大家好，这篇是有关Learning from data第五章习题的详解，这一章主要介绍了机器学习中三个学习原则。

我的github地址：
https://github.com/Doraemonzzz

个人主页：
http://doraemonzzz.com/

参考资料:
https://blog.csdn.net/a1015553840/article/details/51085129
http://www.vynguyen.net/category/study/machine-learning/page/6/
http://book.caltech.edu/bookforum/index.php
http://beader.me/mlnotebook/

Chapter 5 Three Learning Principles

Part 1: Exercise

Exercise 5.1 (Page 168)

Consider hypothesis sets $\mathcal{H}_1$ and $\mathcal{H}_{100}$ that contain Boolean functions on 10 Boolean variables, so $\mathcal{X} = \{-1, +1\}^{10}$. $\mathcal{H}_1$ contains all Boolean functions which evaluate to $+1$ on exactly one input point, and to -1 elsewhere; $\mathcal{H}_{100}$ contains all Boolean functions which evaluate to $+1$ on exactly 100 input points, and to -1 elsewhere.

(a) How big (number of hypotheses) are $\mathcal{H}_1$ and $\mathcal{H}_{100}$?

(b) How many bits are needed to specify one of the hypotheses in $\mathcal{H}_1$?

(c) How many bits are needed to specify one of the hypotheses in $\mathcal{H}_{100}$?

首先分析下题目，输入空间是$\mathbb{R}^{10}$，每个分量为$-1, +1$，$\mathcal{H}_1$和 $\mathcal{H}_{100}$为布尔函数，将$\mathcal{X} = \{-1, +1\}^{10}$映射到$-1$或者$+1$，所以$\mathcal{X} = \{-1, +1\}^{10}$一共有$2^{10}$种可能的输入，有了这些准备工作，可以看下后面的题目。

(a)$\mathcal{H}_1$可以理解为从$\mathcal{X} = \{-1, +1\}^{10}$挑1个点将其映射为$+1$，其余点映射为$-1$，所以$|\mathcal{H}_1| = C_{2^{10}}^1 = 2^{10}$。$\mathcal{H}_{100}$可以理解为从$\mathcal{X} = \{-1, +1\}^{10}$挑100个点将其映射为$+1$，其余点映射为$-1$，所以$|\mathcal{H}_{100}| = C_{2^{10}}^{10}$。

(b)$\mathcal{H}_1$只要记录那1个映射为$+1$的点即可，所以需要10 bits。

(c)$\mathcal{H}_{100}$只要记录那100个映射为$+1$的点即可，所以需要$10 \times 100 = 1000$ bits。

Exercise 5.2 (Page 170)

Suppose that for 5 weeks in a row, a letter arrives in the mail that predicts the outcome of the upcoming Monday night football game. You keenly watch each Monday and to your surprise, the prediction is correct each time. On the day after the fifth game, a letter arrives, stating that if you wish to see next week's prediction, a payment of 50.00 is required. Should you pay?

(a)How many possible predictions of win-lose are there for 5 games?

(b)If the sender wants to make sure that at least one person receives correct predictions on all 5 games from him, how many people should he target to begin with?

(c)After the first letter 'predicting' the outcome of the first game, how many of the original recipients does he target with the second letter?

(d)How many letters altogether will have been sent at the end of the 5 weeks?

(e)If the cost of printing and mailing out each letter is 0.50, how much would the sender make if the recipient of 5 correct predictions sent in the 50.00?

(f)Can you relate this situation to the growth function and the credibility of fitting the data?

(a)对于5场比赛来说，一共有$2^5 = 32$种可能。

(b)要使得有人收到的5次预测全对，只要给32个人同时写信即可，方法如下，第一天告诉一半的人A队胜利，告诉另一半人B队胜利，那么第一天必然有16封信是正确的，对这16个人重复此操作，到第五天肯定有人收到的5封信都是正确的。

(c)由(b)的方法可知，第二封信只要寄给16个人。

(d)第一次要寄送2^5封信，第二次2^4，以此类推可得一共要寄送

$$2^5 + 2^4 + 2^3 + 2^2 + 2^1 + 2^0 = 2^6 - 1 = 63$$

(e)由(d)可知成本为

$$63 \times 0.5 = 31.5$$

如果有人出了50元，那么可以赚

$$50 - 31.5 = 18.5$$

(f)这个问题说明，只要假设足够多以及足够复杂，那么总能使得$E_{\text{in}} = 0$。

Exercise 5.3 (Page 172)

In an experiment to determine the distribution of sizes of fish in a lake, a net might be used to catch a representative sample of fish . The sample is then analyzed to find out the fractions of fish of different sizes. If the sample is big enough, statistical conclusions may be drawn about the actual distribution in the entire lake. Can you smell sampling bias?

