学习理论与模型选择

王裕 经济系

2019.04.21

本文首先介绍了与模型选择问题相关的学习理论，随后详细介绍了模型选择的基本思想与几种常用方法。

一、学习理论

1、empirical risk minimization（经验风险最小化）

假设有 m 个样本的训练集，并且每个样本都是相互独立地从概率分布 D 中生成的。对于假设 h，定义 training error 训练误差（或者叫 empirical risk 经验风险）为：h 误分类样本占整个训练集的比重：

$$\hat{\varepsilon}(h) = \frac{1}{m}\sum_{i=1}^{m} 1\{h(x^{(i)}) \neq y^{(i)}\}$$

另外，定义 generalization error 泛化误差为：从生成训练集的概率分布 D 中生成新的样本，假设 h 误分类的概率

$$\varepsilon(h) = P_{(x,y)\sim\mathcal{D}}(h(x) \neq y)$$

值得注意的是：假设训练集与新的样本都相互独立地由同一个分布 D 产生（IID 独立同分布）是学习理论里重要的基础。

当我们选择模型参数时使用如下方法：

$$\hat{\theta} = \arg\min_{\theta} \hat{\varepsilon}(h_\theta)$$

就是所谓的经验风险最小化(empirical risk minimization)，经验风险最小化是一个非凸优化问题难以求解，而 logistic 回归与 SVM 可以看做这个问题的凸优化近似。

为了使得定义更加一般化（不仅限于线性分类器参数θ的选择），我们定义 hypothesis class H 为一个学习算法所考虑的所有分类器（函数 h），则经验风险最小化变为：

$$\hat{h} = \arg\min_{h\in\mathcal{H}} \hat{\varepsilon}(h)$$

2、最小化经验风险与最小化泛化误差的关系

2.1 对于有限的 H

这里将直接给出公式推导的结论，具体的步骤可参考 cs229 课程笔记。首先，根据 union bound Lemma 与 Cherno bound Lemma 可以推出：

$$P(\forall h \in \mathcal{H}. |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| \leq \gamma) \geq 1 - 2k\exp(-2\gamma^2 m)$$

该式被称为 uniform convergence，其中 k 为 H 中包含的模型假设数量。再将使得经验风险最小化的假设与使得泛化误差最小化的假设带入，可得：

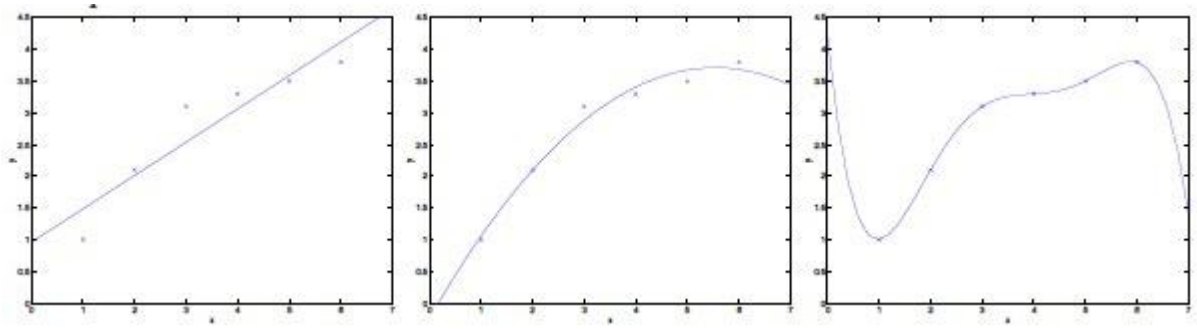
$$\begin{aligned}\varepsilon(\hat{h}) &\leq \hat{\varepsilon}(\hat{h}) + \gamma \\ &\leq \hat{\varepsilon}(h^*) + \gamma \\ &\leq \varepsilon(h^*) + 2\gamma\end{aligned}$$

也就是说，在 uniform convergence 的前提下，通过最小化经验风险选出的模型的泛化误差与最小化泛化误差模型的分类准确度只差 2γ。

由此，有定理：设 H 中包含的假设数量为 k 样本量为 m，则至少在 1-δ的概率下我们有：

$$\varepsilon(\hat{h}) \leq \left(\min_{h\in\mathcal{H}} \varepsilon(h)\right) + 2\sqrt{\frac{1}{2m}\log\frac{2k}{\delta}}$$

