统计学习理论课程内容总结

张超
计研094 2009210934

1 统计学习理论

1.1 函数估计模型

“学习问题可以看做是利用有限数量的观测来寻找待求的依赖关系的问题”[4]。一般地，可以把从样本学习的过程描述为下面的函数估计模型。在有定义概率测度$F(x)$的空间X上，对按照未知但确定的概率分布函数$F(x)$产生随机向量$x \in X$，会按照确定的未知条件概率分布函$F(y|x)$得到一个输出值y；一个学习机器，可以实现一系列函数集$f(x,\alpha), \alpha \in \Lambda$。学习问题就是基于由$l$个符合联合概率分布$F(x,y) = F(x)F(y|x)$ 的独立同分布的观测$(x_1,y_1),\cdots,(x_l,y_l)$构成的训练集，从给定的函数集$f(x,\alpha), \alpha \in \Lambda$ 中选出能够最好逼近$F(y|x)$的函数。

1.2 风险最小化和学习问题的一般表示

为了选择最好的逼近函数，需要定义度量给定输入x时学习机器的输出$f(x,\alpha)$与真实输出y间的损失的函数。记$(X,Y) = Z$，样本$(x_i,y_i) = z_i$，分布$F(x,y) = F(z)$，损失函数集为$Q(z,\alpha) = L(y,f(x,\alpha)), \alpha \in \Lambda$。称损失的数学期望

$$R(\alpha) = \int Q(z,\alpha)\, dF(z) \tag{1}$$

为度量损失的风险泛函。则学习问题变为在训练集上寻找能够最小化$R(\alpha)$的函数$f(x,\alpha_0)$。

1.3 经验风险最小化归纳原则

由于分布$F(z)$未知，常用如下的归纳原则：使用经验风险泛函

$$R_{\text{emp}}(\alpha) = \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \tag{2}$$

来替换风险泛函$R(\alpha)$，并用使经验风险泛函最小的函数$Q(z,\alpha^{'})$来逼近使风险泛函最小的函数$Q(z,\alpha_0)$。称这一归纳原则为经验风险最小化归纳原则，简称为ERM。

1.4 学习理论研究的四个问题

在现实中，训练集的样本总是有限的，甚至往往并不能很好的反应数据分布$F(z)$的真实情况，这时经验风险最小可能并不能够很好的近似期望风险最小。于是，在有限样本时就需要考虑以下问题：经验风险最小时期望风险是否最小？使经验风险最小的解的期望风险如何？当存在多个使得经验风险最小的解时，哪一个可以使得期望风险更（最）小？

进一步的，把问题一般化为：“一个基于ERM原则的学习过程一致的充要条件是什么？这个学习过程收敛的速度有多快？如何控制这个学习过程的收敛速度（推广能力）？怎样构造能够够控制推广能力的算法？”[4]这就是统计学习理论需要研究的四个问题，对这四个问题的回答构成了统计学习理论的部分：

(1)学习过程一致性理论。
(2)学习过程收敛速度的非渐进理论。
(3)控制学习过程的推广能力的理论。
(4)构造学习算法的理论。

1.5 学习过程的一致性

ERM原则的一致性

$Q(z,\alpha^{'})$是使(2)式的经验风险最小化的函数，若

$$R(\alpha^{'}) \xrightarrow[l\to\infty]{\text{P}} \inf R(\alpha)$$
$$R_{\text{emp}}(\alpha^{'}) \xrightarrow[l\to\infty]{\text{P}} \inf R(\alpha)$$

即两个序列收敛于同一个极限，则称ERM原则对函数集$Q(z,\alpha), \alpha \in \Lambda$和概率分布函数$F(z)$是一致的。$\xrightarrow[def]{abc}$

ERM原则一致性的充要条件

可以证明[5]，当$Q(z,\alpha), \alpha \in \Lambda$使得$A \leqslant R(\alpha) \leqslant B$时，ERM原则一致性的充要条件是一致单边收敛

$$\lim_{l\to\infty} P\left\{\sup_{\alpha\in\Lambda}\left(R(\alpha) - R_{\text{emp}}(\alpha)\right) > \varepsilon\right\} = 0, \forall\varepsilon > 0 \tag{3}$$

成立。即把一致性问题转化为了一致单边收敛的问题。

为讨论一致单边收敛的充要条件，先补充以下定义。

函数集的熵

对示性函数集$Q(z,\alpha), \alpha \in \Lambda$和样本$z_1,\cdots z_l$，定义$N^{\Lambda}(z_1,\cdots,z_l)$为示性函数集中函数对给定样本进行分类得到的分类数目，用它可以表示函数集在给定数据集上的多样性。