我认为这题是有样本偏差的，最明显的一个问题是，很小的鱼抓不到，所以这个不是随机抽样。

Exercise 5.4 (Page 174)

Consider the following approach to learning. By looking at the data, it appears that the data is linearly separable, so we go ahead and use a simple perceptron, and get a training error of zero after determining the optimal set of weights. We now wish to make some generalization conclusions, so we look up the d_{vc} for our learning model and see that it is $d + 1$. Therefore, we use this value of d_{vc} to get a bound on the test error.

(a) What is the problem with this bound, is it correct?

(b) Do we know the d_{vc} for the learning model that we actually used? It is this d_{vc} that we need to use in the bound.

(a)因为模型是在我们看了数据之后得到的，所以实际上进行了data snooping，因此这个上界不正确。

(b)实际的d_{vc}无法得到。

Exercise 5.5 (Page 176)

Assume we set aside 100 examples from that will not be used in training, but will be used to select one of three final hypotheses g_1, g_2, g_3 produced by three different learning algorithms that train on the rest on the data. Each algorithm works with a different $\mathcal{H}$ of size 500. We would like to characterize the accuracy of estimating $E_{\text{out}}(g)$ on the selected final hypothesis if we use the same 100 examples to make that estimate.

(a) What is the value of M that should be used in (1.6) in this situation?

(b) How does the level of contamination of these 100 examples compare to the case where they would be used in training rather than in the final selection?

(a)公式1.6在课本的24页，M为假设的数量，由之前的讨论可知，应该要把所有的假设都考虑进来，所以$M = 500 \times 3 = 1500$

(b)如果将测试集加入训练集，那么会增加污染。

Part 2: Problems

Problem 5.1 (Page 178)

The idea of *falsifiability* - that a claim can be rendered false by observed data - is an important principle in experimental science.

Axiom of Non-Falsifability. If the outcome of an experiment has no chance of falsifying a particular proposition, then the result of that experiment does not provide evidence one way or another toward the truth of the proposition.

Consider the proposition "There is $h \in \mathcal{H}$ that approximates f as would be evidenced by finding such an h with in sample error zero on $x_1, \cdots, x_N$." We say that the proposition is falsified if no hypothesis in $\mathcal{H}$ can fit the data perfectly. (a) Suppose that $\mathcal{H}$ shatters $x_1, \cdots, x_N$. Show that this proposition is not falsifiable for any f .

(b) Suppose that f is random ($f(x) = \pm 1$ with probability $\frac{1}{2}$. independently on every x) , so $E_{out}(h) = \frac{1}{2}$ for every $h \in \mathcal{H}$. Show that

$$P[\text{falsification}] \geq 1 - \frac{m_{\mathcal{H}}(N)}{2^N}$$

(c) Suppose $d_{vc} = 10$ and $N = 100$. If you obtain a hypothesis h with zero E_{in} on your data, what can you 'conclude' from the result in part (b)?

首选看下题目中的命题

可以通过在x_1，...，x_N上找到样本误差为零的h来证明$\mathcal{H}$中有h可以近似f

如果$\mathcal{H}$中没有假设h可以使得$E_{in}=0$则这个命题是可证伪的。

(a)因为$\mathcal{H}$可以shatter $x_1,\cdots,x_N$，所以对于每个f，存在$h\in\mathcal{H}$，使得$E_{\text{in}}(h)=0$，从而题目中的命题是不可证伪的。

(b)我们知道N个点一共有2^N种表示方法，即一共有2^N种f。那么对于某个$h\in\mathcal{H}$，h正好等于f的概率为$\frac{1}{2^N}$，而$\mathcal{H}$一共有$m_{\mathcal{H}}(N)$个有效的h，所以$\mathcal{H}$中存在h正好等于f的概率小于等于$\frac{m_{\mathcal{H}}(N)}{2^N}$，即不可证伪的概率小于等于$\frac{m_{\mathcal{H}}(N)}{2^N}$，因此

