式(7-2)中的数学期望。定义经验风险(empirical risk)为在训练样本上损失函数的平均

$$R_{\text{emp}}(\alpha)=\frac{1}{l}\sum_{i=1}^{l}L(y_i,f(\boldsymbol{x}_i,\alpha)) \tag{7-6}$$

历史上大部分机器学习方法实际上都是用最小化经验风险来替代最小化期望风险的目标。

例如,感知器的学习目标是:

$$\min \quad J_P(\alpha)=\sum^{y_j\in Y^k}(-\alpha^{\mathrm{T}}y_j) \tag{7-7}$$

线性回归的学习目标是:

$$\min \quad E(\boldsymbol{w})=\frac{1}{N}\sum_{j=1}^{l}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j-y_j)^2 \tag{7-8}$$

罗杰斯特回归的学习目标是:

$$\min \quad E(\boldsymbol{w})=\frac{1}{N}\sum_{j=1}^{l}\ln(1+\mathrm{e}^{-y_j\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_j}) \tag{7-9}$$

多层感知器神经网络的学习目标是:

$$\min \quad E(\boldsymbol{w})=\frac{1}{2}\sum_{j=1}^{l}(y-\hat{y}_{\text{MLP}})^2 \tag{7-10}$$

统计学习理论把这种以在训练样本上最小化错误或风险的策略称为经验风险最小化(empirical risk minimization)原则,简称 ERM 原则。20 世纪 50 年代,以感知器为代表的机器学习方法取得了巨大的发展,当时的方法大都采用了 ERM 原则,研究的重点放在如何设计合适的候选函数集和如何设计有效的算法实现经验风险最小化。

Vapnik 把这些研究者称作应用分析学派,并指出,经验风险最小化原则并不是毋庸置疑的,我们不应该只关注如何设计经验风险最小化的算法,而应该研究经验风险最小化这个原则是否合理,应该寻找最优化学习机器推广能力的新原则。他把这样的研究称作理论分析学派。

实际上,以期望风险最小为目标来分析经验风险最小化原则,会发现这其实是想当然的做法,合理性并没有充分的理论保证。

首先,$R_{\text{emp}}(\alpha)$和$R(\alpha)$都是$f(\boldsymbol{x},\alpha)$的泛函,概率论中的大数定律只说明了随机变量的均值在样本倾向于无穷大时会收敛于其期望,但这个定律对泛函是否仍然成立?这一点在当时并没有数学上的结论。

其次,即使我们类比随机变量的情况认为$R_{\text{emp}}(\alpha)$在样本倾向于无穷大时会充分接近$R(\alpha)$,这并不是我们需要的结果。我们需要的是$R_{\text{emp}}(\alpha)$在l个样本上取得极小值的解α_l收敛于使$R(\alpha)$取得最小值的解α^*。通常,两个函数充分接近并不能保证它们的极值点也充分接近,这里实际上提出来一个更基本的问题:所谓用$R_{\text{emp}}(\alpha)$近似$R(\alpha)$或当样本趋向无穷多时$R_{\text{emp}}(\alpha)$收敛于$R(\alpha)$,应该用什么来作为两个函数接近程度的度量?

再次,即使我们有办法证明或通过一定条件保证在样本趋向于无穷多时,使经验风险最小的解也使期望风险最小,在实际问题中需要多少样本才能达到接近无穷多的效果?如果样本远非无穷多而是非常有限,经验风险最小化是否还可行?得到的解是否还有推广能力?

在机器学习领域多数研究者都将注意力集中到如何更好地求解最小经验风险问题上