用随机熵表示函数集在给定数据上的多样性，定义为

$$H^{\Lambda}(z_1,\cdots,z_l) = \ln N^{\Lambda}(z_1,\cdots,z_l)$$

可以理解为用$Q(z,\alpha),\alpha\in\Lambda$在由样本$z_1,\cdots z_l$的所有可能分类情况构成的$l$维超立方体上可以得到的立方体的不同的定点数目。

于是函数集在规模为l的样本集上的多样性定义为

$$H^{\Lambda}(l)=E\ln N^{\Lambda}(z_1,\cdots,z_l) \tag{4}$$

称作VC熵。

对于损失函数集是有界实函数集的情况，可以令随机变量$N^{\Lambda}(\varepsilon;z_1,\cdots,z_l)$是$l$为超立方体中最小$\varepsilon-$网格的数目[5]。于是定义

$$H^{\Lambda}(\varepsilon;z_1,\cdots,z_l)=\ln N^{\Lambda}(\varepsilon;z_1,\cdots,z_l)$$

为随机$\varepsilon-$熵。

于是，完全类似的也可得到实函数损失函数集的VC熵。

一致单边收敛成立的充要条件

考虑一致双边收敛为

$$\lim_{l\to\infty}P\{\sup_{\alpha\in\Lambda}|R(\alpha)-R_{\mathrm{emp}}(\alpha)|>\varepsilon\}=0,\forall\varepsilon>0 \tag{5}$$

可以看做大数定律在泛函空间的推广。

显然一致双边收敛包含一致单边收敛，即一致双边收敛是一致单边收敛成立的充分条件。

又可以证明一致双边收敛成立的重要条件为[5]

$$\lim_{l\to\infty}\frac{H^{\Lambda}(\varepsilon;l)}{l}=0,\forall\varepsilon>0$$

即随着观测数目的增加，VC熵与观测数目的比值应该趋于零。

这样，$\lim\limits_{l\to\infty}\dfrac{H^{\Lambda}(\varepsilon;l)}{l}=0$就是一致单边收敛的一个充分条件。

再定义函数集$Q^*(z,\alpha^*),\alpha^*\in\Lambda^*$为对任意的$Q(z,\alpha)$存在一个$Q^*(z,\alpha^*)$使得

$$Q(z,\alpha)-Q^*(z,\alpha^*)\geqslant 0,\forall z,$$

$$\int(Q(z,\alpha)-Q^*(z,\alpha^*))\,dF(z)\leqslant\delta$$

即$Q^*(z,\alpha^*)$是不超过$Q(z,\alpha)$且与其非常接近的函数。

这样，可以得到[5]完全一致有界的损失函数集$Q(z,\alpha),\alpha\in\Lambda$经验风险一致单边收敛的充要条件为$\forall\delta,\eta,\varepsilon>0,\exists Q^*(z,\alpha^*),\alpha^*\in\Lambda^*$，且有

$$\lim_{l\to\infty}\frac{H^{\Lambda^*}(\varepsilon;l)}{l}<\eta \tag{6}$$

又有前述ERM原则一致性的充要条件为经验风险一致单边收敛，于是知(6)式即为ERM原则一致收敛的充要条件。

1.6 无限样本学习过程的收敛速度

度量上面(3)式的单边收敛速度即为学习过程的收敛速度。

对示性损失函数集，定义退火的VC熵为，

$$H^{\Lambda}_{\mathrm{ann}}(l)=\ln EN^{\Lambda}(z_1,\cdots,z_l) \tag{7}$$

由[5]可知，

$$\lim_{l\to\infty}\frac{H^{\Lambda}_{\mathrm{ann}}(l)}{l}=0$$

是学习过程收敛速度快的一个充分条件。

1.7 一致性和收敛速度不依赖于概率测度的情况

定义生长函数为

$$G^{\Lambda}(l)=\ln\sup_{z_1,\cdots,z_l}N^{\Lambda}(z_1,\cdots,z_l)$$

仍由[5]知，

$$\lim_{l\to\infty}\frac{G^{\Lambda}(l)}{l}=0$$

是ERM学习过程对任何概率测度具有一致性的重要条件，且是ERM学习过程收敛速度快的充分条件。

最后，还可以得到

$$H^{\Lambda}(l)\leqslant H^{\Lambda}_{\mathrm{ann}}(l)\leqslant G^{\Lambda}(l) \tag{8}$$

