第5章 机器学习的性能与评估

CHAPTER 5

正如第1章中“没有免费午餐定理”所述，没有一种机器学习模型是对所有问题普遍适用的，正因为如此，机器学习领域中出现众多的不同模型。第3章和第4章介绍了基本的回归和分类模型，在第6章之后仍将继续介绍多种常用的机器学习模型。在这样一个节点，本章偏离机器学习模型和学习算法的介绍，以一章的篇幅集中介绍机器学习中训练和性能评价的一些基本问题，这是一个很大的问题，本章的介绍仅仅是入门和概要性质的。

当用模型表示一类问题时，对模型性能的理想描述是期望风险，即从完整的统计意义上刻画模型相对于目标的偏差。但在机器学习领域，缺乏对目标完整的概率描述，因此无法获得期望风险，而是从有限数据中学习模型，评价准则也以经验风险代替期望风险。由于数据集的代表能力有限，以经验风险最优确定的模型对真实目标的总体表达能力如何，即泛化能力如何？这是一个非常关键的问题。

泛化性能好是一个机器学习模型可用的基本要求，因此必须对泛化性能进行评价。一种比较实际的评价泛化性能的方法是通过数据集进行测试，将数据集划分为训练集和测试集，用训练集学习模型，在测试集上近似估计其泛化性能；另一种评价方法是给出理论上的泛化界并研究泛化误差与数据集规模的关系，这是机器学习理论讨论的基本问题。遗憾的是，目前这两类方法之间仍存在鸿沟。利用机器学习理论，对于在要求的泛化误差下给出的样本规模并不能很精确地指导许多实际机器学习模型的训练，逾越这道鸿沟仍需努力。

本章由两部分组成，第1部分包括5.1节和5.2节，讨论如何利用实际数据集有效地训练和评价一个机器学习模型；第2部分由5.3节和5.4节组成，讨论机器学习理论中的一些基本概念和结论。

5.1 模型的训练、验证与测试

在1.3节介绍机器学习的基本元素时，为了讨论问题方便，将数据集分为训练集和测试集。本节结合实际机器学习系统的学习过程，对数据的划分和作用再做一些更深入的讨论。

在机器学习的许多模型中，存在一些称为超参数的量，超参数是不能直接通过训练过程确定得到的，例如多项式拟合的阶 M、KNN的参数 K 或正则项的控制参数 λ。可以通过学习理论或贝叶斯框架下的学习确定这些超参数，但目前实际中更常用的是通过验证过程确定。

在最简单的情况下，不需要确定超参数，数据集仍划分为训练集和测试集，或两个集合独立地产生自同一个数据生成分布，训练集训练模型，通过测试误差近似评价泛化性能。

更复杂的情况下可将数据集划分为 3 个集合：训练集、验证集(validation set)和测试集。若数据集数据量充分，可以直接按一定比例划分 3 个集合，例如训练集占 80%，验证集占 10%，测试集占 10%，各集合的比例可根据数据集的总量做适当调整，数据集划分如图 5.1.1(a)所示。在一般的学习过程中，在超参数的取值空间内，按一定方式(等间隔均匀取值或随机取值)取一个(或一组，复杂模型可能有多个超参数)超参数值，用这组确定的超参数，通过训练集训练模型，将训练得到的模型用于验证集，计算验证(集)误差。取不同的超参数，重复这个过程，最后确定效果最好的超参数及对应的模型。然后用测试集测试性能，计算测试误差，估计泛化性能。若测试性能达不到要求，还可能回到原点，选择不同的模型，重复以上过程，直到达到要求或在可能选择的模型中取得最好结果。在整个过程中测试集是不参与学习过程的，这样才能够可信地评估模型的泛化能力。

一些情况下，数据集规模较小，若固定地分成 3 个集合，则每个集合数据量小使训练过程和验证过程都缺乏可靠性。这种情况下可采用交叉验证(cross validation)方法。数据集仍划分为测试集和训练集，测试集留作最后的测试用。将训练集分为 K 折(K folds)，用于训练和验证，对于一组给出的超参数，做多轮训练，每次训练留出一折作为验证集，其余作为训练集，进行一次训练和验证，然后循环操作，过程如图 5.1.1(b)所示。做完一个循环，将每次验证集的误差做平均，作为验证误差。选择一组新的超参数重复该过程，直到全部需要实验的超参数取值完成后，比较所有超参数取值下的验证误差，确定超参数的值，其后再用全部训练集样本训练出模型。将以上学习过程确定的模型用于测试集计算测试误差，评价是否达到目标。

图 5.1.1　用于训练和测试的数据集划分

在数据集样本相当匮乏的情况下，以上交叉验证可取其极限情况，每轮只留一个样本作为验证集，称为留一验证(leave-one out cross validation，LOOCV)。

以上介绍了用数据集获得一个机器学习模型的基本方法。实践中可能还有各种灵活的组合方式。后续章节介绍的各种算法在实际中应用时，一般用上述某种方式完成学习和测试过程。

5.2 机器学习模型的性能评估

一个机器学习模型确定后,性能是否符合任务的需求,需要对其进行评估。一般来讲,对于较复杂的实际任务,性能评估方法可能与任务是相关的,因此性能评估方式有很多,本书作为以机器学习算法为主的基本教材,不对各种与任务相关的评估方法做过多讨论,本节只对几个最基本的性能评价方法进行概要介绍,并只讨论监督学习中的回归和分类的性能评估。

对一个机器学习模型 $h(\boldsymbol{x})$ 做准确的性能评估是困难的,实际中一般是在样本集(例如测试集)上对其进行性能评估,当样本集中样本数充分多且可充分表示实际样本分布时,用在样本集上的评估作为近似的泛化性能评估。本节为了叙述简单,假设样本集中的标注 y 均是标量。样本集表示为

$$\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N} \tag{5.2.1}$$

1. 回归的性能评估

对于回归问题,模型 $h(\boldsymbol{x})$ 的输出和样本标注均为实数,在样本集上评价其性能的常用方法是均方误差,即

$$E_{\text{mse}}(h)=\frac{1}{N}\sum_{i=1}^{N}[h(\boldsymbol{x}_i)-y_i]^2 \tag{5.2.2}$$

均方误差重点关注了大误差的影响,在一些应用中,也可能采用平均绝对误差或最大误差,分别表示为

$$E_{\text{abs}}(h)=\frac{1}{N}\sum_{i=1}^{N}|h(\boldsymbol{x}_i)-y_i| \tag{5.2.3}$$

$$E_{\infty}(h)=\max_{1\leqslant i\leqslant N}\{|h(\boldsymbol{x}_i)-y_i|\} \tag{5.2.4}$$

