Part 1 Foundations

2 A Gentle Start

2.1 A FORMAL MODEL - THE STATISTICAL LEARNING FRAMEWORK

1. The learner's input:
    1. Domain set : set of objects $\mathcal{X}$
    2. Label set : set of labels $\mathcal{Y}$
    3. Training data : finite sequence $S = \mathcal{X} \times \mathcal{Y}$
2. The learner's output: $h : \mathcal{X} \to \mathcal{Y}$ comes from learning algorithm $A(S)$
3. A simple data-generation model : $\mathcal{X} \sim \mathcal{D}$, exists but unknown
4. Measures of success, true error : $h : \mathcal{X} \to \mathcal{Y},\ real\ labeling\ function f$

   $L_{\mathcal{D},f}(h) := \underset{x\sim\mathcal{D}}{\mathbb{P}}[h(x) \neq f(x)] := \mathcal{D}(\{x : h(x) \neq f(x)\})$
5. The learner is blind to $\mathcal{D}$ and f

2.2 EMPIRICAL RISK MINIMIZATION

1. **Empirical Risk Minimization(ERM)** , training error instead of true error:

$L_s(h) := \frac{|\{i \in [m] : h(x_i) \neq y_i\}|}{m}, [m] = \{1, \ldots, m\}$

2. ERM rule causes overfitting

2.3 EMPIRICAL RISK MINIMIZATION WITH INDUCTIVE BIAS

1. hypothesis class $\mathcal{H}$, gotten before seeing data
2. inductive bias rule : $ERM_{\mathcal{H}}(S) \in \underset{h \in \mathcal{H}}{argmin} L_s(h)$
3. If $\mathcal{H}$ is a finite class then $ERM_{\mathcal{H}}$ will not overfit, provided it is based on a sufficiently large training sample.
    1. The Realizability Assumption : $\exists h^* \in \mathcal{H} \;\; s.t. L_{(\mathcal{D},f)}(h^*) = 0$, which implies we always can get ERM hypothesis that $L_S(h_S) = 0$
    2. The i.i.d. assumption: $S \sim \mathcal{D}^m$, The examples in the training set are independently and identically distributed(i.i.d) according to the distribution $\mathcal{D}$
    3. the probability of getting a nonrepresentative sample : δ

       the confidence parameter of prediction : $1 - \delta$
    4. accuracy parameter : ϵ, (failure of the learner : $L_{(\mathcal{D},f)}(h_S) > \epsilon$)
    5. Let $S|_x = (x_1, \ldots, x_m)$ be the training set, then :

       the upper bound : $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_s) > \epsilon\})$, 该符号对应的结果是一个概率值

       the set of "bad" hypotheses : $\mathcal{H}_B = \{h \in \mathcal{H} : L_{(\mathcal{D},f)}(h) > \epsilon\}$

       the set of misleading samples : $M = \{S|_x : \exists h \in \mathcal{H}_B, L_S(h) = 0\} = \underset{h \in \mathcal{H}_B}{\cup} \{S|_x : L_S(h) = 0\}$

       then, $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq \mathcal{D}^m(M) \leq \sum_{h \in \mathcal{H}_B} \mathcal{D}^m(\{S|_x : L_S(h) = 0\})$
    6. $h_s \in \mathcal{H}$, 但由于训练数据会有多个h 满足$L_S(h) = 0$, 再由于算法的不同，得到不同的h_s
    7. **proof**

       $\mathcal{D}^m(S|_x : L_s(h) = 0) = \mathcal{D}^m(\{S|_x : \forall i, h(x_i) = f(x_i)\}) = \prod \mathcal{D}(x_i : h(x_i) = f(x_i))$

       $\mathcal{D}(x_i : h(x_i) = f(x_i)) = 1 - L_{(\mathcal{D},f)}(h) \leq 1 - \epsilon$

       $\mathcal{D}^m(S|_x : L_s(h) = 0) \leq (1 - \epsilon)^m \leq e^{-\epsilon m}$

       $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq |\mathcal{H}_B| e^{-\epsilon m} \leq |\mathcal{H}| e^{-\epsilon m}$
    8. If we want $\mathcal{D}^m(\{S|_x : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq \delta$, then we can choose $m \geq \frac{\log(|\mathcal{H}|/\delta)}{\epsilon}$

3 A Formal Learning Model

3.1 PAC LEARNING(Probably Approximately Correct)

1. **Definition** :

   A hypothesis class $\mathcal{H}$ is **PAC learnable** if :
   $\exists m_{\mathcal{H}}(x_1, x_2) \to N, \forall \epsilon, \delta \in (0,1), \forall \{S : |S| \geq m_{\mathcal{H}}(\epsilon, \delta), S \sim \mathcal{D}^m\}$,满足$\mathcal{D}^{|S|}(\{S : L_{(\mathcal{D},f)}(h_s) > \epsilon\}) \leq \delta$
2. $m_{\mathcal{H}}(\epsilon, \delta) = \lceil \frac{\log(|\mathcal{H}|/\delta)}{\epsilon} \rceil$
3. infinite hypothesis classes $\mathcal{H}$

   solution：

$\mathbb{P}(\frac{r^2-max(x_i^2+y_i^2)}{r^2}>\epsilon)\leq\delta$

$\mathbb{P}(max(x_i^2+y_i^2)<(1-\epsilon)r^2)\leq\delta$

$[\mathbb{P}(x_i^2+y_i^2<(1-\epsilon)r^2)]^m=(1-\epsilon)^m\leq e^{-\epsilon m}\leq\delta$

$m\geq\lceil\frac{log(1/\delta)}{\epsilon}\rceil\ and\ m_{\mathcal{H}}(\epsilon,\delta)\leq\lceil\frac{log(1/\delta)}{\epsilon}\rceil$

3.2 A MORE GENERAL LEARNING MODEL

- Removing the Realizability Assumption
- Learning Problems beyond Binary Classification

1. **Agnostic PAC Learning**
    1. $L_{\mathcal{D}}(h):=\underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}[h(x)\neq y]:=\mathcal{D}(\{(x,y):h(x)\neq y\})$

       $L_s(h):=\frac{|\{i\in[m]:h(x_i)\neq y_i\}|}{m},[m]=\{1,\dots,m\}$
    2. **The Bayes Optimal Predictor**:$\forall\mathcal{D}=\mathcal{X}\times\{0,1\},f_{\mathcal{D}}(x)=1\{\mathbb{P}[y=1|x]>0.5\}$
    3. $\forall g(x)\in\mathcal{H},L_{\mathcal{D}}(f)\leq L_{\mathcal{D}}(g)$

       proof:

       $L_{\mathcal{D}}(g)=\underset{(x,y)\sim\mathcal{D}}{\mathbb{P}}[g\neq y|x]=\mathbb{P}[g=1,y=0|x]+\mathbb{P}[g=0,y=1|x]$

       $=p(y=0|x)p(g=1|x)+p(y=1|x)p(g=0|x)$

       $L_{\mathcal{D}}(f)-L_{\mathcal{D}}(g)=p(y=0|x)[p(f=1|x)-p(g=1|x)]+p(y=1|x)[p(f=0|x)-p(g=0|x)]$
       $=[1-2p(y=1|x)][p(f=1|x)-p(g=1|x)]<0$
    4. **definition**

