大间隔与推广能力

什么叫推广 (generalization)?

学习的目的不仅是要对训练样本能够正确分类，而是要能够对所有可能的样本正确分类，这种能力叫做推广。

什么叫经验风险 (empirical risk)?
在某个参数w下对所有训练样本的分类决策损失称作经验风险。

$$R_{\mathrm{emp}}(w) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f(x_i, w))$$

$f(x_i, w)$ 是判别函数

$L(y, f(x, w))$ 是损失

线性可分的情况下，通过感知器算法，已经能使经验风险为0

什么是期望风险 (expected risk)？

在权值w下未来所有可能出现的样本的错误率或风险

$$R(w) = \int L(y, f(x, w)) \mathrm{d}F(x, y)$$

$F(x, y)$ 表示所有可能出现的样本及其类别的联合概率模型

经验风险只是在给定训练样本上对期望风险的估计

补充：

什么是测试误差？

在测试集上的误差

什么是泛化误差？

是在所有样本总体上的误差，会超出测试集，遇到新样本

什么是过拟合？

过拟合（overfitting）是指过于紧密或精确地匹配特定数据集，以致于无法良好地拟合其他数据或预测未来的观察结果的现象。

什么是欠拟合？

欠拟合（underfitting）它是指相较于数据而言，模型参数过少或者模型结构过于简单，
以至于无法捕捉到数据中的规律的现象。
发生欠拟合时，模型的偏差大，方差小。

什么是VC维？

什么是置信范围？

这些写上的文字课后看

统计学习理论指出，有限样本下，经验风险与期望风险是有差别的，期望风险可能大于经验风险，但它们之间满足下面的规律

$$R(w) \leqslant R_{emp}(w) + \varphi\left(\frac{h}{N}\right) \tag{4-87}$$

其中，$\varphi(h/N)$称作置信范围，它与样本数 N 成反比，而与一个重要的参数 h 成正比。这个参数 h 是依赖于模式识别算法的设计的，称作 VC 维（VC Dimension），它反映了所设计的学习机器（函数集）的复杂性，确切的定义请参考 Vapnik 所著的《统计学习理论的本质》或《统计学习理论》。

式(4-87)给出了有限样本下期望风险的上界。它告诉我们，在训练误差相同的情况下,学习机器的复杂度越低(VC维越低),则期望风险与经验风险的差别就越小,因而学习机器的推广能力就越好。

在线性可分的问题中,我们能得到很多使Remp(w)为0的解,要使方法有最好的推广能力，就应该设法使 $\varphi(h/N)$ 最小。由于训练样本集是给定的,即N固定,能够调整的是算法的VC维。

统计学习理论中的另一个重要的结论是,对于规范化的分类超平面,如果权值满足||w|l≤A,那么这种分类超平面集合的VC维有下面的上界

$$h \leqslant \min([R^2 A^2], d) + 1$$

其中,　　是样本特征空间中能包含所有训练样本的最小超球体的半径,d是样本特征的维数。对于给定的样本集,这两项均是确定的。在求最大间隔分类超平面时,最大化分类间隔也就等价于最小化A^2 ,实际上是使VC维上界最小。根据式(4-87),这样就是试图使期望风险的置信范围尽可能小，即在经验风险都最小化为0的情况下追求期望风险的上界的最小化。

因此,支持向量机中最大分类间隔的准则,是为了通过控制算法的VC维实现最好的推广能力。在这个意义下,所得的分类超平面是最优的。

线性不可分情况

svm内容很深

$$y_i[(w \cdot x_i)+b]-1 \geqslant 0, \quad i=1,2,\cdots,N$$

不可能被所有样本同时满足。

如何补偿呢？

为每个样本引入松弛变量 ξ_k

$$y_i[(w \cdot x_i)+b]-1+\xi_i \geqslant 0$$

假定某个样本 x_k 不满足式(4-90)的条件，即 $y_k[(w \cdot x_k)+b]-1<0$，那么总可以在不等式的左侧加上一个正数 ξ_k，使得新的不等式 $y_k[(w \cdot x_k)+b]-1+\xi_k \geqslant 0$ 成立。