尽管存在一些其他评价函数,在回归问题中,均方误差评价使用最多。

2. 分类的性能评估

分类的性能评估比回归要复杂。这里讨论只有两类的情况,将我们更关心的一类称为正类,另一类称为负类。

评价分类的最基本准则是分类错误率和分类准确率,当对一个样本 $(\boldsymbol{x}_i,y_i)$ 做分类测试时,若分类器输出 $h(\boldsymbol{x}_i)$ 与样本标注相等,则分类正确,否则产生一个分类错误。对于式(5.2.1)的样本集中所有样本,统计对 $h(\boldsymbol{x})$ 能够进行正确分类和错误分类的比例,可得到在样本集上的分类错误率和分类准确率,分别表示为

$$E=\frac{1}{N}\sum_{i=1}^{N}I[h(\boldsymbol{x}_i)\neq y_i] \tag{5.2.5}$$

$$\text{Acc}=\frac{1}{N}\sum_{i=1}^{N}I[h(\boldsymbol{x}_i)=y_i]=1-E \tag{5.2.6}$$

其中,$I(x)$ 是示性函数,x 是逻辑变量。当 x 为真时,$I(x)=1$,否则 $I(x)=0$。当一个样本集中正类样本和负类样本分布均衡(大致数目相当),各种分类错误(将正类错分为负类或反之)代价相当时,分类准确率(或分类错误率)可较好地评价分类器的性能。但当式(5.2.1)所示的样本集中正类样本和负类样本分布很不均衡时,分类正确率不能客观反映分类器性

能，甚至引起误导。例如，在一个检测某癌症的数据集中有 10 000 个样本，正类样本（患癌症）的数目只有 300，其余的均为负类样本（非患者），对于这样的样本集，若一个分类器简单地将所有样本分类为负类，则分类准确率仍为 0.97，这个指标相当好，但对本任务该分类器毫无用处。

为了进一步讨论怎样构造更合理的评价方法，对于一个分类器 $h(\boldsymbol{x})$，将式(5.2.1)的样本集分为 4 类：①真正类，即样本为正样，分类器将其分为正类；②真负类，即样本是负样，分类器将其分类为负类；③假负类，即样本为正样，分类器将其分为负类；④假正类，即样本是负样，分类器将其分类为正类。样本集中各类的数目如表 5.2.1 所示。

表 5.2.1　样本的类型

标注的真实类型	分类器返回的类型	
	正类	负类
正类	N_{TP}	N_{FN}
负类	N_{FP}	N_{TN}

用表 5.2.1 的符号，样本总数 $N=N_{\mathrm{FP}}+N_{\mathrm{FN}}+N_{\mathrm{TP}}+N_{\mathrm{TN}}$，可重写分类错误率和分类准确率为

$$E=\frac{N_{\mathrm{FP}}+N_{\mathrm{FN}}}{N_{\mathrm{FP}}+N_{\mathrm{FN}}+N_{\mathrm{TP}}+N_{\mathrm{TN}}}$$

$$\mathrm{Acc}=\frac{N_{\mathrm{TP}}+N_{\mathrm{TN}}}{N_{\mathrm{FP}}+N_{\mathrm{FN}}+N_{\mathrm{TP}}+N_{\mathrm{TN}}}$$

对于前述癌症的例子，若将所有样本均分类为负类，则 $N_{\mathrm{FP}}=N_{\mathrm{TP}}=0$，$N_{\mathrm{FN}}=300$，$N_{\mathrm{TN}}=9700$，$E=0.03$，$\mathrm{Acc}=0.97$。在这种情况下，分类错误率和分类准确率几乎无法告诉我们分类器的实际效用。下面定义两个更有针对性的性能评价：精度(precision)和查全率(recall)。

精度定义为真正类 N_{TP} 与被分类器识别为正类的所有样本 $N_{\mathrm{TP}}+N_{\mathrm{FP}}$ 的比例，即

$$\mathrm{Pr}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FP}}} \tag{5.2.7}$$

查全率定义为正类样本被分类器正确识别为正类的概率，即真正类数目 N_{TP} 与正类样本总数 $N_{\mathrm{TP}}+N_{\mathrm{FN}}$ 之比

$$\mathrm{Re}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{5.2.8}$$

通过一个例子说明两个参数的意义。

例 5.2.1　针对癌症的例子，设一个分类器对样本集的分类情况如表 5.2.2 中“/”上侧所示。

表 5.2.2　样本的类型

标注的真实类型	分类器返回的类型	
	正类	负类
正类	210/260	90/40
负类	200/400	9500/9300

可计算出精度和查全率分别为 Pr≈0.51,Re=0.7,分类准确率 Acc=0.971。

若改变分类器参数,使输出为正类的概率提高,则可能同时负类样本被判定为正类的数目也增加了,如表 1.3.2 中"/"下侧的数据,则精度和查全率为 Pr≈0.39,Re≈0.87,分类准确率 Acc=0.956。

例 5.2.1 给出了两组数据,所代表的分类器均比将所有样本都分类为负类的"负分类器"有价值,但就分类准确率来讲,第 1 组数据没有改善,第 2 组数据反而下降了。对两组数据自身做比较可见,第 2 组("/"之下)数据将更多的正类样本(癌症)做了正确分类,但同时也将更多的负类样本判别为正类,因此查全率提高但精度降低。对于该任务,可以认为第 2 组数据表示的分类器更有用,它将更多的患者检查出来,以免耽误治疗,对于将负类样本判别为正类的错误分类,一般可通过后续检查予以改正。在这个应用任务中查全率的提高是更有意义的,但也有的任务希望有更高的精度。在实际中精度和查全率往往是矛盾的,哪个指标更重要往往取决于具体任务的需求。

可以将精度和查全率综合在一个公式中,即如下的 F_β

$$F_\beta=\frac{(\beta^2+1)\times \mathrm{Pr}\times \mathrm{Re}}{\beta^2\times \mathrm{Pr}+\mathrm{Re}} \tag{5.2.9}$$

对于 F_β,当 $\beta>1$ 时,查全率将得到更大权重;当 $0\leqslant\beta<1$ 时,精度得到更大权重。当取 $\beta=1$ 时,查全率和精度有等重的比例,得到一个简单的综合性能指标

$$F_1=\frac{2\times \mathrm{Pr}\times \mathrm{Re}}{\mathrm{Pr}+\mathrm{Re}} \tag{5.2.10}$$