       A hypothesis class $\mathcal{H}$ is **agnostic PAC learnable** if :
       $\exists m_{\mathcal{H}}(x_1,x_2)\to N,\forall\epsilon,\delta\in(0,1),\forall\{S:|S|\geq m_{\mathcal{H}}(\epsilon,\delta),S\sim\mathcal{D}^m\}$,满足
       $\mathcal{D}^{|S|}(\{S:L_{\mathcal{D}}(h_s)>\underset{h'\in\mathcal{H}}{min}L_{\mathcal{D}}(h')+\epsilon\})\leq\delta$
2. **The Scope of Learning Problems Modeled**
    1. risk function : $L_{\mathcal{D}}(h):=\underset{z\sim\mathcal{D}}{\mathbb{E}}[l(h,z)]$
    2. empirical risk: $L_S(h):=\frac{1}{m}\sum l(h,z_i)$
    3. $l_{0-1}(h,(x,y))=1\{h(x)\neq y\},l_{sq}(h,(x,y)):=(h(x)-y)^2$
    4. **definition**

       A hypothesis class $\mathcal{H}$ with respect to a set of Z and general loss function l is **agnostic PAC learnable** if : $\exists m_{\mathcal{H}}(x_1,x_2)\to N,\forall\epsilon,\delta\in(0,1),\forall\{S:|S|\geq m_{\mathcal{H}}(\epsilon,\delta),S\sim\mathcal{D}^m\}$,满足
       $\mathcal{D}^{|S|}(\{S:L_{\mathcal{D}}(h_s)>\underset{h'\in\mathcal{H}}{min}L_{\mathcal{D}}(h')+\epsilon\})\leq\delta$

4 Learning via Uniform Convergence

4.1 UNIFORM CONVERGENCE IS SUFFICIENT FOR LEARNABILITY

$(\mathcal{X},\mathcal{Y})=Z$

1. ϵ**-representative sample** :

   A training set S is called ϵ-representative(w.r.t. domain Z, hypothesis class $\mathcal{H}$, loss function l and distribution $\mathcal{D}$) if $\forall h\in\mathcal{H},|L_S(h)-L_D(h)|\leq\epsilon$

2. Any output of $ERM_{\mathcal{H}}(S)$,namely, any $h_S \in argmin_{h\in\mathcal{H}}(h)$, satisfies $L_{\mathcal{D}}(h_S) \leq \min\limits_{h\in\mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

   $L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \frac{\epsilon}{2} \leq L_S(h) + \frac{\epsilon}{2} \leq L_{\mathcal{D}}(h) + \epsilon$

3. **Uniform convergence**

   $\exists m_{\mathcal{H}}^{\mathcal{UC}}(x_1, x_2) \to N, \forall \epsilon, \delta \in (0,1), \forall \{S : |S| \geq m_{\mathcal{H}}^{\mathcal{UC}}(\epsilon, \delta), S \sim \mathcal{D}^m\}$,满足
   $\mathbb{P}(S\ is\ \epsilon - representative) \geq 1 - \delta$

4. If a class $\mathcal{H}$ has the uniform convergence property with a function $m_{\mathcal{H}}^{UC}$, then the sample complexity $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$, the $ERM_{\mathcal{H}}$ paradigm is a successful agnostic PAC learner for $\mathcal{H}$.

   从上面的定义可以直接得出来。

4.2 FINITE CLASS ARE AGNOSTIC PAC LEARNABLE

1. **Hoeffding's Inequality**

   w.r.t. $\{\theta_1, \ldots, \theta_m\}$is i.i.d, $\mathbb{E}[\theta_i] = \mu$, $\mathcal{P}[a \leq \theta_i \leq b] = 1$, $\epsilon > 0$

   then $\mathbb{P}[|\frac{1}{m}\sum \theta_i - \mu| > \epsilon] \leq 2exp(-2m\epsilon^2/(b-a)^2)$

   proof:

   Step 1: **MARKOV'S INEQUALITY**

   $\mathbb{P}[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}$

   $\because Z \geq 0, \int_{x=0}^{\infty} \mathbb{P}[Z \geq x]dx = \int_0^{\infty}\int_x^{\infty} p(z)dzdx = \int_0^{\infty}\int_0^{z} p(z)dxdz = \int_0^{\infty} zp(z)dz = \mathbb{E}[Z]$

   $\therefore \forall a \geq 0, \mathbb{E}[Z] \geq \int_{x=0}^{a} \mathbb{P}[Z \geq x]dx \geq \int_{x=0}^{a} \mathbb{P}[Z \geq a]dx \geq a\mathbb{P}[Z \geq a]$

   Step 2: $\mathbb{E}[e^{\lambda X}] \leq e^{\frac{\lambda^2(b-a)^2}{8}}, s.t. \mathbb{E}(X) = 0$

   $\because e^{\lambda x} \leq \frac{b-x}{b-a}e^{\lambda a} + \frac{x-a}{b-a}e^{\lambda b}$

   $\therefore \mathbb{E}(e^{\lambda x}) \leq \frac{b-\mathbb{E}(x)}{b-a}e^{\lambda a} + \frac{\mathbb{E}(x)-a}{b-a}e^{\lambda b} = e^{L(h)}, L(h) = -hp + log(1-p+pe^h), h = \lambda(b-a), p = \frac{-a}{b-a}$

   $\because L(0) = L'(0) = 0, L''(h) \leq 1/4, \therefore L(h) \leq \frac{h^2}{8}$

   Step 3:

   $X_i = Z_i - \mathbb{E}[Z_i], X = \frac{1}{m}\sum X_i$

   $\mathbb{P}[X \geq \epsilon] = \mathbb{P}[e^{\lambda X} \geq e^{\lambda\epsilon}] \leq e^{-\lambda\epsilon}\mathbb{E}(e^{\lambda X})$

   $\mathbb{E}[e^{\lambda X}] = \mathbb{E}[\prod e^{\lambda X_i/m}] = \prod \mathbb{E}[e^{\lambda X_i/m}] \leq \prod e^{\frac{\lambda^2(b-a)^2}{8m^2}} = e^{\frac{\lambda^2(b-a)^2}{8m}}$

   $\mathbb{P}[X \geq \epsilon] \leq e^{-\lambda\epsilon + \frac{\lambda^2(b-a)^2}{8m}} \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$

   Similarly, $\mathbb{P}[X \leq -\epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$

   Then, $\mathbb{P}[|X| \leq \epsilon] \leq e^{-\frac{2m\epsilon^2}{(b-a)^2}}$

2. **proof**:

   $\because l(h, z_i) \in [0,1], L_S(h) = \frac{1}{m}\sum l(h, z_i)$ and $L_{\mathcal{D}} = \mathbb{E}(l(h, z_i))$

   $\therefore \mathcal{D}^m(S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon) = \mathbb{P}[|\frac{1}{m}\sum \theta_i - \mu| > \epsilon] \leq 2exp(-2m\epsilon^2)$

   $\therefore as for\ \mathcal{D}^m(\{S : \exists h \in \mathcal{H}, |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon\}) \leq \sum\limits_{h\in\mathcal{H}} \mathcal{D}^m(S : |L_S(h) - L_{\mathcal{D}}(h)| > \epsilon)$