1.8 有限样本下学习过程的收敛速度

上面的讨论都是基于样本趋于无穷时的极限情况，以下的给出有限样本时学习过程的收敛速度的界。又因为这里下界并不重要，所以只给出上界。

示性损失函数收敛速度依赖于分布的上界

由[4]知下面的不等式成立。

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}\leqslant 4\exp\left\{\left(\frac{H^{\Lambda}_{\mathrm{ann}}(2l)}{l}-\varepsilon^2\right)l\right\} \tag{9}$$

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\}\leqslant 4\exp\left\{\left(\frac{H^{\Lambda}_{\mathrm{ann}}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right\} \tag{10}$$

其中，公式(10)成立的条件是$\lim\limits_{l\to\infty}\dfrac{H^{\Lambda}(l)}{l}=0$。

实损失函数集上的推广

用实损失函数的示性函数来得到相应的推广。

定义实损失函数集$Q(z,\alpha),\alpha\in\Lambda$的示性函数集为$I(z,\alpha,\beta)\alpha\in\Lambda,\beta\in\mathcal{B}$为

$$I(z,\alpha,\beta)=\begin{cases}1 & \text{if } Q(z,\alpha)\geqslant\beta\\ 0 & \text{else.}\end{cases},$$

(1)当$A\leqslant Q(z,\alpha)\leqslant B,\alpha\in\Lambda$是完全有界实函数的集合时，有

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\}\leqslant 4\exp\left\{\left(\frac{H^{\Lambda,\mathcal{B}}_{\mathrm{ann}}(2l)}{l}-\varepsilon^2\right)l\right\} \tag{11}$$

(2)当$0 \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$是完全有界非负实函数的集合时，有

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\} \leqslant 4\exp\left\{\left(\frac{H_{\text{ann}}^{\Lambda,\mathcal{B}}(2l)}{l}-\frac{\varepsilon^2}{4}\right)l\right\} \quad (12)$$

而当$Q(z,\alpha)$为示性函数时，(11)和(12)分别退化为(9)和(10)。

收敛速度与分布无关的上界

由(8)可知，类似(11)和(12)的不等式对生长函数也成立，得到的不等式将不依赖于分布函数$F(z)$。

(1)当$A \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$是完全有界实函数的集合时，有

$$P\left\{\sup_{\alpha\in\Lambda}\left|\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)\right|>\varepsilon\right\} \leqslant 4\exp\left\{\left(\frac{G^{\Lambda,\mathcal{B}}(2l)}{l}-\frac{\varepsilon^2}{(B-A)^2}\right)l\right\} \quad (13)$$

(2)当$0 \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$是完全有界非负实函数的集合时，有

$$P\left\{\sup_{\alpha\in\Lambda}\frac{\int Q(z,\alpha)\,dF(z)-\frac{1}{l}\sum_{i=1}^{l}Q(z_i,\alpha)}{\sqrt{\int Q(z,\alpha)\,dF(z)}}>\varepsilon\right\} \leqslant 4\exp\left\{\left(\frac{G^{\Lambda,\mathcal{B}}(2l)}{l}-\frac{\varepsilon^2}{4B}\right)l\right\} \quad (14)$$

1.9 ERM学习机器推广能力的概念性界

为探讨采用ERM原则的学习机器的推广能力，需要讨论以下两个问题：

(1)得到最小经验风险$R_{\text{emp}}(\alpha')$的函数$Q(z,\alpha')$的期望风险$R(\alpha')$是多少？

(2)对任意给定的函数集，$R_{\text{emp}}(\alpha')$与真正的期望风险最小$\inf\limits_{\alpha\in\Lambda} R(\alpha)$的接近程度怎样？

以下仍将讨论限制在完全有界函数集和完全有界非负函数集上，并记

$$\mathscr{E}=4\frac{G^{\Lambda,\mathcal{B}}(2l)-\ln(\frac{\eta}{4})}{l}$$

完全有界实函数集

设$A \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$，则

(1)下面不等式至少以$1-\eta$的概率同时对$Q(z,\alpha), \alpha \in \Lambda$的所有函数成立：

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha)+\frac{B-A}{2}\sqrt{\mathscr{E}} \quad (15)$$

$$R_{\text{emp}}(\alpha)-\frac{B-A}{2}\sqrt{\mathscr{E}} \leqslant R(\alpha)$$

并且有这个不等式与(13)等价。

(2)下面不等式至少以$1-2\eta$的概率对使经验风险最小的函数$Q(z,\alpha')$成立：

$$R(\alpha')-\inf_{\alpha\in\Lambda}R(\alpha) \leqslant (B-A)\sqrt{\frac{-\ln\eta}{2l}}+\frac{B-A}{2}\sqrt{\mathscr{E}} \quad (16)$$