当调整一个分类器的参数使其性能变化时,常利用 P-R 曲线或 ROC 曲线评价分类器在不同参数下的表现。

P-R 曲线是以精度为纵轴、以查全率为横轴的曲线,一般随着查全率提高,精度下降,图 5.2.1 为一个典型 P-R 曲线的示意图。

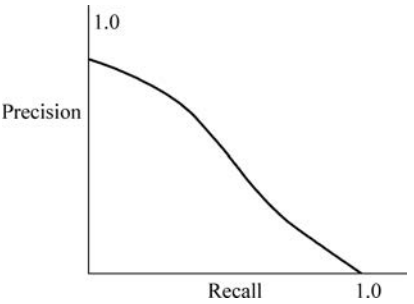

图 5.2.1 P-R 曲线示意

ROC 最初来自雷达检测技术,是接收机工作特性的缩写(receiver operating characteristic)。对于分类器,可定义正类样本分类准确率为

$$P_{\mathrm{Ac}}=\frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}}+N_{\mathrm{FN}}} \tag{5.2.11}$$

和负类样本错误率为

$$N_{\mathrm{e}}=\frac{N_{\mathrm{FP}}}{N_{\mathrm{TN}}+N_{\mathrm{FP}}} \tag{5.2.12}$$

当改变分类器参数时,P_{Ac} 和 N_{e} 都变化,以 P_{Ac} 为纵轴,以 N_{e} 为横轴,可画出一条曲线,则称为 ROC 曲线。一个理想的分类器,可取到 $P_{\mathrm{Ac}}=1$ 和 $N_{\mathrm{e}}=0$ 点,但现实中难以实现这样的分类器。实际分类器曲线示例如图 5.2.2 所示。

对于一个实际分类器,若控制参数将所有样本分类为负类,则对应$(N_{\mathrm{e}},P_{\mathrm{Ac}})=(0,0)$;若将所有样本分类为正类,则$(N_{\mathrm{e}},P_{\mathrm{Ac}})=(1,1)$,则曲线可通过坐标原点和(1,1)点。设有两个分类器对应的 ROC 曲线 c_1 和 c_2 画在同一图中,若 c_1 总是位于 c_2 之上,则分类器 1

性能总是优于分类器 2；若两条曲线有交叉，则在不同参数下两个分类器表现各有优劣。对于 ROC 曲线有交叉的不同分类器，一种比较其总体优劣的方法是采用 AUG(area under ROC curve)参数，一个分类器的 AUG 参数表示为其 ROC 曲线之下到坐标横轴之间的面积。

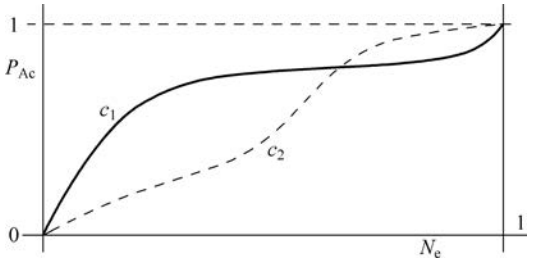

图 5.2.2　分类器 ROC 曲线示例

针对各类应用任务还有许多特别的参数，例如对于医学应用和金融应用，人们关注的性能可能非常不同，本书不再进一步讨论针对实际任务的性能评价。

5.3　机器学习模型的误差分解

本节以回归学习作为对象，讨论模型复杂度与误差的关系，即模型的偏和方差的折中问题。在 3.3 节的多项式基函数例子中(例 3.3.3)，已经看到对于固定规模的训练数据集，随着模型复杂度变化，训练误差和测试误差的变化关系。为了清楚起见，将该图重示于图 5.3.1(a)中。以多项式阶 M 表示模型的复杂度，可以看到，随着 M 增加，训练误差单调减少，但测试误差先减小后上升，也就是模型出现过拟合。用测试误差逼近所学模型的泛化误差，故模型的泛化误差不是随着模型的表达能力增强而减小，一般泛化误差与模型复杂度的关系是一个 U 形曲线。图 5.3.1(b)给出了训练误差与测试误差的一种更一般的示意图，横坐标表示模型的复杂性，类似地，训练误差单调减小，测试误差类似 U 形曲线。对于一个给定的学习任务，可选择对应测试误差 U 形曲线底端的模型复杂度。对于具体例子，图 5.3.1(a)中 M 取 3～7 这样一个较宽的范围，测试误差都处于 U 形的底端，都是可选的模型复杂度。

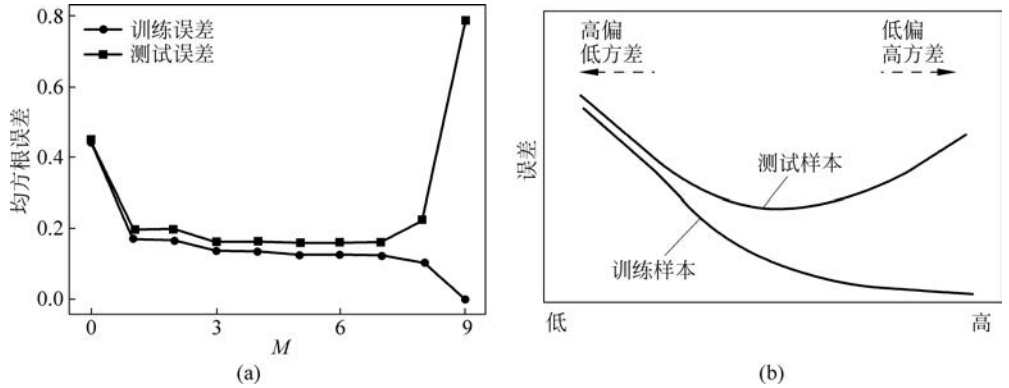

图 5.3.1　模型复杂度与误差的关系

接下来讨论更一般的泛化误差问题,以一般的回归模型作为讨论对象,不限于本书前面讨论的线性回归或基函数回归。假设一个数据集 $\boldsymbol{D}=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^N$ 是对一个联合概率密度函数 $p(\boldsymbol{x},y)$的采样,通过一个数据集学习得到的模型为 $\hat{y}(\boldsymbol{x})$,注意这里模型没有显式地依赖参数 $\boldsymbol{w}$,可表示更一般的模型。定义误差函数为

$$L[\hat{y}(\boldsymbol{x}),y]=[\hat{y}(\boldsymbol{x})-y]^2 \tag{5.3.1}$$

其中,y 表示回归模型要逼近的真实值。针对 $p(\boldsymbol{x},y)$可得到模型的误差期望为

$$\begin{aligned}E(L)&=\iint L[\hat{y}(\boldsymbol{x}),y]p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&=\iint[\hat{y}(\boldsymbol{x})-y]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\end{aligned} \tag{5.3.2}$$

其中,$E(L)$是针对模型 $\hat{y}(\boldsymbol{x})$的泛化误差。

对于回归问题,由第 2 章讨论的决策理论可知,若已知 $p(\boldsymbol{x},y)$,则最优的回归模型为

$$h(\boldsymbol{x})=\int yp(y\mid\boldsymbol{x})\mathrm{d}y=E(y\mid\boldsymbol{x}) \tag{5.3.3}$$