   $\leq \sum\limits_{h\in\mathcal{H}} 2exp(-2m\epsilon^2) \leq 2|\mathcal{H}|exp(-2m\epsilon^2)$

$\therefore we\ choose\ m \geq \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2}, m_{\mathcal{H}}^{UC}(\epsilon,\delta) = \lceil \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \rceil$

$\therefore m_{\mathcal{H}}(\epsilon,\delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2,\delta) \leq \lceil \frac{2log(2|\mathcal{H}|/\delta)}{\epsilon^2} \rceil$

3. **The "Discretization Trick" in infinite size hypothesis classes**

   $h_\theta = sign(x-\theta), \mathcal{X} \to \{-1,1\}$

   64 bits floating point number, d parameters $\to |\mathcal{H}| \leq 2^{64d}$

   sample complexity is $\frac{128d+2log(2/\delta)}{\epsilon^2}$

5 The Bias-Complexity Tradeoff

1. In ch2, the misleading training data can cause overfitting, so we need a hypothesis class to reflect some prior knowledge.In ch5, we elaborate on learning tasks without prior knowledge.
2. **No-Free-Lunch theorem**
3. Decompose the error of an ERM algorithm into **approximation error** and **estimation error**

5.1 THE NO-FREE-LUNCH THEOREM

1. **NO universal learner**

   Let $\mathcal{X}$ be an infinite domain set and let $\mathcal{H}$ be the set of all functions from $\mathcal{X}$ to $\{0,1\}$.Then, $\mathcal{H}$ is not PAC learnable.

2. **No-Free-Lunch**

   Let **A** be any learning algorithm for the task of **binary classification**,

   w.r.t. $l_{0-1}(h,(x,y)) = 1\{h(x) \neq y\}$ over domain $\mathcal{X}$, $m \leq |\mathcal{X}|/2$, then

   - $\exists f : \mathcal{X} \to \{0,1\}, L_{\mathcal{D}}(f) = 0$
   - $\mathbb{P}(L_{\mathcal{D}}(A(S)) \geq 1/8) \geq 1/7$

3. **proof**(construct a special example)

   1. **construction**

      - Let new real data set $C \subseteq \mathcal{X}, |C| = 2m$, so functions set is $\{f_1,\dots,f_T\}, T = 2^{2m}$.
      - Let $\mathcal{D}_i$ be $\mathcal{D}_i(\{(x,y)\}) = 1/|C|, (x,y) \in (C, f_i(C))$.
      - Let S be sample data set,$|S| = m$, so sequences set is $\{S_1,\dots,S_k\}, k = (2m)^m$, and $S_j = (x_1,\dots,x_m), S_j^i = ((x_1,f_i(x_1)),\dots,(x_m,f_i(x_m)))$.
      - Test Set. Let $v_1,\dots,v_p$ be the examples in C that do not appear in S_j, so $p \geq m$.

   2. **key step1** : $\forall A, \underset{i\in[T]}{max}\ \underset{S\sim\mathcal{D}_i^m}{\mathbb{E}}[L_{\mathcal{D}_i}(A(S^i))] \geq 1/4$

      1. $\underset{i\in[T]}{max}\ \underset{S\sim\mathcal{D}_i^m}{\mathbb{E}}[L_{\mathcal{D}_i}(A(S^i))] = \underset{i\in[T]}{max}\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{k}\sum_{j=1}^{k}L_{\mathcal{D}_i}(A(S_j^i))$

         $= \frac{1}{k}\sum_{j=1}^{k}\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i)) \geq \underset{j\in[k]}{min}\frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i))$

      2. $\because L_{\mathcal{D}_i}(h) = \frac{1}{2m}\sum_{x\in C}1\{h(x) \neq f_i(x)\} \geq \frac{1}{2m}\sum_{r=1}^{p}1\{h(v_r) \neq f_i(v_r)\} \geq \frac{1}{2p}\sum_{r=1}^{p}1\{h(v_r) \neq f(v_r)\}$

         $\therefore \forall j, \frac{1}{T}\sum_{i=1}^{T}L_{\mathcal{D}_i}(A(S_j^i)) \geq \frac{1}{T}\sum_{i=1}^{T}\frac{1}{2p}\sum_{r=1}^{p}1\{A(S_j^i)(v_r) \neq f(v_r)\}$

         $= \frac{1}{2p}\sum_{r=1}^{p}\frac{1}{T}\sum_{i=1}^{T}1\{A(S_j^i)(v_r) \neq f(v_r)\} \geq \frac{1}{2}\underset{r\in p}{min}\frac{1}{T}\sum_{i=1}^{T}1\{A(S_j^i)(v_r) \neq f(v_r)\}$

3. $\because \forall r \in [p]$, every f_i has a dual function $f_{i'}$: $\forall c \in C, f_i(c) \neq f_{i'}(c), iff\ c = v_r$ (if and only if)

   $\therefore S_j^i = S_j^{i'}$ and $1\{A(S_j^i)(v_r) \neq f_i(v_r)\} + 1\{A(S_j^{i'}(v_r) \neq f_{i'}(v_r)\} = 1$

   $\therefore \frac{1}{T}\sum_{i=1}^{T} 1\{A(S_j^i)(v_r) \neq f(v_r)\} = \frac{1}{2}$

3. **key step2**

   1. **MARKOV'S INEQUALITY**

      $\mathbb{P}[Z \geq a] \leq \frac{\mathbb{E}[Z]}{a}$

      proof :

      $\because Z \geq 0, \int_{x=0}^{\infty} \mathbb{P}[Z \geq x]dx = \int_0^{\infty}\int_x^{\infty} p(z)dzdx = \int_0^{\infty}\int_0^{z} p(z)dxdz = \int_0^{\infty} zp(z)dz = \mathbb{E}[Z]$

      $\therefore \forall a \geq 0, \mathbb{E}[Z] \geq \int_{x=0}^{a} \mathbb{P}[Z \geq x]dx \geq \int_{x=0}^{a} \mathbb{P}[Z \geq a]dx \geq a\mathbb{P}[Z \geq a]$

   2. **Lemma** : If $Z \in [0,1]\ \mathbb{E}[Z] = \mu$, then

      $\forall a \in (0,1), \mathbb{P}[Z > 1-a] = 1 - \mathbb{P}[Z \leq 1-a] = 1 - \mathbb{P}[1-a \geq Z] \geq 1 - \frac{\mathbb{E}[1-Z]}{a} = 1 - \frac{1-\mu}{a}$

   3. $\because \forall A, \underset{i\in[T]}{max} \underset{S\sim\mathcal{D}_i^m}{\mathbb{E}}[L_{\mathcal{D}_i}(A(S^i))] \geq 1/4$

      $\therefore \exists \mathcal{D}, f, \underset{S\sim\mathcal{D}^m}{\mathbb{E}}[L_{\mathcal{D}}(A(S))] \geq 1/4$

      $\therefore \exists \mathcal{D}, f, \mathcal{P}[L_{\mathcal{D}}(A(S)) > \frac{1}{8}] \geq \frac{1}{7}$

5.2 ERROR DECOMPOSITION

1. We decomposition the error of an $ERM_{\mathcal{H}}$ predictor into:

   $L_{\mathcal{D}}(h_S) = \epsilon_{app} + \epsilon_{est}\ where: \epsilon_{app} = \underset{h\in\mathcal{H}}{min} L_{\mathcal{D}}(h), \epsilon_{est} = L_{\mathcal{D}}(h_S) - \epsilon(app)$

   - **The Approximation Error** : determined by the hypothesis class chosen.