完全有界非负实函数集

设$0 \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$，则

(1)下面不等式至少以$1-\eta$的概率同时对$Q(z,\alpha), \alpha \in \Lambda$的所有函数成立：

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha)+\frac{B\mathscr{E}}{2}\left[1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha)}{B\mathscr{E}}}\right] \quad (17)$$

并且有这个不等式与(14)等价。

(2)下面不等式至少以$1-2\eta$的概率对使经验风险最小的函数$Q(z,\alpha')$成立：

$$R(\alpha')-\inf_{\alpha\in\Lambda}R(\alpha) \leqslant B\sqrt{\frac{-\ln\eta}{2l}}+\frac{B\mathscr{E}}{2}\left[1+\sqrt{1+\frac{4}{\mathscr{E}}}\right] \quad (18)$$

上面不等式(15)、(17)限定了集合$Q(z,\alpha), \alpha \in \Lambda$中的所有函数的风险；(16)、(18)则限制了用ERM原则得到的风险与最小可能的风险之间的接近程度。

另外还注意到，若$\mathscr{E} < 1$，则从相对偏差一致收敛速度得到的界(17)对较小的经验风险值，其置信范围的量级为$\mathscr{E}$，比从一致收敛速度得到的界(15)的量级$\sqrt{\mathscr{E}}$要好得多。

1.10 ERM学习机器推广能力的构造性界

由于退火熵和生长函数都是概念性的，所以前面得到的界无法直接用来构造算法。

所以引入函数集$Q(z,\alpha), \alpha \in \Lambda$的VC维的概念来找到构造性的界。

生长函数的结构

VC维与生长函数之间有重要联系。对任何生长函数，它或者满足等式

$$G^{\Lambda}(l)=l\ln 2$$

或者满足不等式约束

$$G^{\Lambda}(l) \leqslant h\left(\ln\frac{l}{h}+1\right)$$

其中，h是整数，并使得

$$G^{\Lambda}(l)\begin{cases}=l\ln 2 & \text{if } l<h\\ \leqslant h\left[1+\ln\frac{l}{h}\right] & \text{if } l\geqslant h\end{cases}$$

也就是说，生长函数或者是线性的，或者上界是一个对数函数，如图1[2]。

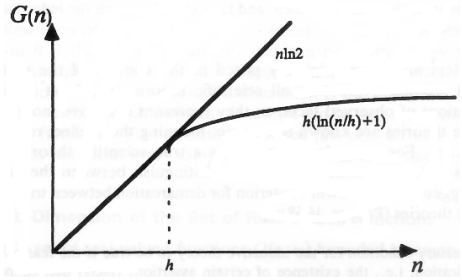

Figure 1: $G^{\Lambda}(l)$的结构

(1)当示性函数集的生长函数线性时，称这个函数集的VC维无穷大。

(2)当示性函数集的生长函数以参数为h的对数函数为界，则称示性函数集的VC维有限且等于h。因为此时，

$$\frac{H^{\Lambda}(l)}{l} \leqslant \frac{H^{\Lambda}_{\text{ann}}(l)}{l} \leqslant \frac{G^{\Lambda}(l)}{l} \leqslant \frac{h\left(\ln\frac{l}{h}+1\right)}{l}, (l>h)$$

所以学习机器所实现的示性函数集的VC维就是ERM方法一致性的一个充分条件，并且不依赖于概率测度。同时，也可看出此时也有快的收敛速度。另外，VC维有限也是ERM学习机器具有与分布无关的一致性的充要条件[5]。

函数集的VC维

(1)示性函数的VC维

一个示性函数集$Q(z,\alpha),\alpha \in \Lambda$的VC维是能够被集合中的函数以所有可能的$2^h$种方式分成两类的向量$z_1,\cdots,z_h$的最大数目$h$。如果对任意的$n$，总存在一个$n$个向量的集合可以被该函数集大打散，则函数集的VC维是无穷大[4]。

(2)实函数集的VC维

设$A \leqslant Q(z,\alpha) \leqslant B, \alpha \in \Lambda$是一个以常数$A$和$B$为界的是函数集（$A$、$B$可以为$\infty$），$I(z,\alpha,\beta),\alpha \in \Lambda, \beta \in \mathcal{B}$ 是其示性函数集。则原实函数集的VC维就定义为它对应的示性函数集的VC维