机器学习中,由于一般不知道准确的 $p(y|\boldsymbol{x})$,无法直接获得最优回归模型 $h(\boldsymbol{x})$,但是从原理上可以将 $h(\boldsymbol{x})$作为一个比较基准,得到如下误差分解

$$\begin{aligned}E(L)&=\iint[\hat{y}(\boldsymbol{x})-y]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&=\iint[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})+h(\boldsymbol{x})-y]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&=\iint[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y+\iint[h(\boldsymbol{x})-y]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y+\\&\quad 2\iint[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})][h(\boldsymbol{x})-y]p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\\&=\int[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})]^2p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\iint[E(y\mid\boldsymbol{x})-y]^2p(\boldsymbol{x},y)\mathrm{d}\boldsymbol{x}\mathrm{d}y\end{aligned} \tag{5.3.4}$$

式中,$\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})$与 y 无关,交叉项积分为 0,误差项由两项组成,其中第 2 项是随机变量的不可完全预测性的结果,这是一个固有量,与模型选择、学习过程均无关。式(5.3.4)最后一行的第 1 项是与模型和学习过程有关的,接下来仔细分析这一项。

假设只给出一个数据集 $\boldsymbol{D}=\{(\boldsymbol{x}_n,y_n)\}_{n=1}^N$,在这个数据集上学习回归模型,学习到的模型是与数据集相关的,为了清楚表示与数据集的相关性,将学习到的模型表示为 $\hat{y}(\boldsymbol{x};\boldsymbol{D})$,若可以获得若干数据集,将每个数据集学习的模型做平均,当数据集数目很大时,这个平均逼近一个期望,用符号 $E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]$表示。使用这个符号,将针对一个指定数据集时的误差项$[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})]^2$ 分解为

$$\begin{aligned}[\hat{y}(\boldsymbol{x};\boldsymbol{D})-h(\boldsymbol{x})]^2&=\{\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]+E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]-h(\boldsymbol{x})\}^2\\&=\{\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]\}^2+\{E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]-h(\boldsymbol{x})\}^2+\\&\quad 2\{\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]\}\{E_D[\hat{y}(\boldsymbol{x};\boldsymbol{D})]-h(\boldsymbol{x})\}\end{aligned} \tag{5.3.5}$$

将以上误差项对所有不同数据集进行平均,即取 $E_D(\cdot)$,则注意到交叉项为 0,故

$$E_D\{[\hat{y}(\boldsymbol{x};\boldsymbol{D})-h(\boldsymbol{x})]^2\}$$

$$=E_D\{[\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))]^2\}+E_D\{[E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))-h(\boldsymbol{x})]^2\}$$

$$=[E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))-h(\boldsymbol{x})]^2+E_D\{[\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))]^2\} \tag{5.3.6}$$

式(5.3.6)的第1项是偏差，从多个数据集分别学习得到模型的 $\hat{y}(\boldsymbol{x};\boldsymbol{D})$ 做平均的结果 $E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))$ 仍与最优模型 $h(\boldsymbol{x})$ 之间存在偏差；第2项是学习得到的模型的方差，即每个数据集训练得到的模型与模型期望之间的偏离程度，这个方差越大，不同数据集训练出的模型的起伏程度越大。由于式(5.3.6)对所有数据集 $\boldsymbol{D}$ 取了期望，其与具体数据集无关，可以看作 $[\hat{y}(\boldsymbol{x})-h(\boldsymbol{x})]^2$，带入式(5.3.4)得

$$\begin{aligned}E(L)=&\int[E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))-h(\boldsymbol{x})]^2p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\\&\int E_D\{[\hat{y}(\boldsymbol{x};\boldsymbol{D})-E_D(\hat{y}(\boldsymbol{x};\boldsymbol{D}))]^2\}p(\boldsymbol{x})\mathrm{d}\boldsymbol{x}+\\&\iint[E(y\mid\boldsymbol{x})-y]^2p(\boldsymbol{x},\boldsymbol{y})\mathrm{d}\boldsymbol{x}\mathrm{d}y\\=&(\text{偏})^2+\text{方差}+\text{固有误差}\end{aligned} \tag{5.3.7}$$

式(5.3.7)说明，对于一个给出的模型，其泛化误差由3部分组成：偏(实际是偏的平方，为了叙述简单，这里称为偏)、方差和固有误差。固有误差与模型、数据集和学习过程均无关，不需要进一步讨论。偏和方差确实与模型选择有关，一般地，若选择比较简单的模型，则偏比较大，这是由于模型的表示能力有限，即使从多次训练获得的模型平均也仍偏离最优模型 $h(\boldsymbol{x})$。若选择比较复杂的模型，可以使偏比较小，但方差变大。设模型是参数模型，则复杂模型具有更多的参数，在给定数据集规模的条件下，每个参数的平均有效样本数较小。第2章讨论过，方差与有效样本数成反比，因此方差变大。即模型简单，方差小，偏大；模型复杂，方差大，偏小。当模型取得比较合适时，既不算复杂也不算简单，即相对折中的模型，可能偏和方差都比较小，总误差最小。图5.3.2给出了偏、方差和泛化误差随模型复杂度的变化曲线。

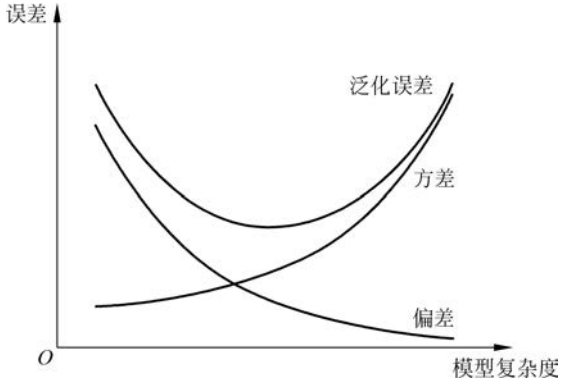

图5.3.2　模型的误差分解

例5.3.1　为了给出误差分解的直观解释，考虑一个简单的学习模型的例子。

设函数 $f(\boldsymbol{x})$ 无法直接观测，为了对该函数进行预测，通过采样获得数据集，采样过程为

$$y=f(\boldsymbol{x})+v \tag{5.3.8}$$

由于无法直接观测 $f(\boldsymbol{x})$，故采样样本存在误差 v，设 v 为零均值、方差为 σ_v^2 的高斯噪声。为了讨论问题简单，采样时各输入 $\boldsymbol{x}_i$ 是预先确定的。由采样数据构成I.I.D数据集