     Under the realizability assumption, the approximation error is zero, while can be large in the agnostic case.

   - **The Estimation Error** : comes from empirical risk.

     The quality of this estimation depends on the training set size and on the size(or complexity) of the hypothesis class.

2. **bias-complexity tradeoff**

   - $|\mathcal{H}| \uparrow, \epsilon_{app} \downarrow, \epsilon_{est} \uparrow$, overfitting
   - $|\mathcal{H}| \downarrow, \epsilon_{app} \uparrow, \epsilon_{est} \downarrow$, underfitting

6 The VC-Dimension

- Goal : figure out which classes $\mathcal{H}$ are PAC learnable, and to characterize exactly the sample complexity of learning a given hypothesis class.
- Vladimir Vapnik and Alexey Chervonenkis discovered the Vapnik-Chervonenkis Dimension

6.1 INFINITE-SIZE CLASSES CAN BE LEARNABLE

6.2 THE VC-DIMENSION

1. **Restriction of $\mathcal{H}$ to C**

$\mathcal{H}_C = \{(h(c_1), \dots, h(c_m)) : h \in \mathcal{H}\} w.r.t. \mathcal{H} = \{h : \mathcal{X} \to \{0,1\}\}, C = \{c_1, \dots, c_m\} \subset \mathcal{X}$

2. **Shattering** , $\mathcal{H}$ **shatters** C : If $\mathcal{H}_C$ is the set of all functions from C to $\{0,1\}$, that is $|\mathcal{H}_C| = 2^{|C|}$

3. Consider No-Free-Lunch

   If $C \subset \mathcal{X}, |C| = 2m$, and C is shattered by $\mathcal{H}$, then $\forall A, \mathop{\mathbb{P}}_{S \sim \mathcal{D}^m}(L_{\mathcal{D}}(A(S)) \geq 1/8) \geq 1/7$

   If someone can explain every phenomenon, his explanations are worthless.

4. **VC-dimension**

   $VCdim(\mathcal{H}) = max\{|C| : C \subset \mathcal{X}, 2^{|C|} = |\mathcal{H}_C|\}$

   - $VCdim(\mathcal{H}) = \infty \Rightarrow \mathcal{H}$ is not PAC learnable.
   - $VCdim(\mathcal{H}) = d < \infty \Rightarrow \mathcal{H}$ is PAC learnable.

6.3 EXAMPLES

1. To show that $VCdim(\mathcal{H}) = d$ we need to show that:
   - There exists a set C of size d that is shattered by $\mathcal{H}$.
   - Every set C of size $d+1$ is not shattered by $\mathcal{H}$.
2. **Threshold Functions** : $\mathcal{H} = \{h_a : a \in \mathbb{R}, h_a = 1\{x < a\}\}$, $VCdim(\mathcal{H}) = 1$, recall **example 6.2** .
3. **Intervals** : $\mathcal{H} = \{h_{a,b} : a, b \in \mathbb{R}, a < b, h_{a,b} = 1\{x \in (a,b)\}\}$, $VCdim(\mathcal{H}) = 2$.
4. **Axis Aligned Rectangles** :

   $\mathcal{H} = \{h_{a_1,a_2,b_1,b_2} : a_1 < a_2, b_1 < b_2, h_a = 1\{a_1 \leq x \leq a_2, b_1 \leq y \leq b_2\}\}, VCdim(\mathcal{H}) = 4.$
5. **Finite Classes** : $|\mathcal{H}_C| \leq |\mathcal{H}|, 2^{|C|} = |\mathcal{H}_C| \Rightarrow VCdim(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$
6. **VC-Dimension and the Number of Parameters**

   The VC-Dimension often equals to the number of parameters, but it's not always true.

   For example, $\mathcal{H} = \{h_\theta : \theta \in \mathbb{R}, h_\theta(x) = \lceil 0.5 sin(\theta x) \rceil\}, VCdim(\mathcal{H}) = \infty$.

   proof :

   1. If $x \in (0,1)$ and its binary expansion is $0.x_1x_2x_3\dots$, then $\forall m, \lceil 0.5 sin(2^m \pi x) \rceil = (1 - x_m)$, provided that $\exists k > m, s.t. \ x_k = 1$.
   2. $C = \{2^1\pi, 2^2\pi, \dots, 2^d\pi\}$ is shattered by $\mathcal{H}$.

6.4 THE FUNDAMENTAL THEOREM OF PAC LEARNING

1. **The Fundamental Theorem of Statistical Learning**

   $w.r.t \ \mathcal{H} = \{h : h : \mathcal{X} \to \{0,1\}\}$, the loss function is 0-1 loss. Then the following are equivalent:

   1. $\mathcal{H}$ has the uniform convergence propperty.
   2. Any ERM rule is a successful agnostic PAC learning.
   3. $\mathcal{H}$ is agnostic PAC learnable.
   4. $\mathcal{H}$ is PAC learnable.
   5. Any ERM rule is a successful PAC learner for $\mathcal{H}$.
   6. $\mathcal{H}$ has a finite VC-dimension.

   Proof is given in **6.5**.

2. **The Fundamental Theorem of Statistical Learning - Quantitative Version**

   $w.r.t \ \mathcal{H} = \{h : h : \mathcal{X} \to \{0,1\}\}$, the loss function is 0-1 loss. If $VCdim(\mathcal{H}) = d < \infty$. Then there are absolute constants C_1, C_2 such that

   1. $\mathcal{H}$ has the uniform convergence property with sample complexity:

$$C_1 \frac{d + log(1/\delta)}{\epsilon^2} \leq m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq C_2 \frac{d + log(1/\delta)}{\epsilon^2}$$

2. $\mathcal{H}$ is agnostic PAC learnable with sample complexity

$$C_1 \frac{d + log(1/\delta)}{\epsilon^2} \leq m_H(\epsilon, \delta) \leq C_2 \frac{d + log(1/\delta)}{\epsilon^2}$$

3. $\mathcal{H}$ is PAC learnable with sample complexity

$$C_1 \frac{d + log(1/\delta)}{\epsilon} \leq m_{\mathcal{H}}(\epsilon, \delta) \leq C_2 \frac{dlog(1/\epsilon) + log(1/\delta)}{\epsilon}$$

Proof is given in **Chapter 28**.

3. **Remark**
   - The fundamental theorem holds for some other learning problems such as regression with absolute loss or the squared loss.
   - It does not hold for all learning tasks.
   - Learnable even though without the uniform convergence property.
   - In some situations, the ERM rule fails but learnability is possible with other learning rules.

6.5 PROOF OF THEOREM 6.7 (In this note 6.4.1)

$1 \to 2$ in Ch.4. $2 \to 3, 3 \to 4$ are trivial $\Rightarrow 2 \to 5$. **No-Free-Lunch** $\Rightarrow 4 \to 6, 5 \to 6$. So the difficult part is to $6 \to 1$.