图2是函数集可以打散向量的情况；图3是线性函数集无法打散向量的情况[1]。

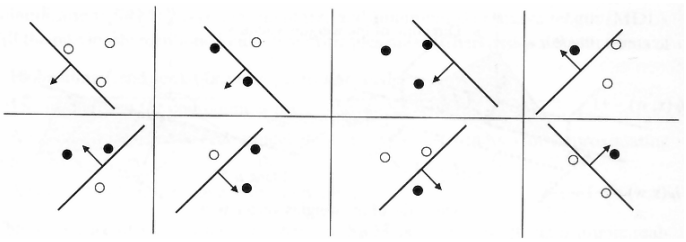

Figure 2: 线性函数集可以打散二维向量的情况

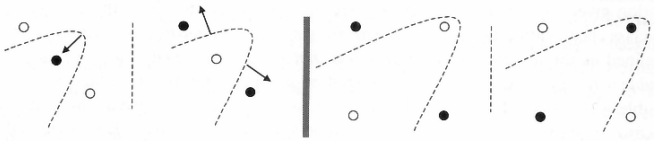

Figure 3: 线性函数集无法打散二维向量的情况

注意，从图2、图3中观察到的线性函数集的VC维等于其自由参数的规律对一般情况并不成立。虽然VC维是构造性的，但目前仍未有一般算法计算或估算任意函数集的VC维。

构造性的与分布无关的界

对VC维有限的函数集，$G^{\Lambda} \leqslant h\left(\ln\frac{l}{h}+1\right)$对$l>h$的情况成立，用这个关系代替生长函数就可得到基于VC维的更便于计算的边界。

使用

$$\mathscr{E} = 4\frac{h\left(\ln\frac{2l}{h}+1\right)-\ln\left(\frac{\eta}{4}\right)}{l}$$

来表达VC维无限的不等式。

另外当损失函数集$Q(z,\alpha),\alpha\in\Lambda$有限时（设其元素数目为$N$），可以，

$$\mathscr{E} = 2\frac{\ln N - \ln\eta}{l}$$

来表达VC维有限的不等式。

(1)对于完全有界的函数集

设$A \leqslant Q(z,\alpha) \leqslant B, \alpha\in\Lambda$，则

(a)下面不等式至少以$1-\eta$的概率同时对$Q(z,\alpha),\alpha\in\Lambda$的所有函数成立：

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \frac{B-A}{2}\sqrt{\mathscr{E}} \qquad (19)$$
$$R(\alpha) \geqslant R_{\text{emp}}(\alpha) + \frac{B-A}{2}\sqrt{\mathscr{E}}$$

(b)下面不等式至少以$1-2\eta$的概率对使经验风险最小的函数$Q(z,\alpha')$成立：

$$R(\alpha') - \inf_{\alpha\in\Lambda} R(\alpha) \leqslant (B-A)\sqrt{\frac{-\ln\eta}{2l}} + \frac{B-A}{2}\sqrt{\mathscr{E}} \qquad (20)$$

(2)完全有界非负实函数集设$0 \leqslant Q(z,\alpha) \leqslant B, \alpha\in\Lambda$，则

(a)下面不等式至少以$1-\eta$的概率同时对$Q(z,\alpha),\alpha\in\Lambda$的所有函数成立：

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \frac{B\mathscr{E}}{2}\left[1+\sqrt{1+\frac{4R_{\text{emp}}(\alpha)}{B\mathscr{E}}}\right] \qquad (21)$$

(b)下面不等式至少以$1-2\eta$的概率对使经验风险最小的函数$Q(z,\alpha')$成立：

$$R(\alpha') - \inf_{\alpha\in\Lambda} R(\alpha) \leqslant B\sqrt{\frac{-\ln\eta}{2l}} + \frac{B\mathscr{E}}{2}\left[1+\sqrt{1+\frac{4}{\mathscr{E}}}\right] \qquad (22)$$

1.11 控制学习过程的推广能力

函数集的结构

设函数的集合S具有由一系列函数子集$S_k = \{Q(z,\alpha), \alpha \in \Lambda_k\}$嵌套构成的结构，满足

$$S_1 \subset S_2 \subset \cdots \subset S_k \cdots \subset S$$

其中，结构的元素满足:

(1) 每个函数集的VC维都是有限的，且$h_1 < h_2 < \cdots < S_k < \cdots$

(2) 结构中的任何元素或者包含一个完全有界的正函数集

$$0 < Q(z,\alpha) < B_k, \alpha < \Lambda_k$$

或者包含对一定的(p,τ_k)满足下列不等式的函数的集合，

$$\sup_{\alpha\in\Lambda_k} \frac{\left(\int Q^p(z,\alpha)\,dF(z)\right)^{\frac{1}{p}}}{\int Q(z,\alpha)\,dF(z)} \leqslant \tau_k, p > 2$$

于是把这种结构称作容许结构，其中的函数子集称作结构的元素。

图4为嵌套函数子集确定的函数集的结构[4]。

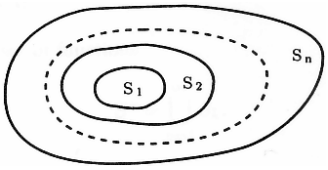

Figure 4: 嵌套函数子集确定的函数集的结构

结构风险最小化

对于前述式

$$\mathscr{E} = 4\frac{h\left(\ln\frac{2l}{h}+1\right) - \ln\left(\frac{\eta}{4}\right)}{l}$$

可以看出，若$\frac{l}{h}$较大，则$\mathscr{E}$较小，经验风险接近期望风险，ERM准则合理；若$\frac{l}{h}$较小，则小的经验风险并不能保证小的期望风险合理。

把前面的不等式(19)、(20)简写为以下形式，

$$R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \Phi(\frac{l}{h}) \tag{23}$$

即期望风险的上界是经验风险和$\Phi(\frac{l}{h})$的和。于是有限样本的情况下，使期望风险最小化，必须最小化等式右端的两项。其中，经验风险取决于函数集中特定的函数，即由训练方法决定；$\Phi(\frac{l}{h})$是置信范围，由函数集的VC维决定，即由学习机器的设计决定。

这时，求$\min R(\alpha) \leqslant R_{\text{emp}}(\alpha) + \Phi(\frac{l}{h})$就称作结构风险最小原则，记做SRM。

图5在函数集结构上解释了结构风险最小化的原则[3]。可以看出，随着函数子集序号的增加，经验风险减小，但同时置信范围增加。于是经验风险的上界是在结构的某个适当的元素上取得的，是在数据逼近的精度和逼近函数复杂性间的折中。

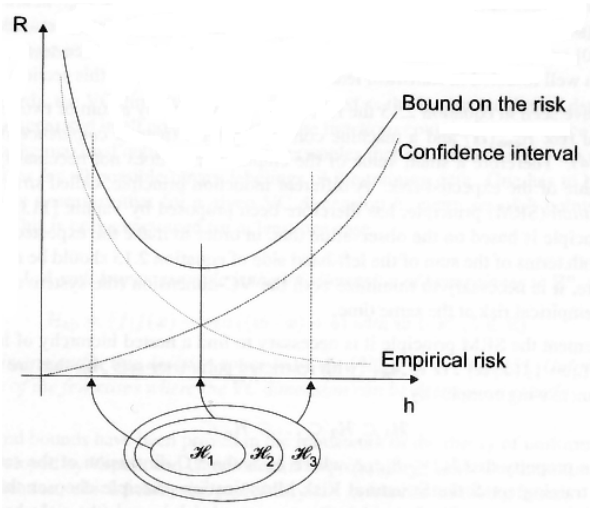

Figure 5: 结构风险最小化

结构风险最小化就是根据函数集的性质而将它划分为一系列的嵌套子集；而学习问题就可重新认识为根据推广能力选择最好的函数子集，并根据经验风险最小在函数子集中选择最好的函数。

2 用于模式识别的支持向量机

实现SRM可以通过保持置信范围固定（选择一定的学习机器），再最小化经验风险的方式得到；也可以通过保持经验风险一定，再最小化置信范围的思路得到。

SVM就是使用后一种思路得到的。

下面的推导和构造将限制在模式识别问题上，但对于回归估计、密度估计等其他机器学习问题也有类似的结果。

对于模式识别问题，定义判别函数为，

$$g(x) = w^T x + w_0 \begin{cases} g(x) > 0 & \Rightarrow x \in \omega_1, \\ g(x) < 0 & \Rightarrow x \in \omega_2. \end{cases}$$

2.1 最优分类超平面

假定训练数据$(x_1,y_1),\cdots,(x_l,y_l), x \in R^n, y \in \{+1,-1\}$可以被超平面$(w\cdot x)-b=0$分开。