$\{\boldsymbol{x}_i, y_i\}_{i=1}^N$，由数据集训练一个模型，作为说明，这里采用 K 近邻回归算法，模型为

$$\hat{y} = \hat{f}(\boldsymbol{x}) = \frac{1}{K}\sum_{l=1}^{K} y_{(l)} \tag{5.3.9}$$

这里用 $y_{(l)}$ 表示对于给定的 $\boldsymbol{x}$，最近邻的 K 个训练集样本的标注值，用(l)表示最近邻样本的下标。

为了讨论误差分解，首先注意到，通过式(5.3.8)的观测，可得到最优模型为

$$h(\boldsymbol{x}) = E(y \mid \boldsymbol{x}) = f(\boldsymbol{x})$$

因此，固有误差为 σ_v^2。由于本例较为简单，直接使用式(5.3.4)的最后一行，则有

$$\begin{aligned} E(L) &= E\{(\hat{y} - h(\boldsymbol{x}))^2\} + \sigma_v^2 \\ &= E\left\{\left[\frac{1}{K}\left(\sum_{l=1}^{K} f(\boldsymbol{x}_{(l)}) + v_l\right) - f(\boldsymbol{x})\right]^2\right\} + \sigma_v^2 \\ &= E\left\{\left[\frac{1}{K}\sum_{l=1}^{K} f(\boldsymbol{x}_{(l)}) - f(\boldsymbol{x})\right]^2\right\} + E\left\{\left(\frac{1}{K}\sum_{l=1}^{K} v_l\right)^2\right\} + \sigma_v^2 \\ &= \left[\frac{1}{K}\sum_{l=1}^{K} f(\boldsymbol{x}_{(l)}) - f(\boldsymbol{x})\right]^2 + \frac{\sigma_v^2}{K} + \sigma_v^2 \end{aligned} \tag{5.3.10}$$

式(5.3.10)从倒数第二行到最后一行，是考虑做预测时，$\boldsymbol{x}$ 是一个给出的固定值。式(5.3.10)的最后一行分别是如式(5.3.7)所示的：偏差、方差和固有误差。对于 K 近邻方法，K 越大代表模型表达力越弱，$K=1$ 代表最强表达能力。显然，对于变化的内在函数 $f(\boldsymbol{x})$，K 越大 $\frac{1}{K}\sum_{l=1}^{K} f(\boldsymbol{x}_{(l)})$ 与 $f(\boldsymbol{x})$ 偏差越大(越多偏离 $\boldsymbol{x}$ 更远的函数值参与平均)，但方差 $\frac{\sigma_v^2}{K}$ 越小。$K=1$ 时则由最近邻的函数 $f(\boldsymbol{x}_{(l)})$ 逼近 $f(\boldsymbol{x})$，因此偏差最小，这时方差 $\frac{\sigma_v^2}{K}$ 最大。不管 K 取何值，最后一项的固有误差 σ_v^2 不变。

对于线性回归模型，也可导出闭式结果说明误差的分解，只是推导过程更加复杂一些。

对于机器学习的模型选择来讲，对于给定的问题和数据集，并不是选择越复杂的模型越好，要选择适中的模型。这是一个基本的原则，在实际中怎样使用这个原则，却不是一个简单的问题。从原理上讲，从最大似然准则过渡到完全的贝叶斯框架下，可以解决模型选择的问题，但对于一般非线性模型，贝叶斯框架下的求解要复杂得多。一个更实际的方法是通过正则化和交叉验证来合理地选择模型。

本节注释 图 5.3.1 的测试误差和图 5.3.2 的泛化误差曲线都呈现一种类似 U 形曲线。对于传统的单一机器学习模型，U 形曲线具有一般性。但在深度学习中，当深度网络复杂度达到一定规模后，测试误差的表现更加复杂，对于集成学习中一些方法，如随机森林和提升算法，测试误差一般也并没有呈现出 U 形曲线，换言之，集成学习更不易出现过拟合问题。机器学习仍在快速发展，一些传统结论可能会被不断补充和修改。

5.4 机器学习模型的泛化性能

5.3 节以回归问题为例，讨论了偏和方差的折中问题，本节将以分类问题为例，讨论机器学习的另一个理论问题——泛化界。偏和方差的折中与泛化界是机器学习理论中关注的

两个基本问题，都是关于泛化误差的，两者之间也有密切的联系。

在机器学习模型的训练过程中，一般只有一组训练集，通过训练误差的最小化学习到一个模型，但真正关心的是泛化误差，即对不存在于训练集中的新样本，模型的预测性能如何？所以我们要关心一个基本问题：训练误差和泛化误差之间有多大的差距？机器学习的概率近似正确(probably approximately correct，PAC)理论对这个问题进行了研究。下面对该理论给出一个极为简要的介绍。

本节以二分类问题为例，讨论 PAC 理论的一些最基本概念和结论。假设所面对的样本可表示为$(\boldsymbol{x},y)$，$\boldsymbol{x}$ 是输入特征向量，$\boldsymbol{x}\in\mathcal{X}$，$\mathcal{X}$表示输入空间，$y\in\{0,1\}$表示类型，$(\boldsymbol{x},y)$满足概率分布 $p_{\mathcal{D}}(\boldsymbol{x},y)$，简写为 $p_{\mathcal{D}}$，故可用$(\boldsymbol{x},y)\sim p_{\mathcal{D}}$ 表示样本服从的分布。从 $p_{\mathcal{D}}$ 中采样得到满足独立同分布的训练样本集

$$\boldsymbol{D}=\{(\boldsymbol{x}_i,y_i)\}_{i=1}^{N} \tag{5.4.1}$$

其中，每个样本$(\boldsymbol{x}_i,y_i)\sim p_{\mathcal{D}}$。

用机器学习理论常用的术语，将一个机器学习模型称为一个假设 h，$h(\boldsymbol{x})$输出$\{0,1\}$表示类型，即 h 完成映射 $h:\mathcal{X}\to\{0,1\}$。在一个机器学习过程中，所有可能选择的假设构成一个假设空间$\mathcal{H}$。对于任意假设 $h\in\mathcal{H}$，其在训练样本集上的分类错误率定义为训练误差或经验风险，经验风险表示为

$$\hat{R}(h)=\frac{1}{N}\sum_{i=1}^{N}I(h(\boldsymbol{x}_i)\neq y_i) \tag{5.4.2}$$

经验风险表示假设 h 在训练集上的误分类率，这里 $I(\cdot)$为示性函数。可定义一般的泛化误差为

$$R(h)=P_{(\boldsymbol{x},y)\sim p_{\mathcal{D}}}[h(\boldsymbol{x})\neq y] \tag{5.4.3}$$

