KKT 条件的 (等式 (3-7))，也就是说，求解出对偶形式，也就解出了原始问题，而对偶形式，就是找到一个 w^* 使得 $\mathcal{L}(w^*, \alpha)$ 取最大值

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

然后根据公式求出 w 和 b

$$b^* = - \frac{\max_{i: y^{(i)} = -1} w^{*T} x^{(i)} + \min_{i: y^{(i)} = 1} w^{*T} x^{(i)}}{2}.$$

由于前面求解得到

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

考虑另一个问题

我们通常考虑问题的出发点是 $w \cdot b$ ，根据求解得到的 α_i ，我们代入原式得到

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

也就是说，以前求来的被分类的样本首先根据 w 和 b 做一次线性运算，然后看求的结果是大于 0 还是小于 0，来判断正例还是负例。
现在看 $f(x)$ ，我们不需要求出 w, b ，只需要求来的样本和训练数据中的所有数据做内积和即可。
那有人会说，与前面所有的样本都做运算是不是太耗时了？
其实不然，我们从 KKT 条件中得到，只有支持向量的 $\alpha_i > 0$ ，其他情况 $\alpha_i = 0$ 。
因此，我们只需要求来的样本和支持向量的内积，然后运算即可。这种写法与下面要提到的核函数 (kernel) 做了很好的铺垫。