有了这个定理，就可以更全面地理解 Andrew 所说的 bias 与 variance（与概率中的方差无关）



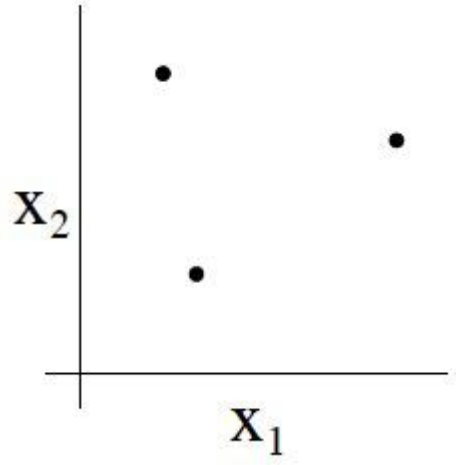
最左边的图对应着 high bias，因为 hypothesis class 中可选的假设少，因此即使大大提升样本量也无法降低最小泛化误差，这种现象对应着欠拟合；
最右边的图对应着 high variance，因为增大了 hypothesis class（增加了参数），因此 min ε(h) 肯定会下降或者保持不变，但是 k 增大将导致γ也增大，尤其是当 m 很小时将有可能导致虽然经验风险很小但是泛化误差依然很大，这种现象对应着过拟合。

而最好的方案是在 high bias 与 high variance 之间做好折中，即对应着中间那副图。
最后还有一个用于求样本复杂度的推论：设 H 中包含的假设数量为 k，则至少在 1-δ的概率下，要使最小化经验风险的泛化误差与最小泛化误差的差值小于等于γ，需要确保样本量：

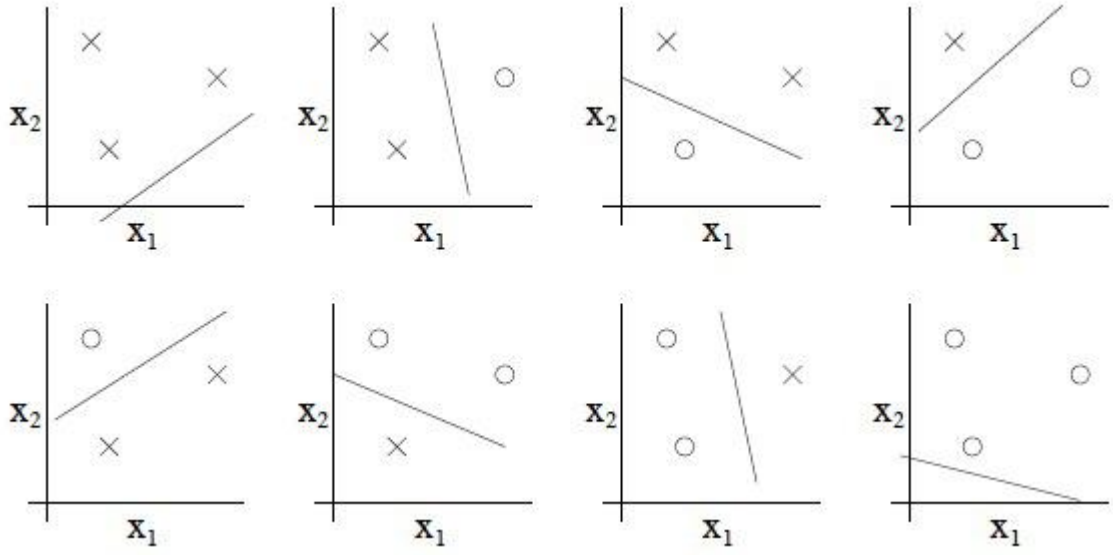
$$\begin{aligned}m &\geq \frac{1}{2\gamma^2}\log\frac{2k}{\delta} \\ &= O\left(\frac{1}{\gamma^2}\log\frac{k}{\delta}\right)\end{aligned}$$

2.2 对于无限的 H

在大多数情况下，H 包含的假设都是无限的，因此有必要把上一小节的结论推广至无限的 H 中，但是证明步骤过于复杂因此 Andrew 也略过了证明过程。在介绍定理前需要先说一下 Vapnik-Chervonenkis dimension(VC 维度)：首先定义 shatter：给定样本集 S，H shatters S 是指 H 可以划分出 S 上所有的标记。而 H 的 VC 维度 VC(H)则定义为：能被 H shatter 的最大集合的样本量。举例来说，对下图集合中的三个点：



h(x) = 1{θ0 + θ1x1 + θ2x2 ≥0} 可以 shatter 这三个点：



并且这个 H 无法 shatter4 个点的集合，因此 VC(H)=3。
定理：对于给定的 H，h∈H，d=VC(H)，在至少 1-δ的概率下，我们有：

$$|\varepsilon(h) - \hat{\varepsilon}(h)| \leq O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

$$\varepsilon(\hat{h}) \leq \varepsilon(h^*) + O\left(\sqrt{\frac{d}{m}\log\frac{m}{d} + \frac{1}{m}\log\frac{1}{\delta}}\right)$$