对于任意样本$(\boldsymbol{x},y)\sim p_{\mathcal{D}}$，不管其是否存在于训练集中，泛化误差均表示其总体统计意义上的误分类率。注意，用 R 表示泛化误差，用“带帽”的符号 $\hat{R}$ 表示经验风险。

若不考虑可实现性，理论上，希望学习到的假设是从$\mathcal{H}$空间找到使泛化误差最小的假设，即

$$h^{*}=\underset{h\in\mathcal{H}}{\operatorname{argmin}}R(h) \tag{5.4.4}$$

实际上，由于无法准确获得 $p_{\mathcal{D}}(\boldsymbol{x},y)$，故无法通过泛化误差优化获得最优假设，实际上总是通过经验风险最小化(empirical risk minimization，ERM)得到一个假设，即

$$\hat{h}=\underset{h\in\mathcal{H}}{\operatorname{argmin}}\hat{R}(h) \tag{5.4.5}$$

我们关心的一个理论问题：通过 ERM 得到的假设 $\hat{h}$ 与真正的泛化误差最小的 h^{*} 之间的泛化误差差距有多大？即 $R(h^{*})$与 $R(\hat{h})$差距有多大？

在继续讨论之前，首先通过一个例子进一步理解以上概念。

例 5.4.1　一个假设空间的例子。在 4.2 节的感知机的例子中，为了与本节分类输出用$\{0,1\}$表示相符，将分类假设修改为

$$h(\boldsymbol{x})=I(\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{x}}\geqslant 0) \tag{5.4.6}$$

这里 $\bar{\boldsymbol{x}}$ 包含了哑元，$\bar{\boldsymbol{w}}$ 包含了偏置系数，是$(K+1)$维参数向量。式(5.4.6)是一个假设，则假设空间为

$$\mathcal{H}=\{h_{\bar{w}} \mid h_{\bar{w}}(\boldsymbol{x})=I(\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{x}}\geqslant 0),\bar{\boldsymbol{w}}\in\mathbf{R}^{K+1}\} \tag{5.4.7}$$

$\mathcal{H}$表示 K 维向量空间中的所有线性分类器集合,其中 $\mathbf{R}^{K+1}$ 为 $K+1$ 维实数集合。由于不同 $\bar{\boldsymbol{w}}$ 构成$\mathcal{H}$的不同成员,故式(5.4.5)可具体化为

$$\hat{h}=\underset{\bar{\boldsymbol{w}}\in\mathbf{R}^{K+1}}{\operatorname{argmin}}\hat{R}(h_{\bar{w}}) \tag{5.4.8}$$

$\hat{h}$ 是 ERM 意义下的假设,一般不能通过学习得到 $h_{\bar{w}}^{*}$。

实际上,感知机的目标函数式(4.2.21)是式(5.4.2)的经验风险的一种逼近,故训练得到的感知机是对 $\hat{h}$ 的一种逼近。

逻辑回归也可做类似理解,其目标函数交叉熵也是式(5.4.2)经验风险函数的一种近似。

为了研究 $R(h^{*})$与 $R(\hat{h})$的关系,给出以下引理。

引理 5.4.1 设 $Z_1,Z_2,\cdots,Z_N$ 是 N 个独立同分布的随机变量,均服从伯努利分布,且 $P(Z_i=1)=\mu$,定义样本均值为 $\hat{\mu}=\dfrac{1}{N}\sum_{i=1}^{N}Z_i$,令 $\varepsilon>0$ 为一个固定值,则

$$P(\mid \mu-\hat{\mu}\mid>\varepsilon)\leqslant 2\exp(-2\varepsilon^{2}N) \tag{5.4.9}$$

该引理说明,对于独立同分布样本,当 N 充分大时,对于给定的 $\varepsilon>0$,均值估计和实际概率值之差大于 ε 的概率是很小的。

接下来,对于$\mathcal{H}$有限的情况,利用引理 5.4.1 导出训练误差和泛化误差的误差界,然后将结论推广到$\mathcal{H}$无限的情况。

5.4.1 假设空间有限时的泛化误差界

首先假设空间$\mathcal{H}$是有限的,即$\mathcal{H}=\{h_1,h_2,\cdots,h_K\}$,假设空间成员数目 $K=|\mathcal{H}|$可能很大,但有限。例如 4.4 节介绍的朴素贝叶斯方法,当特征分量取离散值时,其假设空间是有限的。若每个特征变量取有限值,则第 7 章介绍的决策树假设空间也是有限的。对于例 5.4.1,若 $\bar{\boldsymbol{w}}$ 取值为实数,则假设空间是无限的,但若 $\bar{\boldsymbol{w}}$ 的每个分量是由有限位二进制表示的数值,则其假设空间是有限的(详细讨论见例 5.4.2)。首先讨论$\mathcal{H}$有限的情况。

可从$\mathcal{H}$空间选择一个固定的假设 h_k,利用引理 5.4.1 容易得到对于 h_k,其训练误差和泛化误差的关系。为此定义一个随机变量,对于$(\boldsymbol{x},y)\sim p_{\mathcal{D}}$,定义

$$Z=I[h_k(\boldsymbol{x})\neq y] \tag{5.4.10}$$

即当 h_k 对样本$(\boldsymbol{x},y)$不能正确分类时 $Z=1$。对式(5.4.1)所示的每个样本有 $Z_i=I[h_k(\boldsymbol{x}_i)\neq y_i]$,显然,$h_k$ 的训练误差为

$$\hat{R}(h_k)=\frac{1}{N}\sum_{i=1}^{N}Z_j \tag{5.4.11}$$

对比引理 5.4.1,Z_i 是 I.I.D 的伯努利随机变量,$\hat{R}(h_k)$是对 $R(h_k)$的样本均值估计,则由式(5.4.9)直接得到

$$P[\mid R(h_k)-\hat{R}(h_k)\mid>\varepsilon]\leqslant 2\exp(-2\varepsilon^{2}N) \tag{5.4.12}$$

以上是对于一个固定的 h_k,泛化误差和训练误差之差(绝对值)大于 ε 的概率。对于给定 ε,若样本数 N 足够大,则泛化误差与训练误差相差大于 ε 的概率很小。

利用式(5.4.12)可导出一个更一般的结果。因此,定义事件 A_k 为 $|R(h_k)-\hat{R}(h_k)|>\varepsilon$,则有 $P(A_k)\leqslant 2\exp(-2\varepsilon^2 N)$。利用概率性质,至少存在一个 h(表示为 $\exists h$),其 $|R(h)-\hat{R}(h)|>\varepsilon$ 的概率为

$$\begin{aligned}P[\exists h\in\mathcal{H},|R(h)-\hat{R}(h)|>\varepsilon]&=P(A_1\cup A_2\cup\cdots\cup A_K)\\&\leqslant\sum_{k=1}^{|\mathcal{H}|}P(A_k)\\&\leqslant\sum_{k=1}^{|\mathcal{H}|}2\exp(-2\varepsilon^2N)\\&=2|\mathcal{H}|\exp(-2\varepsilon^2N)\end{aligned}\tag{5.4.13}$$