若向量集被无错误的分开，即保证$R_{\text{emp}}(\alpha)=0$，且距离超平面最近的向量与超平面间的距离（称作间隔）最大，则说这个向量集合被最优超平面分开。

决策函数为$f(x)=\text{sgn}\{w\cdot x-b\}$，把距离分类面最近的决策函数值归一化，即强制

$$(w\cdot x)-b \begin{cases} \geqslant 1 & \text{if } y_i = 1, \\ \leqslant 1 & \text{if } y_i = -1. \end{cases}$$

也就是，

$$y_i[(w\cdot x)-b] \geqslant 1, i = 1,\cdot,l \tag{24}$$

于是，最优超平面就是满足式(24)，并使得

$$\min_{w,b}\Phi(w)=\|w\|^2 \tag{25}$$

成立的超平面。

2.2 构造最优超平面

采用Lagrange方法求解(24)、(25)式描述的最优分类超平面的问题，

$$L(w,b,\alpha)=\frac{1}{2}(w\cdot w)-\sum_{i=1}^{l}\alpha_i\{[(x_i\cdot w)-b]y_i-1\}$$

求解$L(w,b,\alpha)$的鞍点$\min_{w,b}\max_{\alpha}L(w,b,\alpha)$即可得到最优超平面的解。其中解满足以下条件

$$\frac{\partial L(w_0,b_0,\alpha^0)}{\partial b}=0$$

$$\frac{\partial L(w_0,b_0,\alpha^0)}{\partial w}=0$$

分别解得

$$\sum_{i=1}^{l}\alpha_i^0y_i=0 \tag{26}$$

$$w_0=\sum_{i=1}^{l}x_i\alpha_i^0y_i \tag{27}$$

(26)、(27)即为系数和最优超平面必须满足的方程组。

又由Kühn-Tacker条件，最优超平面得充要条件是

$$\alpha_i^0\{[(x_i\cdot w)-b_0]y_i-1\}=0,i=1,\cdots,l \tag{28}$$

考虑到(24)式及系数非负，所以只有满足

$$[(x_i\cdot w)-b_0]y_i=1 \tag{29}$$

的样本对应的系数才可能非零。

称满足(29)式的样本为支持向量，记所有支持向量对应的标号集合为SV，即有

$$w_0=\sum_{i\in SV}x_i\alpha_i^0y_i \tag{30}$$

最后将(30)式代回Lagrange函数，得到超平面问题的对偶问题，

$$\begin{aligned}\max_{\alpha}W(\alpha)=&\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=0}^{l}\alpha_i\alpha_jy_iy_j(x_i\cdot x_j),\\ &\alpha_i\geqslant 0,i=1,\cdots,l,\\ &\sum_{i=1}^{l}\alpha_iy_i=0\end{aligned} \tag{31}$$

为一个简单的非线性优化问题，求解得到系数向量$\alpha_0=(\alpha_1^0,\cdots,\alpha_l^0)$。

于是得到

$$|w_0|^2=\sum_{i=1}^{l}\alpha_i^0$$

在用某一样本待定系数可解得b_0，

$$f(x)=\text{sgn}\left[\sum_{i\in SV}y_i\alpha_i^0(x_i\cdot x_i)-b_0\right]$$

2.3 不可分情况下构造最优超平面

当数据为线性不可分时，使用最优超平面进行分类必然引入分类错误，即$R_{\text{emp}}(\alpha)\neq 0$，引入反应训练样本错分情况的和函数

$$F_\delta(\xi)=\sum_{i=1}^{l}\xi_i^\delta,\delta>0 \tag{32}$$

其中$\xi_i\geqslant 0$为非负松弛随机变量。修改(24)式为

$$y_i[(w\cdot x)-b]\geqslant 1-\xi_i,i=1,\cdot,l \tag{33}$$

$\Delta-$间隔分类超平面

引入条件$(w\cdot w)\leqslant\Delta^{-2}$。

在式(32)和新引入的条件下最小化泛函$F_\delta(\xi)$，并考虑$\delta=1$，称作$\Delta-$间隔分类超平面。

可以证明，$\Delta-$间隔分类超平面是由向量

$$w_0=\frac{1}{C^*}\sum_{i=1}^{l}x_i\alpha_i^0y_i \tag{34}$$

其中$\alpha_i,i=1,\cdots,l$是下面凸优化问题的解，

$$\begin{aligned}\max_{\alpha}W(\alpha,C^*)=&\sum_{i=1}^{l}\alpha_i-\frac{1}{2C^*}\sum_{i,j=0}^{l}\alpha_i\alpha_jy_iy_j(x_i\cdot x_j)-\frac{C^*}{2\delta^2},\\ &\alpha_i\geqslant 0,i=1,\cdots,l,\\ &\sum_{i=1}^{l}\alpha_iy_i=0,C^*\geqslant 0\end{aligned}$$