回忆一下，svm 通过核函数可以使用无限维度的特征，那么是否会出现过拟合？Andrew 给出的答案是不会：可以证明 svm 的 VC 维度是有上界的（即使特征向量的维度是无限的），因此 variance 不会过大。

推论：在至少 1-δ的概率下，要使最小化经验风险的泛化误差与最小泛化误差的差值小于等于γ，则样本量 m 的复杂度为 Oγδ(d)。

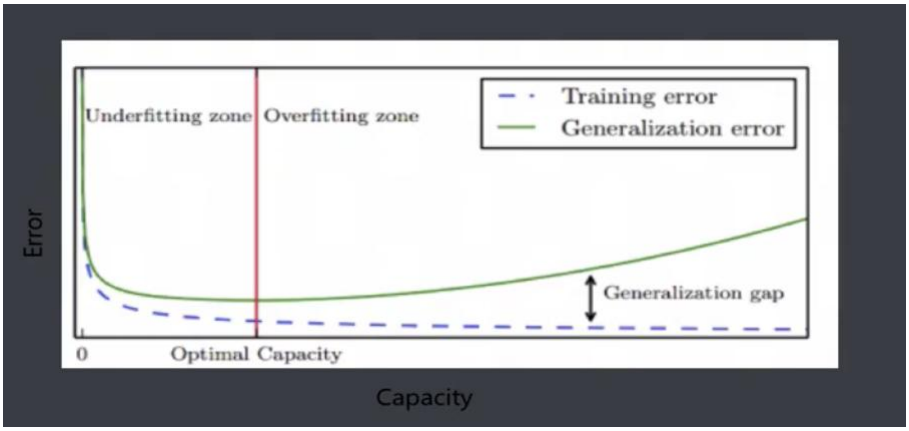
如果严格按照推论得到的公式来计算所需要的样本量，通常会发现根本无法取得如此大的样本量。但是 Andrew 提到一个粗略估计的方法：要使得算法学习得比较好，需要训练集的样本量 m 与 VC(H)是线性关系；而对于大多数 H，VC(H)又大致与模型参数的个数线性相关；联合起来就是，通常需要的训练集样本量应该大致与模型参数的个数为线性关系。

最后，Andrew 还提到了自己的经验：做逻辑回归，训练集样本量为参数个数的 10 倍时，通常就可以拟合出不错的边界，即使只有不到十倍，也还可以接受。

二、模型选择

1、模型选择(Model Selection)

针对某个具体的任务，通常会有多种模型可供选择，对同一个模型也会有多组参数，可以通过分析，评估模型的泛化误差，选择泛化误差最小的模型



红线是最佳模型选择

2、模型选择的核心思想

模型选择核心思想就是从某个模型类中选择最佳模型。注意，它与一般的“调参”意义不同，调参很多时候可能是针对优化算法中的某些参数进行调整，比如步长（学习速率）、mini-batch 大小、迭代次数等，也会涉及到模型 $f(x,\alpha)$ 中调整参数（也称正则参数）α 的选择，但是模型选择不涉及算法中的参数，仅涉及模型目标函数中的调整参数 α 。

从上面叙述可得知模型选择阶段，最标准的方法自然在训练集 $T(X,Y)$ 上训练模型，然后在验证集上获取预测误差 $E\mathrm{rr}_\tau = E_{X^O Y^O}\left(L\left(Y^O, f\left(\hat{X}^O\right)\right) | \tau \right)$，该误差也被称作“样本外（extra-sample）误差”，可真实反映出模型的样本外的预测能力，最后选择最小预测误差所对应

的模型作为最佳模型即可。但通常而言，独立的验证集我们也没有，手头仅有的信息就是训练集，那么要想估计测试误差或者其期望曲线，就只能在训练集上做文章，一般而言可能仅有两种思路：