由于是概率值,由互补性,式(5.4.13)等价表示为,对于所有 h,有

$$P[|R(h)-\hat{R}(h)|\leqslant\varepsilon,\forall h\in\mathcal{H}]\geqslant 1-2|\mathcal{H}|\exp(-2\varepsilon^2N)\tag{5.4.14}$$

在式(5.4.14)中,令

$$\delta=2|\mathcal{H}|\exp(-2\varepsilon^2N)\tag{5.4.15}$$

式(5.4.14)有丰富的内涵,其中有 3 个量:δ、ε、N,以下从几个方面讨论式(5.4.14)的含义,并讨论这 3 个量的关系。

(1) 将式(5.4.14)重写为

$$P[|R(h)-\hat{R}(h)|\leqslant\varepsilon]\geqslant 1-\delta,\quad\forall h\in\mathcal{H}\tag{5.4.16}$$

对于给定的 ε,所有假设 $\forall h\in\mathcal{H}$ 都以不小于 $1-\delta$ 的概率满足 $|R(h)-\hat{R}(h)|\leqslant\varepsilon$,即泛化误差和训练误差之差不大于界 ε。这里 $1-\delta$ 是一个置信概率,当 N 很大时,δ 很小,以很高的概率满足 $|R(h)-\hat{R}(h)|\leqslant\varepsilon$。

(2) 假设空间的元素数目 $|\mathcal{H}|$ 是确定的,若给出 ε 和 δ,则可得到满足以 $1-\delta$ 为概率达到 $|R(h)-\hat{R}(h)|\leqslant\varepsilon$ 所需的样本数目。固定 δ,ε 从式(5.4.15)反解 N 为

$$N=\frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}\tag{5.4.17}$$

由式(5.4.15),N 增大,δ 减小,故可将式(5.4.17)看作满足 δ,ε 约束的最小样本数。故对于给定 δ、ε,样本数可取为

$$N\geqslant\frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}\tag{5.4.18}$$

(3) 在式(5.1.15)中,固定 δ、N,解得

$$\varepsilon=\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}\tag{5.4.19}$$

这给出式(5.4.16)的另一种解释,对于给定的 δ、N,误差界满足

$$|R(h)-\hat{R}(h)|\leqslant\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}\tag{5.4.20}$$

将以上解释总结为定理 5.4.1。

定理 5.4.1　对于假设空间 $\mathcal{H}$,固定 δ、N,则以概率不小于 $1-\delta$,泛化误差与训练误差

满足

$$|R(h)-\hat{R}(h)|\leqslant\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}}$$

或固定 δ、ε,若样本数目取

$$N\geqslant\frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}$$

则以概率不小于 $1-\delta$ 满足 $|R(h)-\hat{R}(h)|\leqslant\varepsilon$。

以上结论对于假设空间中的任意假设 $\forall h\in\mathcal{H}$ 均成立。我们更感兴趣的一个问题是,对于以经验风险最小化学习得到的假设 $\hat{h}$ 和理论上泛化误差最小对应的 h^* 之间的泛化误差的比较。由以上结果,若对 $\forall h\in\mathcal{H}$ 有 $|R(h)-\hat{R}(h)|\leqslant\varepsilon$,则

$$R(h)\leqslant\hat{R}(h)+\varepsilon \tag{5.4.21}$$

对于 $\hat{h}$,可得到如下不等式

$$R(\hat{h})\leqslant\hat{R}(\hat{h})+\varepsilon\leqslant\hat{R}(h^*)+\varepsilon\leqslant R(h^*)+2\varepsilon \tag{5.4.22}$$

式(5.4.22)中第 1 个不等式只是将 $\hat{h}$ 带入式(5.4.21);第 2 个不等式用了 $\hat{R}(\hat{h})\leqslant\hat{R}(h^*)$,这是因为 $\hat{h}$ 是经验误差最小的假设;第 3 个不等式对 h^* 再次使用式(5.4.21)。式(5.4.22)的结论是:ERM 学习得到的假设 $\hat{h}$ 与泛化误差最优的 h^*,其泛化误差之差不大于 2ε。将该结论总结为重要的定理 5.4.2。

定理 5.4.2 对于假设空间 $\mathcal{H}$,固定 δ、N,则以概率不小于 $1-\delta$,泛化误差满足不等式

$$R(\hat{h})\leqslant\min_{h\in\mathcal{H}}R(h)+2\sqrt{\frac{1}{2N}\ln\frac{2|\mathcal{H}|}{\delta}} \tag{5.4.23}$$

或固定 δ、ε,若样本数目取

$$N\geqslant\frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}$$

则以概率不小于 $1-\delta$ 满足 $R(\hat{h})\leqslant\min_{h\in\mathcal{H}}R(h)+2\varepsilon$。

由两个定理可见,若固定 δ、N,式(5.4.19)计算了一个界 ε,对于任意 $h\in\mathcal{H}$,训练误差和泛化误差之差以概率 $1-\delta$ 不大于 ε,同时 EMR 最小化得到的假设 $\hat{h}$ 与最小泛化误差之间的差距不大于 2ε。

例 5.4.2 讨论例 5.4.1 的假设空间

$$\mathcal{H}=\{h_{\bar{w}}\mid h_{\bar{w}}(\boldsymbol{x})=I(\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{x}}\geqslant 0),\bar{\boldsymbol{w}}\in\mathbf{R}^{K+1}\}$$

$\bar{\boldsymbol{w}}$ 有 $K+1$ 个系数,若 $\bar{\boldsymbol{w}}$ 的每个分量用字长 L 的二进制表示(计算机中常取 L 为 16、32 或 64,近期一些研究采用低比特实现神经网络时,甚至取 L 为 8、4),则表示 $\bar{\boldsymbol{w}}$ 所需的二进制位数为 $L(K+1)$,故假设空间 $\mathcal{H}$ 共有 $|\mathcal{H}|=2^{L(K+1)}$ 个元素,由式(5.4.18)可得,对于固定 δ、ε 时,样本数需满足

$$N\geqslant\frac{1}{2\varepsilon^2}\ln\frac{2|\mathcal{H}|}{\delta}=\frac{1}{2\varepsilon^2}\ln\frac{2\times 2^{L(K+1)}}{\delta}=O\left(\frac{KL}{\varepsilon^2}\ln\frac{1}{\delta}\right)=O_{\varepsilon,\delta}(K) \tag{5.4.24}$$