The proof is based on two main claims:

- If $VCdim(\mathcal{H}) = d$, $|\mathcal{H}_C| \sim O(|C|^d)$
- The uniform convergence holds whenever$|\mathcal{H}_C| \sim O(|C|^d)$

1. **Sauer's Lemma and the Growth Function**
   1. **Growth Function** : $r_{\mathcal{H}}(m) = \underset{C \subset \mathcal{X}: |C| = m}{max} |\mathcal{H}_C|$
   2. **Lemma Sauer-Shelah-Perles**

      If $VCdim(\mathcal{H}) \leq d < \infty$, then$\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \mathbb{C}_m^i$. In particular, if $m > d + 1$ then $\tau_{\mathcal{H}}(m) \leq (em/d)^d$(see Lemma A.5 in Appendix A).

      proof $\tau_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} C_m^i$ inductive argument
      1. We proof a stronger claim :

         $\forall C = \{c_1, \ldots, c_m\}, \forall \mathcal{H}, |\mathcal{H}_C| \leq |\{B \subseteq C : \mathcal{H} \; shatters \; B\}| \leq \sum_{i=0}^{d} \mathbb{C}_m^i$ (这里的C_m^i是排列组合数).
      2. **m=1**: $\mathcal{H}_C = \{0\}, \{1\}, or \; \{\{0\}, \{1\}\}$, $B = \emptyset \; or \; c_1$, the left equation always holds.
      3. **If the left equation holds when** $k < m$:
         1. **define**

$$\begin{aligned} C &= \{c_1, \ldots, c_m\}, C' = \{c_2, \ldots, c_m\} \\ Y_0 &= \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \vee (1, y_2, \ldots, y_m) \in \mathcal{H}_C\} \\ Y_1 &= \{(y_2, \ldots, y_m) : (0, y_2, \ldots, y_m) \wedge (1, y_2, \ldots, y_m) \in \mathcal{H}_C\} \\ \mathcal{H}' &= \{h \in \mathcal{H} : \exists h' \in \mathcal{H} s.t. (1 - h'(c_1), h'(c_2), \ldots, h'(c_m)) \\ &\quad = (h(c_1), h(c_2), \ldots, h(c_m))\} \end{aligned}$$

         2. It is easy to verify that $|\mathcal{H}_C| = |Y_0| + |Y_1|$ (P.S. 韦恩图，重点是理解这三个集合的含义。)

3. $|Y_0| = |\mathcal{H}_{C'}| \leq |\{B \subseteq C' : \mathcal{H} shatters B\}| = |\{B \subseteq C' : c_1 \notin B \wedge \mathcal{H} shatters B\}|$
4. $|Y_1| = |\mathcal{H}'_{C'}| \leq |\{B \subseteq C' : \mathcal{H}' shatters B\}| = |\{B \subseteq C' : \mathcal{H}' shatters B \cup \{c_1\}\}|$
   $= |\{B \subseteq C : c_1 \in B \wedge \mathcal{H}' shatters B\}| \leq |\{B \subseteq C : c_1 \in B \wedge \mathcal{H} shatters B\}|$
5. $|\mathcal{H}_C| = |Y_0| + |Y_1| \leq |\{B \subseteq C : \mathcal{H} shatters B\}|$

2. **Uniform Convergence for Classes of Small Effective Size**

   1. **Theorem** : $w.r.t.\, \mathcal{H}, \tau_{\mathcal{H}}$

      $\underset{S\sim\mathcal{D}^m}{\mathbb{E}}[sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)|] \leq \frac{4+\sqrt{log(\tau_{\mathcal{H}}(2m))}}{\sqrt{2m}}$, by using **MARKOV'S INEQUALITY,** we can get

      $\underset{S\sim\mathcal{D}^m}{\mathbb{P}}[sup_{h\in\mathcal{H}}|L_{\mathcal{D}}(h) - L_S(h)| \geq \frac{4+\sqrt{log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}}] \leq \delta$

   2. **proof** $6 \to 1$: we will prove that $m^{UC}_{\mathcal{H}}(\epsilon, \delta) \leq 4\frac{16d}{(\delta\epsilon)^2} log(\frac{16d}{(\delta\epsilon)^2}) + \frac{16d\log(2e/d)}{(\delta\epsilon)^2}$

      1. $m > d, \tau_{\mathcal{H}}(2m) \leq (2em/d)^d, assume \sqrt{d\log(2em/d)} \geq 4, then$

         $|L_{\mathcal{D}}(h) - L_S(h)| \leq \frac{4+\sqrt{log(\tau_{\mathcal{H}}(2m))}}{\delta\sqrt{2m}} \leq \frac{1}{\delta}\sqrt{\frac{2d\log(2em/d)}{m}} \leq \epsilon$

         $m \geq \frac{2dlog(m)}{(\delta\epsilon)^2} + \frac{2dlog(2e/d)}{(\delta\epsilon)^2} \Leftarrow m \geq 4\frac{2d}{(\delta\epsilon)^2}\log(\frac{2d}{(\delta\epsilon)^2}) + \frac{4d\log(2e/d)}{(\delta\epsilon)^2}$

      2. To proof pre-equation, we can proof:

         - Let $a > 0$. Then: $x \geq 2a\log(a) \Rightarrow x \geq a\log(x)$.

           (分类讨论$a \in (0, \sqrt{e}], a \in (\sqrt{e}, \infty)$,considering $a - 2alog(a)$)
         - Let $a \geq 1$ and $b > 0$. Then : $x \geq 4a\log(2a) + 2b \Rightarrow x \geq a\log(x) + b$ (Firstly, we get $x \geq 2b$. Then $x \geq 4alog(2a)$)

7 Nonuniform Learnability

7.1 NONUNIFORM LEARNABILITY

1. h is (ϵ, δ)-competitive with another hypothesis h' if $\mathbb{P}\{L_{\mathcal{D}}(h) \leq L_{\mathcal{D}}(h') + \epsilon\} \geq 1 - \delta$
2. **nonuniformly learnable** :

   $\exists A, m^{NUL}_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \to \mathbb{N}, \forall \epsilon, \delta \in (0,1), \forall h \in \mathcal{H}$:
   $\mathcal{D}^m\{S : L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon, |S| > m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h)\} \geq 1 - \delta$
3. The difference between aPAC and NL is the question of whether the sample size m may depend on h.
4. NL is a relaxation of aPAC.
5. **theorem** A hypothesis class $\mathcal{H}$ of binary classifiers is nonuniformly learnable if and only if it is a countable union of agnostic PAC learnable hypothesis classes.

   proof

   necessity: use following theorem;

   sufficiency: let $\mathcal{H}_n = \{h \in \mathcal{H} : m^{NUL}_{\mathcal{H}}(1/8, 1/7, h) \leq n\}$. Then $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, using the fundamental of statistical learning, $VC(\mathcal{H}_n) < \infty$, and therefore $\mathcal{H}_n$ is agnostic PAC learnable.(If $VC(H_n) = \infty$, then do not exist $m^{NUL}_{\mathcal{H}}(1/8, 1/7, h) \leq n$)
6. **theorem** Let $\mathcal{H}$ be a countable union of hypothesis class $\mathcal{H} = \cup_{n\in\mathcal{N}}\mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys the uniform convergence property. Then, $\mathcal{H}$ is nonuniformly learnable. (The proof will be given in the next section)
7. Nonuniform learnability is a strict relaxation of agnostic PAC learnability.