$$\mathbb{P}[\text{falsification}]\geq 1-\frac{m_{\mathcal{H}}(N)}{2^N}$$

(c)注意

$$m_{\mathcal{H}}(N)\leq N^{d_{vc}}+1$$

所以

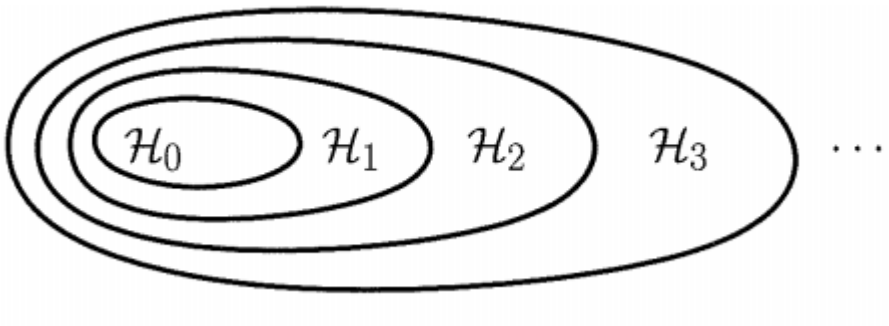
$$\mathbb{P}[\text{falsification}]\geq 1-\frac{m_{\mathcal{H}}(N)}{2^N}\geq 1-\frac{N^{d_{vc}}+1}{2^N}$$

将 $d_{vc}=10$和 $N=100$.带入可得

$$\mathbb{P}[\text{falsification}]\geq 1-\frac{100^{10}+1}{2^{100}}\approx 0.9999999999211139$$

Problem 5.2 (Page 178)

Structural Risk Minimization (SRM) is a useful framework for model selection that is related to Occam 's Razor. Define a structure - a nested sequence of hypothesis sets:



The SRM framework picks a hypothesis from each $\mathcal{H}_i$ by minimizmg E_{in}. That is, $g_i=\underset{h\in\mathcal{H}_i}{\operatorname{argmin}}\ E_{\text{in}}(h)$. Then, the framework selects the final hypothesis by minimizing E_{in} and the model complexity penaltyn Ω. That is, $g^*=\underset{i=1,2...}{\operatorname{argmin}}(\ E_{\text{in}}(g_i)+\Omega(\mathcal{H}_i))$. Note that $\Omega(\mathcal{H}_i)$ should be non decreasing in i because of the nested structure.

(a) Show that the in sample error $E_{\text{in}}(g_i)$ is non increasing in i.

(b) Assume that the framework finds $g^* \in \mathcal{H}_i$ with probability p_i . How does p_i relate to the complexity of the target function?

(c) Argue that the p_i's are unknown but $p_0 \leq p_1 \leq p_2 \leq \cdots \leq 1$.

(d) Suppose $g^* = g_i$.Show that

$$\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon | g^* = g_i] \leq \frac{1}{p_i} 4 m_{\mathcal{H}_i}(2N) e^{-\frac{\epsilon^2 N}{8}}$$

Here, the conditioning is on selecting g_i as the final hypothesis by SRM. [Hint: Use the Bayes theorem to decompose the probability and then apply the VC bound on one of the terms]

You may interpret this result as follows: if you use SRM and end up with g_i , then the generalization bound is a factor $\frac{1}{p_i}$ worse than the bound you would have gotten had you simply started with $\mathcal{H}$.

(a)首先由假设

$$\mathcal{H}_0 \subset \mathcal{H}_1 \subset \mathcal{H}_2 \ldots$$

由集合的性质可知

$$E_{\text{in}}(g_0) \geq E_{\text{in}}(g_1) \geq E_{\text{in}}(g_2) \geq \ldots$$

(b)这部分可以参考(d)，p_i越小，模型复杂度越大。

(c)首先我们知道

$$\begin{gathered} \text{对于} i < j, \text{如果} g^* \subset \mathcal{H}_i, \text{那么} g^* \subset \mathcal{H}_j \\ \text{即对于} i < j, \ g^* \subset \mathcal{H}_i \Rightarrow g^* \subset \mathcal{H}_j \\ \text{从而} \mathbb{P}(g^* \subset \mathcal{H}_i) \leq \mathbb{P}(g^* \subset \mathcal{H}_j) \end{gathered}$$

所以

$$p_0 \leq p_1 \leq p_2 \leq \cdots \leq 1$$

(d)利用贝叶斯公式以及187页VC Bound

$$\begin{aligned} \mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon | g^* = g_i] &= \frac{\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon, g^* = g_i]}{\mathbb{P}(g^* = g_i)} \\ &\leq \frac{\mathbb{P}[|E_{\text{in}}(g_i) - E_{\text{out}}(g_i)| > \epsilon]}{\mathbb{P}(g^* = g_i)} \\ &\leq \frac{1}{p_i} 4 m_{\mathcal{H}_i}(2N) e^{-\frac{\epsilon^2 N}{8}} \end{aligned}$$

Problem 5.3 (Page 179)

In our credit card example, the bank starts with some vague idea of what constitutes a good credit risk. So, as customers $x_1, x_2, \ldots, x_N$ arrive, the bank applies its vague idea to approve credit cards for some of these customers. Then, only those who got credit cards are monitored to see if they default or not.