1、从训练集划分点数据出来形成验证集来近似测试误差；

2、对训练误差进行某种转化来近似测试误差。

第一种思路是非常自然的思路，只要对训练集进行合适的划分，我们就有可能近似出预测误差 Err_τ 。但是对原始训练集 τ 划分为新的训练集 τ_{new} 和验证集，不同的划分比例可能使得新训练集与原训练集相差较大，进而使得 Err_τ 差异很大，因此用这种划分的方式来估计条件期望形式的预测误差 Err_τ 比较困难。那么此时我们可以不估计 Err_τ 转为估计其期望，即平均预测误差 Err ，通过重复抽样的方式来多次估计预测误差 Err_τ ，然后取其平均即可，这种方式我们可以称其为 **“重复抽样法”**：通过训练集多次切分、抽样来模拟训练集、验证集，计算多个“样本外误差”，然后求其平均预测误差，这是一种密集计算型方法，比如交叉验证（Cross Validation）、自助法（bootstrap）等。

第二种思路相比第一种思路更加考虑计算效率，因为重复抽样需要计算多次估计，因此做一次模型选择可能需要花费不少时间，如果单单从训练集的训练误差就可以近似出测试误差 Err_τ ，那么模型选择效率便会大大提高。这种方式以统计学中的 AIC、BIC 等为代表，深刻剖析训练误差与之前提到的“样本内（in-sample）误差”、预测误差 Err_τ 间的关系，给出了预测误差估计的解析式，因此第二种思路我们可以称之为 **“解析法”**。

这两种思路在统计学习和机器学习中都有大量应用，相比较而言，统计学习更喜欢第二种解析法，这样容易计算，并且会较好的理论性质（似然角度）；而机器学习则更喜欢第二种重复抽样法和从 VC 维衍生出来的结构风险最小化法，不需要计算基于分布的似然，普适性更好。

3、模型选择的方法

3.1 几种方法的介绍

留出法(Hold-out)：将已知数据集分成两个互斥的部分，其中一部分用来训练模型，另一部分用来测试模型，评估其误差，作为泛化差的估计。

1）两个数据集的划分尽可能保持数据分布一致性，避免因数据划分过程引入人为的偏差。

2）数据分割存在多种形式会导致不同的训练集，测试机划分，单次留出法结果往往存在偶然性，其稳定性较差，通常会进行若干次随机划分，重复实现评估取平均值作为评估结果。

3）数据集拆分成两部分，每部分的规模设置会影响评估结果，测试，训练的比例一般是 2:8，3:7 等。

4）数据划分一般使用分层法，比如一个训练集里分男人和女人，那么按照男人和女人来划分数据。

交叉验证法(Cross Validation)：将数据集划分为 k 个大小相似的互斥的数据子集，子集数据尽可能保证数据分布的一致性。

留一法(Leave-One-Out. LOO):是 k 折交叉验证的特殊形式，将数据集分成两个，其中一个数据集记录条数为 1，作为测试集使用，其余记录作为训练集训练模型，训练出的模型和使用全部数据集训练得到的模型接近，其评估结果比较准确，缺点是当数据集比较大的时候，训练次数和计算规模比较大。

自助法(Bootstrapping):是一种产生样本的抽样方法，其实质是有放回的随机抽样，即从已知数据集中随机抽取一条记录，然后将该记录放入测试集同时放回原数据集，继续下一次抽样，直到测试集中的数据条数满足要求．通过有放回的抽样获得的训练集去训练模型，不在训练集中的数据（约总数量的 1/3）去用于测试，这样的测试结果被称为包外估计（Out-of-Bag Estimate，OOB）

3.2 几种方法的适用场景

留出法：

优点：实现简单，方便，在一定程度上能够评估泛化误差；测试集和训练集分开，缓解了过拟合。

缺点：一次划分，苹果结果偶然性大；数据被拆分以后，用于训练，测试的数据更少了。

交叉验证法：

优点：K 可以根据实际情况设置，充分利用了所有样本；多次划分，评估结果相对稳定。

缺点：计算比较繁琐，需要进行 k 次训练和评估。

自助法：

优点：样本量比较小的时候可以通过自助法产生多个自助样本集，且有约 36.8%的测试样本

对于总体的理论分布没有要求。

缺点：无放回抽样引起额外的偏差。

3.3 几种方法的选择

已知数据集梳理充足时，通常采用留出法或者 K 折交叉验证法。
对于已知数据集比较小且难以有效划分训练集/测试集的时候，采用自助法。
对于已知数据集比较小且可以有效划分训练集/测试集的时候，采用留一法。

参考文献：
https://cosx.org/2015/08/some-basic-ideas-and-methods-of-model-selection

https://www.cnblogs.com/logosxxw/p/4761087.html

https://www.jianshu.com/p/c5fec07c8f8d