软间隔分类超平面

为简化计算，引入软间隔分类超平面。软间隔超平面是在约束条件(33)式下由使泛函

$$\Phi(w,\xi)=\frac{1}{2}(w\cdot w)+C\left(\sum_{i=1}^{l}\xi_i\right)$$

最小化的向量w决定的（C给定）。

$$\begin{aligned}\max_{\alpha}W(\alpha)=&\sum_{i=1}^{l}\alpha_i-\frac{1}{2}\sum_{i,j=0}^{l}\alpha_i\alpha_jy_iy_j(x_i\cdot x_j),\\ &\sum_{i=1}^{l}\alpha_iy_i=0,\\ &0\leqslant\alpha_i\leqslant C,i=1,\cdots,l\end{aligned}$$

类似线性可分时，有

$$\alpha_i^0\{[(x_i\cdot w)-b_0]y_i-1+\xi_i\}=0,i=1,\cdots,l$$

2.4 非线性分类问题

对非线性问题，可以考虑通过适当的非线性变换转化为线性判别函数。即通过事先选定的非线性映射将输入向量x映射到一个高维特征空间Z，从而在这个空间中构造最优分类超平面。

但是，这样实现这样转化的非线性映射可能非常复杂；同时，变换后的空间维数可能非常高。可以知道，在高维空间中，最优超平面仍然有好的推广型，并可以被找到。下面处理高维空间带来的维度灾难问题。

内积的回旋

在特征空间Z中构造最优分类超平面，并不需要以显示形式来考虑特征空间，而只需要能够计算支持向量与特征空间中向量的内积。

在Hilbert空间中内积的一个一般表达为

$$(z_i \cdot z) = K(x, x_i)$$

其中，z是输入向量x在高维特征空间中的像。

$K(x, x_i)$可以使满足下面Mercer准则的任意对称函数。

Mercer准则：要保证L_2下的对称函数$K(u, v)$能以正的系数$\alpha_k > 0$展开为

$$K(u, v) = \sum_{k=1}^{\infty} \alpha_k \phi_k(u) \phi_k(v)$$

的充分必要条件是，对使得$\int g^2(u)\, du$的所有$g \neq 0$，条件

$$\iint K(u, v) g(u) g(v)\, du\, dv > 0$$

成立。

另外Z空间中的内积可以转化为原X空间中的核函数，

$$(z_i, z_j) = (\phi(x_i), \phi(x_j)) \to k(x_i, x_j)$$

于是变换后的特征空间中的最优分类面只需要在原空间中求解，而并不需要实际实现变换。

2.5 构造SVM

利用内积回旋构造在输入空间中的非线性决策函数

$$f(x) = \operatorname{sgn}\left(\sum_{i \in SV} y_i \alpha_i K(x_i, x_j) \right) \tag{35}$$

它等价于在高维特征空间$\phi_1(x), \cdots, \phi_N(x)$中的线性决策函数。

系数α_i可以通过类似

$$\max_{\alpha} W(\alpha) = \sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j y_i y_j K(x_i . x_j), \tag{36}$$

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \alpha_i \geqslant 0, i = 1, 2, \cdots, l. \tag{37}$$

来求得。其中，用内积的回旋$K(x_i, x_j)$来取代内积$(x_i \cdot x_j)$。

这样构造的(35)式类型的决策函数的学习机器叫做支持向量机，记做SVM。

即SVM是通过事先选定的非线性映射将输入映射x映射到一个高维特征空间Z，并在这个空间中构造最优分类超平面，使得经验风险为零的同时置信区间也最小，从而期望风险最小。

致谢

这篇总结中的很多内容参考了张学工老师《统计学习理论导论》课程的课件。

参考文献

[1] V. Kecman. Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models. 2001, The MIT Press.

[2] Filip M. Mulier, Vladimir S. Cherkassky. Learning from Data 1998, Wiley-Interscience.

[3] Rychetsky, Matthias. Algorithms and Architectures for Machine Learning based on Regularized Neural Networks and Support Vector Approaches 2001, Shaker Verlag.

[4] V. Vapnik. The Nature of Statistical Learning Theory. 1995, Springer.

[5] V. Vapnik. Statistical Learning Theory. 1998, John Wiley and Sons.