7.2 STRUCTURAL RISK MINIMIZATION

1. **denote** $\epsilon_n(m,\epsilon) = min\{\epsilon \in (0,1) : m^{UC}_{\mathcal{H}_n}(\epsilon,\delta) \leq m\}$
2. **weight function** : $\omega : \mathbb{N} \to [0,1], \sum_{n=1}^{\infty} \omega(n) \leq 1$
3. **theorem** : $\mathcal{H} = \cup \mathcal{H}_n, \mathcal{H}_n \ has \ m^{UC}_{\mathcal{H}_n}. \forall \delta, \mathcal{D}, n, h$

   $\mathcal{D}^m\{S : |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, \omega(n)\cdot\delta)\} \geq 1-\delta$

   proof :

   $\forall h \in \mathcal{H}_n, |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m,\delta_n)$

   $\forall h \in \mathcal{H}, \mathcal{D}^m\{S : |L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(m, \omega(n)\cdot\delta)\} \geq 1 - \sum \delta_n \geq 1-\delta$
4. **denote** $n(h) = min\{n : h \in \mathcal{H}_n\}$
5. $\mathcal{D}^m[L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot h)]ge1-\delta$, less constraints, higher probability.
6. **Structural Risk Minimizaiton(SRM)** :
   - **prior knowledge** : $\mathcal{H} = \cup_n \mathcal{H}_n, \mathcal{H}_n \ has \ m^{UC}_{\mathcal{H}_n}, \sum \omega(n) \leq 1$
   - **input** : training set $S \sim \mathcal{D}^m$, confidence δ
   - **output** : $h \in argmin_{h\in\mathcal{H}}[L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot\delta)]$
7. **theorem** $\omega(n) = \frac{6}{n^2\pi^2}$, $m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) \leq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, \frac{6\delta}{(\pi n(h))^2})$

   proof :

   $\mathbb{P}\{L_{\mathcal{D}}(h) \leq L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot h)\} \geq 1-\delta$

   if $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon/2, \omega(n(h))\delta)$, then $\epsilon_{n(h)}(m, \omega(n(h))\cdot h) \leq \epsilon/2$

   Uniform convergence $L_S(h) \leq L_{\mathcal{D}}(h) + \frac{\epsilon}{2}$

   $L_{\mathcal{D}}(A(S)) \leq L_S(h) + \epsilon_{n(h)}(m, \omega(n(h))\cdot h) \leq L_{\mathcal{D}}(h) + \epsilon$
8. *No-Free-Lunch-for-Nonuniform-Learnability*

   $\forall\{\mathcal{X}, |\mathcal{X}| = \infty\}$, the class of all binary valued functions over $\mathcal{X}$ is not a countable union of classes of finite VC-dimension.(Exercise 7.5)

   proof : If $\mathcal{H}$ shatters an infinite set . Then, for any sequence of classes $\mathcal{H}_n : n \in \mathbb{N}$ such that $\mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, there exists some n for which $VCdim(\mathcal{H}_n) = \infty$.

   1. Assume $\exists\{\mathcal{H}_n\}, \mathcal{H} = \cup_{n\in\mathbb{N}}\mathcal{H}_n$, and $\forall\mathcal{H}_n, VCdim(\mathcal{H}_n) < \infty$.
   2. Subproblem: For $K \subseteq \mathcal{X}, |K| = \infty$ we can always construct subsets sequence $\{K_n\}, K_n \subseteq K; \forall K_n, |K_n| > VCdim(\mathcal{H}_n); \forall n \neq m, K_n \cap K_m = \emptyset$.

      subproof :

      First, let $K_1 \subseteq K, |K_1| = VDdim(\mathcal{H}_1) + 1$.

      Second, suppose that $K_1, \ldots, K_{r-1}$ has been chosen, because $|K| = \infty$, we always can choose $K_r \subseteq K \backslash (\cup_{i=1}^{r-1} K_i)$ such that $|K_r| = VCdim(\mathcal{H}_r) + 1$.
   3. The subproblem implies that $\forall n \in \mathbb{N}, \exists f_n \notin \mathcal{H}_n$.
   4. $\exists f = [f_1, f_2, \ldots, f_n], f \in \mathcal{H}, but \ \forall n : f_n \notin \mathcal{H}_n$, which makes a contradiction.
9. $\forall\{\mathcal{X}, |\mathcal{X}| = \infty\}$, there exists no nonuniform learner w.r.t. the class of all deterministic binary classifiers.
10. Assume $VCdim(\mathcal{H}_n) = n$, then $m^{UC}_{\mathcal{H}_n}(\epsilon,\delta) = C\frac{n+log(1/\delta)}{\epsilon^2}$ (Ch6)

    If $\omega(n) = \frac{6}{n^2\pi^2}$, then $m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h) - m^{UC}_{\mathcal{H}_n}(\epsilon/2,\delta) \leq 4C\frac{2log(2n)}{\epsilon^2}$

    The gap between $m^{NUL}_{\mathcal{H}}$ and $m^{UC}_{\mathcal{H}_n}$ increases with the index of the class, which reflecting the value of knowing a good priority order on the hypotheses in $\mathcal{H}$.

7.3 MINIMUM DESCRIPTION LENGTH AND OCCAM'S RAZOR

1. Let $\mathcal{H}$ be a countable hypothesis class. Then $\mathcal{H}$ can be rewrited as $\mathcal{H} = \cup_{n\in\mathbb{N}}\{h_n\}$, each singleton classes has the uniform convergence property with rate $m^{UC}(\epsilon,\delta) = \frac{log(2/\delta)}{2\epsilon^2}$, and $\epsilon_n(m,\delta) = \sqrt{\frac{log(2/\delta)}{2m}}$. SRM rule becomes:
   - $argmin_{h_n\in\mathcal{H}}[L_S(h) + \sqrt{\frac{-log(\omega(n))+log(2/\delta)}{2m}}]$
   - $argmin_{h_n\in\mathcal{H}}[L_S(h) + \sqrt{\frac{-log(\omega(h))+log(2/\delta)}{2m}}]$
2. **the description of h** : Fix some finite set Σ of symbols, the description function $d = \mathcal{H} \to \Sigma^* \subseteq \Sigma$, its length is denoted by $|h|$.
   - σ is always used to represent $d(h)$
   - prefix-free
3. **Kraft Inequality** : If $S \subseteq \{0,1\}^*$ is a prefix-free set of strings, then $\sum\limits_{\sigma\in S}\frac{1}{2^{|\sigma|}} \leq 1$
4. $\omega(h) = \frac{1}{2^{|h|}}$
5. **theorem** : $\mathcal{H}$, prefix-free description language $d : \mathcal{H} \to \{0,1\}^*$, then