要素1

$$f(x)=\mathrm{sgn}(\vec{w}^T x+b)$$

要素2

$$\min_{w,b,\xi_i} \frac{1}{2}(w \cdot w) + C\left(\sum_{i=1}^{N}\xi_i\right)$$

$$\text{s. t.} \quad y_i[(w \cdot x_i) + b] - 1 + \xi_i \geqslant 0, i = 1,2,\cdots,N$$

且

$$\xi_i \geqslant 0, \quad i = 1,2,\cdots,N$$

要素3

拉格朗日泛涵

$$\min_{w,b,\xi_i} \max_{\alpha} L(w,b,\alpha) = \frac{1}{2}(w \cdot w) + C\sum_{i=1}^{N} \xi_i - \sum_{i=1}^{N} \alpha_i \{[y_i(w \cdot x_i) + b] - 1 + \xi_i\} - \sum_{i=1}^{N} \gamma_i \xi_i$$

$\alpha_i \geq 0$，$\gamma_i \geq 0$ 拉格朗日乘子

解释准则的含义

ξ_i

所有样本的松弛因子之和 $\sum_{i=1}^{N}\xi_i$ 可以反映在整个训练样本集上的错分程度，错分样本数越多，则 $\sum_{i=1}^{N}\xi_i$ 越大；同时，如果样本错误的程度越大（即在错误的方向上远离分类面），则 $\sum_{i=1}^{N}\xi_i$ 也越大。显然，我们希望 $\sum_{i=1}^{N}\xi_i$ 尽可能小。因此，可以在线性可分情况下的目标函数 $\frac{1}{2}\|w\|^2$ 上增加对错误的惩罚项，定义下面的广义最优分类面的目标函数

C

这个目标函数反映了我们的两个目标：一方面希望分类间隔尽可能大（对于分类正确的样本来说），另一方面希望错分的样本尽可能少且错误程度尽可能低。参数 C 是一个常数，反映在这两个目标之间的折中。（注意，这里样本被错分的定义不是 $y_j[(w\cdot x_j)+b]<0$，而是 $y_j[(w\cdot x_j)+b]-1<0$，即第一类样本只要 $g(x)$ 小于 1 就算作错误，第二类样本只要 $g(x)$ 大于 -1 就算作错误。）

C 是一个需要人为选择的参数。通常，如果选择较小的 C，则表示对错误比较容忍而更强调对于正确分类的样本的分类间隔；相反，若选择较大的 C，则更强调对分类错误的惩罚。实际应用中，如果样本线性可分，则 C 的大小只是影响算法的中间过程而不影响最后结果，因为 $\sum_{i=1}^{N}\xi_i$ 最终会为 0。在线性不可分情况下，有时需要试用不同的 C 来达到更理想的结果。

对 w, b, ξ_i 求导 $= 0$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w^* = \sum \alpha_i^* y_i x_i$$

$$\therefore \frac{\partial L}{\partial b} = 0 \Rightarrow \sum y_i \alpha_i^* = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \gamma_i = 0$$

$$C \cdot \sum \xi_i - \sum \alpha_i \xi_i - \sum \gamma_i \xi_i = 0$$

对偶问题

$$\max_{\alpha} Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i\alpha_j y_i y_j (\boldsymbol{x}_i \cdot \boldsymbol{x}_j)$$

