斯坦福 ML 公开课笔记 9

好久没有更新 ML 课程笔记了，主要是开学事情比较多，再加上这几天看了很久的 MIT 算法导论公开课，再加上各种上课、coding 和一些电视剧的诱惑，几乎将本系列阻断。窃以为对于专业理论的学习学的扎实比学得多更重要，这也是我一直以来坚持写博客的动力。尤其是对于本系列的内容来说，如果没有这些博客总结，我可能就不能一路顺利的理解到第 9 个视频，很可能会中途而废，当然，Ng 的原意是要把讲解的公式推导自己推下来，我这里用博客总结代替有异曲同工之妙。这里跟大家分享一下我的心得体会，为好久没更新找找借口，为坚持下去找找理由。

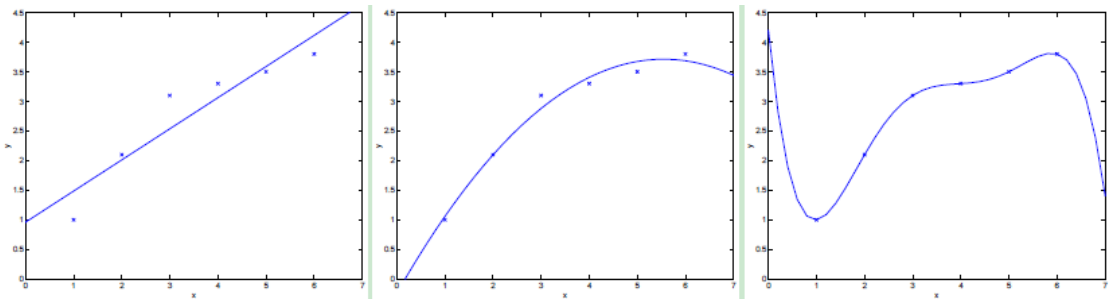
本篇与前面不同，主要内容不是算法，而是机器学习的另一部分内容——学习理论。主要包括偏差/方差（Bias/variance）、经验风险最小化（Empirical Risk Minization，ERM）、联合界（Union bound）、一致收敛（Uniform Convergence）。

Ng 对学习理论的重要性很是强调，他说理解了学习理论是对机器学习只懂皮毛的人和真正理解机器学习的人的区别。学习理论的重要性在于通过它能够针对实际问题更好的选择模型，修改模型。

偏差/方差

偏差与方差对应的仍然是过拟合与欠拟合的问题，本篇主要解决的问题就在于构建一个模型，对何时出现过拟合和欠拟合进行说明。

关于过拟合与欠拟合的问题，对于 MLers 来说，应该是耳熟能详的了。下面再简要介绍一下，如下图：



以回归问题为例，机器学习的目标是从训练集中得到一个模型，使之能对测试集进行分类，这里，训练集与测试集都是分布 D 的样本。机器学习的关注点在于模型在测试集上的分类效果，这也称为泛化能力（generalization ability）。上图中的左图和右图没有较好的泛化能力。对于左图来说，用一个线性模型去拟合二次模型，即便在一个有很多样本的样本集中训练仍然会有很大的泛化误差，这种情况称之为欠拟合，对应着高偏差。对于右图来说，用一个高阶模型去拟合二次模型，从数据中得到的模型结构很可能碰巧是该训练集特有的，这种模型仍然有很大的泛化误差，这种情况称之为过拟合，对应于高方差。

在机器学习中，对偏差和方差的权衡是学习理论着重要解决的问题。

经验风险最小化

本篇都是以二类分类问题为例，对 ERM 问题进行说明。

首先，我们定义数据集：

$$S = \{x^{(i)}, y^{(i)}\}, 1 \le i \le m$$

其中，$x^{(i)}, y^{(i)}$是 i.i.d（独立同分布变量）。$y \in \{0,1\}$。

其次，分类模型为：

$$h_\theta(x) = g(\theta^T x)$$

$$g(z) = I\{z \geq 0\}, g\epsilon\{0,1\}$$

由定义可知，$h_\theta(x)$的输出只能为 0 或者 1。

定义训练误差：

$$\hat{\varepsilon}(h_\theta) = \hat{\varepsilon}_S(h_\theta) = \frac{1}{m}\sum_{i=1}^{m} I\{h_\theta(x^{(i)}) \neq y^{(i)}\}$$