或记为

$$N \sim O_{\varepsilon,\delta}(K) \tag{5.4.25}$$

式中，$O(\cdot)$表示量级；$O_{\varepsilon,\delta}(\cdot)$表示量级函数，其中$\delta$、$\varepsilon$是其参数。式(5.4.25)尽管有比例系数存在，但需要的样本数N与模型的参数数目K是呈线性关系的。

*5.4.2　假设空间无限时的泛化误差界

将定理5.4.2的结论推广到假设空间$\mathcal{H}$无限的情况。如前所述，例5.4.1所示的线性模型或后续章节介绍的支持向量机和神经网络等模型，当参数取实数时，假设空间是无限的，即$|\mathcal{H}|=\infty$，这种情况下式(5.4.23)变得无意义。需要对其进行推广。这里对假设空间无限的情况，只做一个非常扼要的介绍。

当$|\mathcal{H}|=\infty$时，为了表示假设空间的表示能力(或容量)，给出两个概念：打散(shatters)和VC维(Vapnik-Chervonenkis dimension)，这里只给出其简单、直观性的介绍。

首先给出打散的概念，对于一个包含d个点的集合$S=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_d\}$，其中$\boldsymbol{x}_i\in\mathcal{X}$，称$\mathcal{H}$可打散$S$是指：对点集$S$对应加上一个任意标注集$\{y_1,y_2,\cdots,y_d\}$，则必存在$h\in\mathcal{H}$，使$h(x_i)=y_i$，$i=1,2,\cdots,d$。

对于一个假设空间$\mathcal{H}$，其VC维的定义为：至少存在一个最大元素数为d的点集合S，$\mathcal{H}$可打散S，则$\mathcal{H}$的VC维为d，记为$VC(\mathcal{H})=d$。这里d是最大能被$\mathcal{H}$打散的点集的元素数，对于有$d+1$个元素的点集合，$\mathcal{H}$均不可能打散它。

例5.4.3　图5.4.1给出二维平面上的3个点组成的点集和其对应的各种标注，直线表示判决线，假设空间为二维线性分类器，即

$$\mathcal{H}=\{h(\boldsymbol{x})=\theta_1x_1+\theta_2x_2+\theta_0\mid\theta_0,\theta_1,\theta_2\in R\}$$

图中每条判决线属于$\mathcal{H}$，对这个3个元素的点集，$\mathcal{H}$可将其打散。如果是4个点，则对于标注是异或运算，$\mathcal{H}$不能正确分类，故$\mathcal{H}$不能打散4个点的点集，因此，平面线性分类器的VC维为3，即$VC(\mathcal{H})=3$。

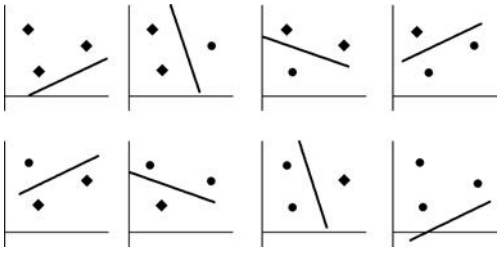

图5.4.1　可由线性分类器打散的点集

若用VC维d表示一个假设空间的容量，则可得到与定理5.4.2类似的结论，这里只给出对应的定理。

定理5.4.3　对于假设空间$\mathcal{H}$，若其VC维为$d=VC(\mathcal{H})$，则对于所有$h\in\mathcal{H}$，以概率不小于$1-\delta$，有不等式

$$|R(h)-\hat{R}(h)|\leqslant O\left(\sqrt{\frac{d}{N}\ln\frac{N}{d}+\frac{1}{N}\ln\frac{1}{\delta}}\right) \tag{5.4.26}$$

对于$\hat{h}$有不等式

$$R(\hat{h}) \leqslant \min_{h \in \mathcal{H}} R(h) + O\left(\sqrt{\frac{d}{N}\ln\frac{N}{d} + \frac{1}{N}\ln\frac{1}{\delta}}\right) \tag{5.4.27}$$

对于以概率不小于 $1-\delta$ 满足 $R(\hat{h}) \leqslant \min_{h \in \mathcal{H}} R(h) + 2\varepsilon$，样本数目需要求

$$N \sim O_{\varepsilon,\delta}(d) \tag{5.4.28}$$

注意，这里只用了量级函数 $O(\cdot)$ 而忽视了表达式中的一些具体常数系数，对于了解性能界，这就够了。从式(5.4.28)看，对于无限假设空间，所需样本数与假设空间的 VC 维呈线性关系。

机器学习理论给出对学习性能的整体性洞察，是非常有意义的，但目前对于一类实际模型的学习算法(如逻辑回归、神经网络、决策树等)，其给出的一些要求(如样本数等)与具体算法的实际需求还有较大距离，在指导一类具体算法的参数选择上的实际意义还有待改善，故实际中仍更常用交叉验证等技术确定机器学习模型的各类参数。

5.5 本章小结

本章介绍了基本的机器学习性能与评估，从两个方面进行讨论。首先从实用方面，介绍了利用数据集通过交叉验证和测试的实际技术训练一个机器学习模型的基本过程，并介绍了几个实际中常用的机器学习性能评估指标。然后从理论上讨论了机器学习的性质。为了讨论上的直观，结合回归介绍了偏和方差的折中问题，结合分类介绍了泛化界定理。

本书更侧重于机器学习算法的介绍，有关机器学习理论的讨论非常简略，对机器学习理论更感兴趣的读者，可参考 Mohri 等的教材(参考文献[32])和 Shai S-S 等的著作(参考文献[44])，这两本书对机器学习理论进行了较为深入的讨论，均由张文生等译成了中文。Vapnik 对于统计学习理论给出了一个简明版的读本(参考文献[50])，已由张学工译成中文。

习题

1. 什么是交叉验证？假设有 1000 个带标注的样本，设计一个参数模型 $\hat{\boldsymbol{y}} = h(\boldsymbol{x};\boldsymbol{w})$，讨论你可能选择的样本集划分方式和交叉验证过程。

2. 对于如下表示的线性分类器的假设空间

$$\mathcal{H} = \{h_{\bar{w}} \mid h_{\bar{w}}(\boldsymbol{x}) = I(\bar{\boldsymbol{w}}^{\mathrm{T}}\bar{\boldsymbol{x}} \geqslant 0), \bar{\boldsymbol{w}} \in \mathbf{R}^{K+1}\}$$

设 $\bar{\boldsymbol{w}}$ 有 11 个系数，若 $\bar{\boldsymbol{w}}$ 的每个分量用字长 8 比特的二进制表示，若得到模型的经验误差和泛化误差的差距以 0.95 的概率不大于 0.05，则样本数应至少取多少？