For simplicity, suppose that the first N customers were given credit cards. Now that the bank knows the behavior of these customers, it comes to you to improve their algorithm for approving credit. The bank gives you the data $(x_1, y_1), \ldots (x_N, y_N)$.

Before you look at the data , you do mathematical derivations and come up with a credit approval function. You now test it on the data and, to your delight, obtain perfect prediction.

(a) What is M, the size of your hypothesis set?

(b) With such an M, what does the Hoeffding bound say about the probability that the true performance is worse than 2% error for $N = 10000$?

(c) You give your g to the bank and assure them that the performance will be better than 2 error and your confidence is given by your answer to part (b) . The bank is thrilled and uses your g to approve credit for new clients. To their dismay, more than half their credit cards are being defaulted on. Explain the possible reason(s) behind this outcome.

(d) Is there a way in which the bank could use your credit approval function to have your probabilistic guarantee? How? [Hint: The answer is yes!]

(a)因为已经用数学推导了一个数学模型，所以此处$M = 1$

(b)回顾Hoeffding不等式

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N}$$

此处$M = 1, N = 10000, \epsilon = 0.02$

带入可得

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \leq 2Me^{-2\epsilon^2 N} \approx 0.000671$$

(c)之所以结果那么差，问题在于我们使用的数据是在原有模型筛选过之后的，相当于data snooping。

(d)如果要提升结果，可以参考原有的模型以及后来的到结果，假设原来的模型为f，我们要参考

$$f(x) \text{ AND } g(x)$$

只有两个模型都通过，才批准，这样模型的结果至少不会比原来的模型差。

Problem 5.4 (Page 180)

The S&P 500 is a set of the largest 500 companies currently trading. Suppose there are 10, 000 stocks currently trading, and there have been 50,000 stocks which have ever traded over the last 50 years (some of these have gone bankrupt and stopped trading) . We wish to evaluate the profitability of various ' buy and hold' strategies using these 50 years of data (roughly 12, 500 trading days) .

Since it is not easy to get stock data , we will confine our analysis to today's S&P 500 stocks, for which the data is readily available.

(a) A stock is profitable if it went up on more than 50% of the days. Of your S & P stocks, the most profitable went up on 52% of the days ($E_{in} = 0.48$).

(i) Since we picked the best among 500, using the Hoeffding bound,

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > 0.02] \leq 2 \times 500 \times e^{-2\times 12500 \times 0.02^2} \approx 0.045.$$

There is a greater than 95% chance this stock is profitable. Where did we go wrong?

(ii) Give a better estimate for the probability that this stock is profitable. [Hint: What should the correct M be in the Hoeffding bound?]

(b) We wish to evaluate the profitability of 'buy and hold ' for general stock trading. We notice that all of our 500 S&P stocks went up on at least 51% of the days.

(i) We conclude that buying and holding a stocks is a good strategy for general stock trading. Where did we go wrong?

(ii) Can we say anything about the perfrmance of buy and hold trading?

(a)(i)我的理解是，这里我们人为把数据减少为500，这是一种data snooping，这里的M应该是历史上有交易的股票数量，所以$M = 50000$。

(a)(ii)将$M = 50000$带入

$$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > 0.02] \leq 2 \times 50000 \times e^{-2\times 12500 \times 0.02^2} \approx 4.539992976248485$$

(b)(i)我感觉还是data snooping的问题，应该考虑全部股票，而不是只考虑这500个。

(b)(ii)如果只考虑这500个股票，是无法做结论的，只有考虑全部股票才能下结论。

Problem 5.5 (Page 180)

You think that the stock market exhibits reversal , so if the price of a stock sharply drops you expect it to rise shortly thereafter. If it sharply rises, you expect it to drop shortly thereafter.

To test this hypothesis, you build a trading strategy that buys when the stocks go down and sells in the opposite case. You collect historical data on the current S&P 500 stocks, and your hypothesis gave a good annual return of 12%.

(a) When you trade using this system, do you expect it to perform at this level? Why or why not?

(b) How can you test your strategy so that its performance in sample is more reflective of what you should expect in reality?

(a)实际使用时不一定有训练时的结果，因为实际情况不一定是大涨大跌，有可能小幅波动。

(b)要在大盘的各种情形下训练模型，不止是大涨大跌。

Problem 5.6 (Page 180)

One often hears "Extrapolation is harder than interpolation." Give a possible explanation for this phenomenon using the principles in this chapter. [Hint: training distribution versus testing distribution.]

之所以外推更难，因为测试数据和训练数据的分布不一定一致，或者训练数据有data snooping。