   $\mathcal{D}^m\{\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{|h|+ln(2/\delta)}{2m}}\} \geq 1-\delta$
6. **Minimum Description Length (MDL)**
   - **prior knowledge** :
     - $\mathcal{H}$ is a countable bypothesis class
     - $\mathcal{H}$ is described by a prefix-free language over $\{0,1\}$
     - For every $h \in \mathcal{H}$, $|h|$ is the length of the representation of h
   - **input** : A training set $S \sim D^m$, confidence δ
   - **output** : $h \in argmin_{h\in\mathcal{H}}[L_S(h) + \sqrt{\frac{|h|+ln(2/\delta)}{2m}}]$
7. Pre theorem conveys a philosophical message : A short explanation (that is, a hypothesis that has a short length) tends to be more valid than a long explanation.
8. The more complex a hypothesis h is , the larger the sample size it has to fit to guarantee that it has a small true risk $L_{\mathcal{D}}(h)$.
9. Choosing a description language (or, equivalently, some weighting of hypotheses) is a weak form of committing to a hypothesis.
10. Rather than committing to a single hypothesis, we spread out our commitment among many.
11. As long as it is done independently of the training sample, our generalization bound holds.
12. Just as the choice of a single hypothesis to be evaluated by a sample can be arbitrary, so is the choice of description language.

7.4 OTHER NOTIONS OF LEARNABLITY-CONSISTENCY

1. Weak consistency : convergence in probablity
2. Strong consistency: sure convergence
3. **Definition** (Consistency) : A learning rule A is consistent w.r.t. $\mathcal{H}$ and $\mathcal{P}$

   domain set Z, probability distributions set $\mathcal{P}$,

   $\exists m^{CON}_{\mathcal{H}} : (0,1)^2 \times \mathcal{H} \times \mathcal{P} \to \mathbb{N}$ such that, $\forall\epsilon,\delta \in (0,1), \forall h \in \mathcal{H}, \forall\mathcal{D} \in \mathcal{P}$, if $m \geq m^{NUL}_{\mathcal{H}}(\epsilon,\delta,h,\mathcal{D})$ then $\mathcal{D}^m\{S|L_{\mathcal{D}}(A(S)) \leq L_{\mathcal{D}}(h) + \epsilon\} \geq 1-\delta$
4. If $\mathcal{P}$ is the set of all distributions, then A is universally consistent w.r.t. $\mathcal{H}$.

5. **Memory algorithm** : memorize the training examples, and, given a test point x, it predicts the majority label among all labeled instances of x that exist in the training sample.

   Not nonuniformly learnable, but universally consistent for every countable domain $\mathcal{X}$ and a finite label set $\mathcal{Y}$. (exercise 7.6)

   proof :

   1. Let $\{x_i : i \in \mathbb{N}\}$ be an enumeration of the elements of $\mathcal{X}$, and $i \leq j \Rightarrow \mathcal{D}(x_i) \geq \mathcal{D}(x_j)$.
   2. It's easy to verify $\lim_{n\to\infty} \sum_{i\geq n}^{\infty} \mathcal{D}(x_i) = 0$($S_n \to 1 \Rightarrow S - S_n \to 0$).
   3. $\forall \eta > 0, \exists N \in \mathbb{N}$ such that $\forall i > N, \mathcal{D}(\{x_i\}) < \eta$, then

      $$\mathbb{P}_{S\sim\mathcal{D}^m}[\exists x_i : \mathcal{D}(\{x_i\}) > \eta, x_i \notin S] \leq \sum_{i=1}^{N} \mathbb{P}[x_i \notin S] \leq N(1-\eta)^m \leq Ne^{-\eta m}$$

   4. $\forall \epsilon > 0, \exists N \in \mathbb{N}$ such that $\sum_{n\geq N} \mathcal{D}(\{x_n\}) < \epsilon$, which also means that $\forall n > N, \mathcal{D}(\{x_n\}) < \epsilon$.

      Let $\eta = \mathcal{D}(\{x_n\})$, then, $\forall k \in [N], \mathcal{D}(\{x_k\}) \geq \eta$.

      $\mathbb{P}_{S\sim\mathcal{D}^m}[\mathcal{D}(\{\exists x_i : x_i \notin S\}) > \epsilon] \leq \mathbb{P}_{S\sim\mathcal{D}^m}[\exists x_i \in [N] : x_i \notin S] \leq Ne^{-\eta m}$

7.5 DISCUSSING THE DIFFERENT NOTIONS OF LEARNABLITY

1. What Is The Risk of the Learned Hypothesis?
   - PAC learning and nonuniform learning gives us an upper bound on the true risk of the learned hypothesis based on its empirical risk.
   - Consistency guarantees do not provide such a bound, but estimate the risk of the output predictor using a validation set.(Ch11)
2. How Many Examples Are Required to Be as Good as the Best Hypothesis in $\mathcal{H}$?
   - PAC learning gives a crisp answer
   - nonuniform learning this number depends on the best hypothesis in $\mathcal{H}$
   - consistency it also depends on the underlying distribution
   - In this sense, PAC learning is the only useful definition of learnability
   - If $\mathcal{H}$ has a large approximation error, PAC's risk may still be large. This reflects the fact that the usefulness of PAC learning relies on the quality of our prior knowledge.
   - If PAC fails, we change the $\mathcal{H}$.
   - If nonuniform algorithm fails, we change a different weighting function.
3. How to Learn? How to Express Prior Knowledge?
   - The definition of PAC learning yields the limitation of learning(via the No-Free-Lunch theorem) and the necessity of prior knowledge.
     - Choose $\mathcal{H}$ by prior knowledge.
     - $ERM_{\mathcal{H}}$
   - nonuniform learnability
     - Encode prior knowledge by specifying weights over(subsets of) hypothesis of $\mathcal{H}$.
     - SRM (pays *estimation error* and *do not know the low bound of m*).
   - consistent algorithm
     - Does not yield a natural learning paradigm or a way to encode prior knowledge.
     - In fact, in many cases there is no need for prior knowledge at all.(*Memorize algorithm*).
     - Weak requirement
4. Which Learning Algorithm Should We Prefer?
   - w.r.t. the set of all functions from $\mathcal{X} \to \mathcal{Y}$, which gives us a guarantee that for enough training examples, we will always be as good as the Bayes Optimal predictor.

- problems:
    - the sample complexity of the consistent algorithm, non enough examples
    - it's not very hard to make any PAC or nonuniform learner consistent:

      Firstly, we run nonuniform learned predictor, obtain the bound on the true risk;

      Then, if the bound is small enough we are done, otherwise, we revert to Memorize algorithm.

5. The "contradiction" between "No-Free-Lunch" and "Memory algorithm"
    - **No-Free-Lunch** : Let $\mathcal{X}$ be a countable infinite domain and let $\mathcal{Y} = \{\pm 1\}$, then for $\forall A$, and a training set size, m, $\exists \mathcal{D}, h^* : \mathcal{X} \to \mathcal{Y}$, A is likely to return a classifier with a large error.
    - **Memorize algorithm** : $\forall \mathcal{D}, h^* : \mathcal{X} \to \mathcal{Y}, then\ \exists m$, memorize algorithm is likely to return a classifier with a small error.

8 The Runtime of Learning

Key words : samples, computational resources, sample complexity, computational complexity.