那么经验误差最小化即为：

$$\hat{\theta} = \underset{\theta}{\operatorname{argmin}}\, \hat{\varepsilon}_S(h_\theta)$$

即选择使训练误差最小的参数。

对于 ERM 来说，因为它是非凸的，故而一般的算法无法优化它，因为它是 NP 的。但值得注意的是，logistic 回归与 SVM 都是这种方法的凸性近似。

再看另外一种等价的 ERM 的定义，假设模型集合：

$$\mathcal{H} = \{h_\theta : \theta \in R^{n\times 1}\}$$

其中，h_θ为分类模型，输出为 0,1。其训练误差的定义为：

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

那么 ERM 的定义为：

$$\hat{h} = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\, \hat{\varepsilon}(h)$$

而我们关心的泛化能力的定义为：

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

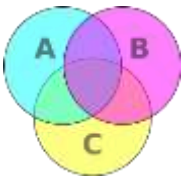
接下来的任务，是证明最优化 ERM 能带来较小的泛化误差。首先，我们引入两个引理。

联合界与 Hoeffding 不等式

首先是联合界定理，令 A1,A2,…,Ak 是 k 个事件，这 k 个事件可以相互独立也可以不相互独立，那么我们会得到：

$$P(A_1 \cup A_2 \cup \ldots \cup A_k) \leq P(A_1) + P(A_2) + \cdots + P(A_k)$$

该定理可以用文氏图来说明：



如左图所示，圆 A,B,C 分别表示事件 A,B,C 发生的概率，那么因为有重叠部分，所以 A,B,C 任意一个发生的概率肯定小于 A,B,C 分别发生的概率之和。

接下来是 Hoeffding 不等式引理，令 Z1,Z2,…,Zm 为 m 个独立同分布(i.i.d)变量，它们都服从 Bernoulli 分布，即：

$$P(Z_i = 1) = \varphi, P(Z_i = 0) = 1 - \varphi$$

我们使用这 m 个 i.i.d 的平均值来估计φ。得到：

$$\hat{\varphi} = \frac{1}{m}\sum_{i=1}^{m} Z_i$$

那么 Hoeffding 不等式的定义即为对于任意的固定数值$\gamma > 0$，存在：

$$P(|\hat{\varphi} - \varphi| > \gamma) < 2\exp(-2\gamma^2 m)$$

这个定理的意义在于，当样本数目 m 增大时，我们对参数的估计将越来越逼近真实值。

一致收敛

我们使用 ERM 的第二种定义来证明该定理。首先推导当模型集合是有限集合的时候成

立的定理。

定义模型集合为：

$$\mathcal{H} = \{h_1, h_2, \dots, h_k\}$$

首先，我们证明对于所有的 h 来说，$\hat{\varepsilon}$都是ε的一个很好的估计；其次，我们证明使用 ERM 方法得到的$\hat{h}$的泛化误差是有上限的。

证明第一个：

从模型集合中任意选择一个假设 h_j，那么会有：

$$P(Z_i = 1) = \varepsilon(h_j) \sim \text{Bernoulli}(\varphi)$$

而训练误差的定义是 m 个$I(Z_i = 1)$之和，即为 m 个服从 Bernoulli 分布的随机变量之和，因而根据 Hoeffding 不等式引理，得到：

$$P\left(\left|\varepsilon(h_j) - \hat{\varepsilon}(h_j)\right| > \gamma\right) \le 2\exp(-2\gamma^2 m)$$

由此，第一个定理得证。

令事件 A_j 为$\left|\varepsilon(h_j) - \hat{\varepsilon}(h_j)\right| > \gamma$，那么有

$$P(A_j) \le 2\exp(-2\gamma^2 m)$$

那么可以推导出至少存在一个假设 h_i，使$|\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma$成立的概率为：

$$P(\exists h \in \mathcal{H}, |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) = P(A_1 \cup A_2 \cup \dots \cup A_k)$$

$$\le P(A_1) + P(A_2) + \cdots + P(A_k) \le 2k\exp(-2\gamma^2 m)$$

这里使用了联合界引理。用 1 减去不等式两侧，得到：

$$P(\neg h \in \mathcal{H}, |\varepsilon(h_i) - \hat{\varepsilon}(h_i)| > \gamma) \ge 1 - 2k\exp(-2\gamma^2 m)$$