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

且

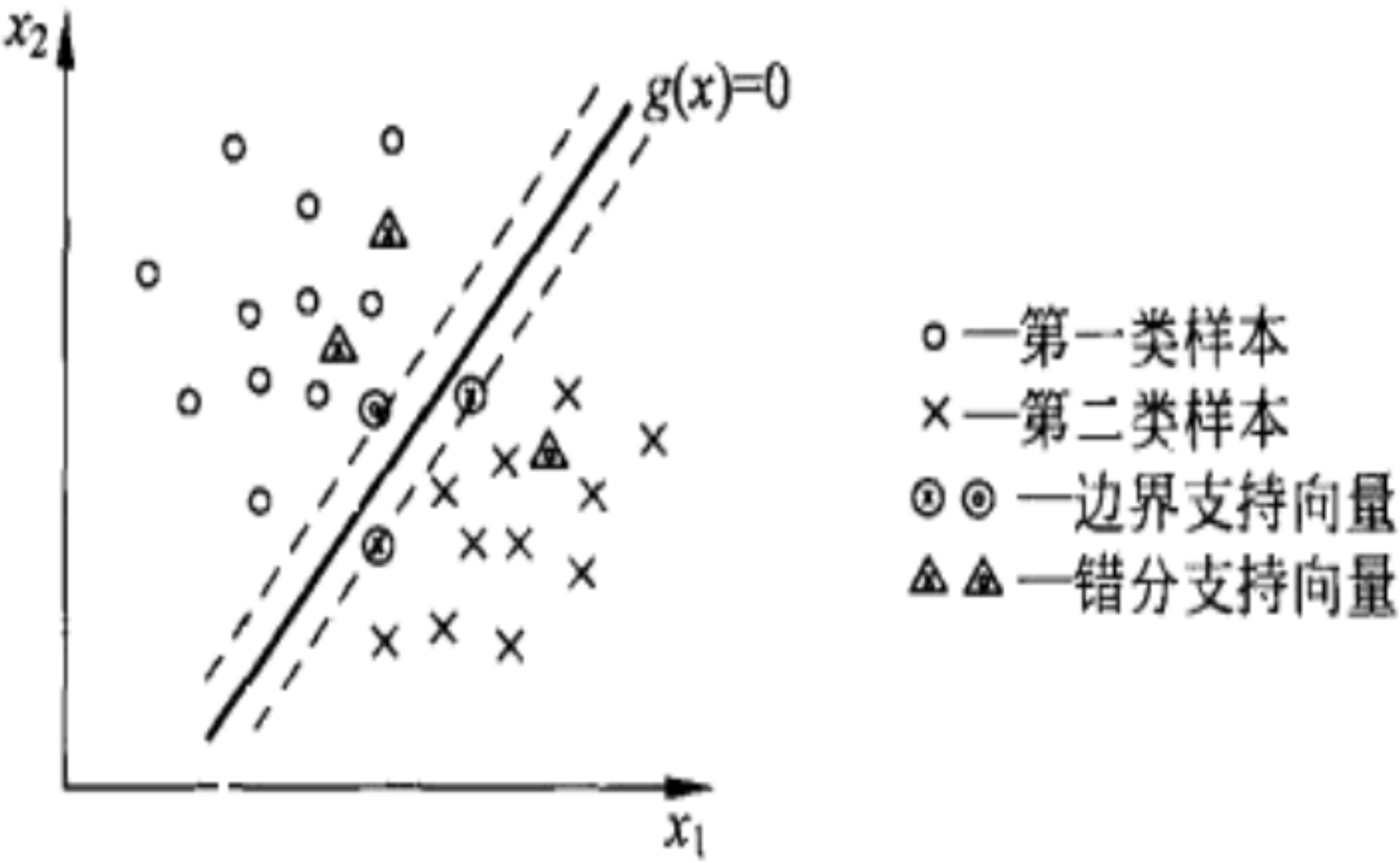
$$0 \leqslant \alpha_i \leqslant C, \quad i = 1,\cdots,N$$

根据库恩-塔克条件,式(4-97)的鞍点满足以下两套条件

$$\alpha_i\{y_i[(\boldsymbol{w}\cdot\boldsymbol{x}_i)+b]-1+\xi_i\} = 0, \quad i = 1,2,\cdots,N$$

$$\gamma_i\xi_i = (C-\alpha_i)\xi_i = 0, \quad i = 1,2,\cdots,N$$

从式(4-104)可以得到,只有对拉格朗日乘子达到上界 $\alpha_i=C$ 的样本才有 $\xi_i>0$,它们是被错分的样本(包括在两条平行的边界面之间的样本),其余样本对应的 $\xi_i=0$。



我们仍然只考虑 $\alpha_i > 0$ 的样本

ξ_i $\begin{cases} >0 & \text{分类错误} \Rightarrow \alpha_i = C \\ =0 & \text{分类正确的但处在边界面上} \Rightarrow 0 < \alpha_i < C \end{cases}$

又叫边界向量：

由于广义最优分类面可以兼容线性可分情况下的最优分类面，所以人们通常采用的支持向量机都是考虑广义最优分类面的形式。