ERM implementation is computationally hard. It follows that hardness of implementing ERM does not imply hardness of learning.

8.1 COMPUTATIONAL COMPLEXITY OF LEARNING

1. Abstract machine, Turning machine : there exists a constant c such that any such "operation" can be performed on the machine using c seconds.
2. For problem "rectangles learning", we can fix ϵ,δ and varying the dimension, we also can fix the dimension δ and varying the ϵ.Then, we can analyze the asymptotic runtime as a function of variables of that sequence.
3. "Cheating" : training easy but predict hard.
4. **Definition 8.1** *The Computational Complexity of a Learning Algorithm*.
    - $f : (0,1)^2 \to \mathbb{N}$, a learning task$(Z, H, l)$, learning algorithm $\mathcal{A}$. $\forall \mathcal{D}, \epsilon, \delta$.We say $\mathcal{A}$ solves the learning task in time $O(f)$ for:
        - $\mathcal{A}$ terminates after performing at most $cf(\epsilon, \delta)$ operations
        - The output of $\mathcal{A}$, denoted $h_{\mathcal{A}}$, can predict the label of a new example while performing at most $cf(\epsilon, \delta)$ operations
        - $\mathcal{A}$ is PAC learnable.
    - Consider a sequence of learning problems, $(Z_n, \mathcal{H}_n, l_n)_{n=1}^{\infty}$,$\exists g : \mathbb{N} \times (0,1)^2 \to \mathbb{N}$, for $\forall n$, $\mathcal{A}$ solves the problem $(Z_n, \mathcal{H}_n, l_n)$ in time $O(g(n, \epsilon, \delta))$
5. $\mathcal{A}$ is efficient algorithm w.r.t. a sequence $(Z_n, \mathcal{H}_n, l_n)$ if its runtime is $O(p(n, 1/\epsilon, 1/\delta))$ for some polynomial p.

8.2 IMPLEMENTING THE ERM RULE

1. $ERM_{\mathcal{H}}$ rule
2. **Finite Classes** :
    1. $m_{\mathcal{H}}(\epsilon, \delta) = c \log(c|\mathcal{H}|/\delta)/\epsilon^c$, where $c = 1$ in the realizable case and $c = 2$ in the nonrealizable case.
    2. If $\mathcal{H}$ can be the set of all predictors that can be implemented by a C++ program written in at most 10000 bits of code, then $|\mathcal{H}| \leq 2^{10000}$ and the sample complexity is only $c(10000 + log(c/\delta))/\epsilon^c$.
    3. exhaustive search : the runtime of exhaustive search is $k|\mathcal{H}|/m$, not efficient.
3. **Axis Aligned Rectangles**
    1. $\mathcal{H}_n = \{h_{(a_1,\ldots,a_n,b_1,\ldots,b_n)} : \forall i, a_i \leq b_i\}$

$h_{(a_1,\ldots,a_n,b_1,\ldots,b_n)} = 1\{\forall i, x_i \in [a_i, b_i]\}$

Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$, then the total runtime is $O(mn)$

2. Solving the ERM problem for many common hypothesis classes in the agnostic setting is NP-hard.
3. There exist efficient learning algorithms on specific hypothesis classes.
4. If m is fixed, but n is not fixed in axis aligned rectangles, then the runtime of the procedure is $m^{O(n)}$.

4. **Boolean Conjunctions**
    1. $\mathcal{X} = \{0,1\}^n, \mathcal{Y} = \{0,1\}, h(x) = 1\{x_{i1} \wedge \cdots \wedge x_{ik} \wedge \neg x_{j1} \wedge \cdots \wedge \neg x_{jr}\}$
    2. The sample complexity of learning $\mathcal{H}_C^n$ is $O(d\log(3/\delta)/\epsilon)$
    3. Efficiently Learnable in the Realizable Case: $O(mn)$
    4. Agnostic Case : NP hard
5. **learning 3-Term DNF**
    1. 3-term disjunctive normal form formulae
    2. $h(x) = A_1(x) \vee A_2(x) \vee A_3(x)$, A_1, A_2, A_3 are former Boolean conjunctions.
    3. the sample complexity of learning $\mathcal{H}_{3DNF}^n$ is $O(3n\log(3/\delta)/\epsilon)$
    4. From the computational perspective, this learning problem is hard, unless $RP = NP$, there is no polynomial that *properly* learns 3-Term DNF. *Properly* means the algorithm outputs a hypothesis that is a 3-term DNF formula. (Execise 8.4)

        proof :

        1. **Definition 8.2** : The complexity class Randomized Polynominal(RP) time is the class of all decision problems (YES or NO problem) for which there exists a probabilistic algorithm with these properties:
            - On any input instance the algorithm runs in polynomial time in the input size.
            - If the correct answer is NO, the algorithm must return NO.
            - If the correct answer is YES, the algorithm returns YES with probability $a \geq 1/2$ and returns NO with probability $1 - a$.(The constant 1/2 can be replaced by any constant in $(0, 1)$)
        2. Clearly the class RP contains the class P. It is also knows that RP is contained in the class NP. It is not known whether any equality holds among these there complexity classes, but it is widely believed that NP is strictly larger than RP.
        3. Subproblem : If a class $\mathcal{H}$ is properly PAC learnable by a polynominal time algorithm, then the $ERM_{\mathcal{H}}$ problem is NP-hard, unless NP=RP, there exists no polynomial time proper PAC learning algorithm for $\mathcal{H}$.

            Subproof : For $S \in (\mathcal{X} \times \{\pm 1\}^m)$ with RA assumption, let $\delta = 0.3$ and $\epsilon = 1/|S|$, then $\mathbb{P}[L_{\mathcal{D}}(A(S)) < \epsilon] \geq 1 - \delta = 0.7 > 0.5$

8.3 EFFICIENTLY LEARNABLE, BUT NOT BY A PROPER ERM

1. $A_1 \vee A_2 \vee A_3 = \bigwedge_{u \in A_1, v \in A_2, w \in A_3} (u \vee v \vee w)$
2. The basic idea is to replace the original hypothesis class of 3-term DNF formula with a larger hypothesis class so that the new class is easily learnable.
3. *representation independent* : The learning algorithm might return a hypothesis that does not belong to the original hypothesis class.
4. $\psi : \{0,1\}^n \to \{0,1\}^{(2n)^3}$
5. 3-term-DNF formulae over $\{0,1\}^n \subset$ Conjunctions over $\{0,1\}^{(2n)^3}$

8.4 HARDNESS OF LEARNING

1. cryptographic assumptions :

   $f : \{0,1\}^n \to \{0,1\}^n$, polynomial $p(\cdot)$, $\forall A$ randomized polynomial time algorithm,
   $\mathbb{P}[f(A(f(x))) = f(x)] \leq \frac{1}{p(n)}$

2. *trapdoor one way function* : secret key

   for some polynomial function p, $\forall n, \exists s_n, |s_n| \leq p(n), \forall x, input(f(x), s_n)\ outputs\ x.$

3. Let F_n be a family of trapdoor functions over $\{0,1\}^n$, then $\mathcal{H}_F^n = \{f^{-1} : f \in F_n\}$ has no efficient learner for this class.

   (Kearns and Vazirani 1994, Ch6)