该式即为一致收敛定理，它的意义在于，至少有$1 - 2k\exp(-2\gamma^2 m)$的概率，使得模型集合（这里的称呼有点混乱，我称$\mathcal{H}$为‘模型集合’，却称模型集合中的 h 为‘假设’，大家理解在我的说法里‘模型’与‘假设’其实是一个东西就可以了）中的所有假设，其泛化误差都在训练误差的γ范围内。

一致收敛的推论

在一致收敛中，有三个参数，m, γ,概率。这三个参数是相互关联的，我们可以通过固定其中两个，来推出第三个。其中固定 m, γ来求概率已经得出了，下面依次对另外两种参数关联进行说明。

第一个，给定γ和$\sigma > 0$，需要多少样本，可以保证在至少有 1-σ的概率，使得泛化错误率在训练错误率的γ范围内？

对下式进行求解即得到答案：

$$1 - 2k\exp(-2\gamma^2 m) \ge 1 - \sigma$$

得到：

$$m \ge \frac{1}{2\gamma^2} log \frac{2k}{\sigma}$$

这个推论的意义为，一个算法或者模型要达到一个确定的性能时，需要的样本数目。也称为算法的样本复杂度。

第二个，给定 m 和$\sigma > 0$，泛化错误率会落在训练错误率的什么范围内？

$$1 - 2k\exp(-2\gamma^2 m) \ge 1 - \sigma \Rightarrow \gamma = \sqrt{\frac{1}{2m} log \frac{2k}{\sigma}}$$

下面我们看看在在一致收敛成立的情况下，我们通过 ERM 方法得到的假设$\hat{h}$的泛化能力到底如何？

首先，定义：

$$h^* = \underset{h\in\mathcal{H}}{\operatorname{argmin}}\,\varepsilon(h)$$

即h^*为$\mathcal{H}$中泛化误差最小的假设。

我们可以推出：

$$\varepsilon(\hat{h}) \le \hat{\varepsilon}(\hat{h}) + \gamma \le \hat{\varepsilon}(h^*) + \gamma \le \varepsilon(h^*) + \gamma + \gamma = \varepsilon(h^*) + 2\gamma$$

其中，第一个不等号成立是一致收敛定理的应用；第二个不等号成立是$\hat{h}$的定义决定，其本身为训练误差最小的假设；第三个不等号成立仍然是一致收敛定理的应用。

这表明，在一致收敛定理成立的时候，通过 ERM 得到的训练误差最小的假设在泛化能力上至多比泛化能力最好的假设差2γ。

将这些推论综合一下，我们得到一个定理：

令$|\mathcal{H}| = k$,给定 m 和$\sigma > 0$，那么至少有 1-σ的概率能够成立如下公式：

$$\varepsilon(\hat{h}) \le \min_{h\in\mathcal{H}} \varepsilon(h) + 2\sqrt{\frac{1}{2m} log \frac{2k}{\sigma}}$$

该定理反映了偏差和方差的权衡。可以想象，当选择一个复杂的模型假设时，$|\mathcal{H}| = k$会变大，导致不等式后的第二项变大，意味着方差变大；但是第一项却会变小，因为使用一个更加大的模型集合$\mathcal{H}$意味着可供选择的假设变多了，在多的那部分中可能有比原来还要小的模型，这样偏差就会变小。选择一个最优值，使得偏差与方差之和最小，才能得到一个好的模型。

同样的，该定理还有另外形式的推论：

令$|\mathcal{H}| = k$,给定γ和$\sigma > 0$，那么至少有 1-σ的概率使$\varepsilon(\hat{h}) \le \min_{h\in\mathcal{H}} \varepsilon(h) + 2\gamma$成立的前提是：

$$m \ge \frac{1}{2\gamma^2} log \frac{2k}{\sigma} = O(\frac{1}{\gamma^2} log \frac{2k}{\sigma})$$