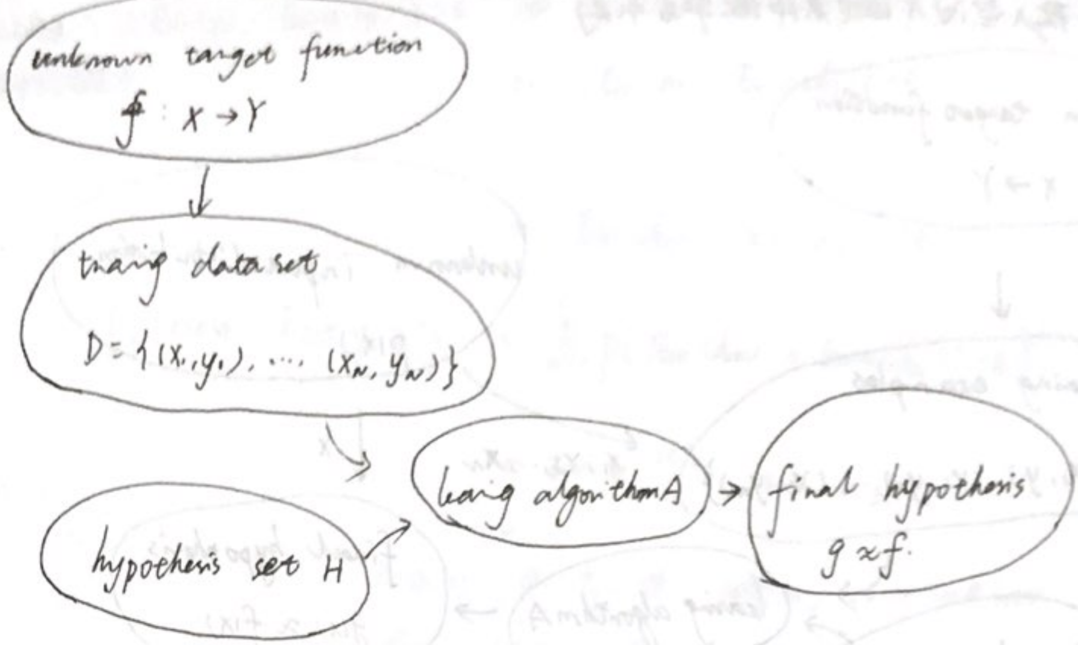
学习问题：leaing problem.



hypothesis set：假设目标函数是线性函数.
leaing algorithm：梯度下降. 牛顿法等.

学习是否可信：Is leaing feasible?

* How could a limited data set reveal enough information to pin down the entire target function?

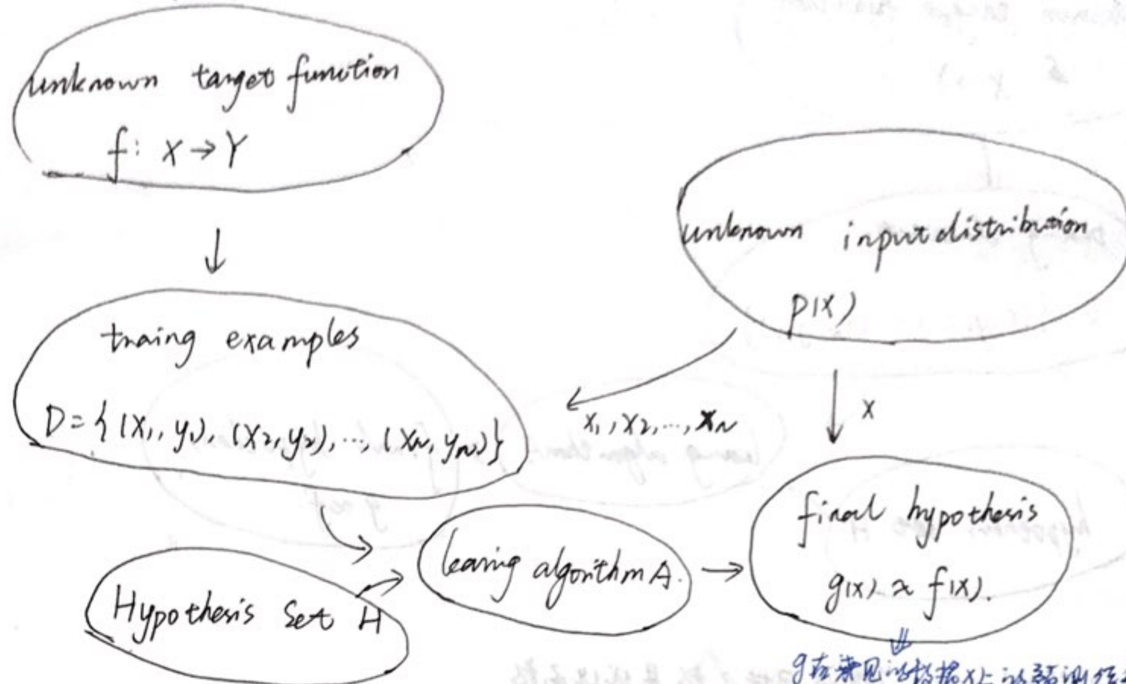
* learning is feasible ⇒ we can lean something from seen dataset, which we didn't known before.

↓

reveal in ~~probability~~ way

probabilistic

△假设输入空间X满足某种概率分布则.



g在未见过的数据X上的预测结果
尽可能接近目标函数f

Hoeffding Inequality

$$P[|\nu-\mu|<\varepsilon] \le 2e^{-2\varepsilon^2 N}, \quad \varepsilon>0.$$

tolerance → ε; # samples → N.

ν → hypothesis → in-sample error

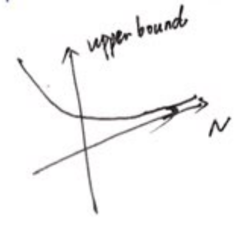
μ → target function → out-sample error.

使用某种方式来衡量g和f的接近程度.

$$\Rightarrow \begin{cases} P[|E_{in}(g)-E_{out}(g)|>\varepsilon] \le 2e^{-2\varepsilon^2 N}, \quad \varepsilon>0. \quad (1) \\ E_{in}(g)=\frac{1}{N}\sum_{n=1}^{N} \mathbb{I}[g(x_n) \neq f(x_n)] \quad \text{in-sample error} \\ E_{out}(g)=P[g(x) \neq f(x)] \quad \text{out-sample error.} \end{cases}$$

⇒ 为了压低上界，使得Ein和Eout尽可能接近，可以增大N.

⇒ 从e^{-x}的曲线来看，当N达到一定数量，增大N带来的好处很少.



⇓ 意味着.

我们不能选g, $\because |E_{in}(g) - E_{out}(g)| > \varepsilon \Rightarrow |E_{in}(h_1) - E_{out}(h_1)| > \varepsilon$

但是我们能选h.

$$\text{or } |E_{in}(h_2) - E_{out}(h_2)| > \varepsilon$$

$$\cdots$$

$$\text{or } |E_{in}(h_M) - E_{out}(h_M)| > \varepsilon$$

$$\therefore P[|E_{in}(g) - E_{out}(g)| > \varepsilon \le \sum_{m=1}^{M} P[|E_{in}(h_m) - E_{out}(h_m)| > \varepsilon]$$

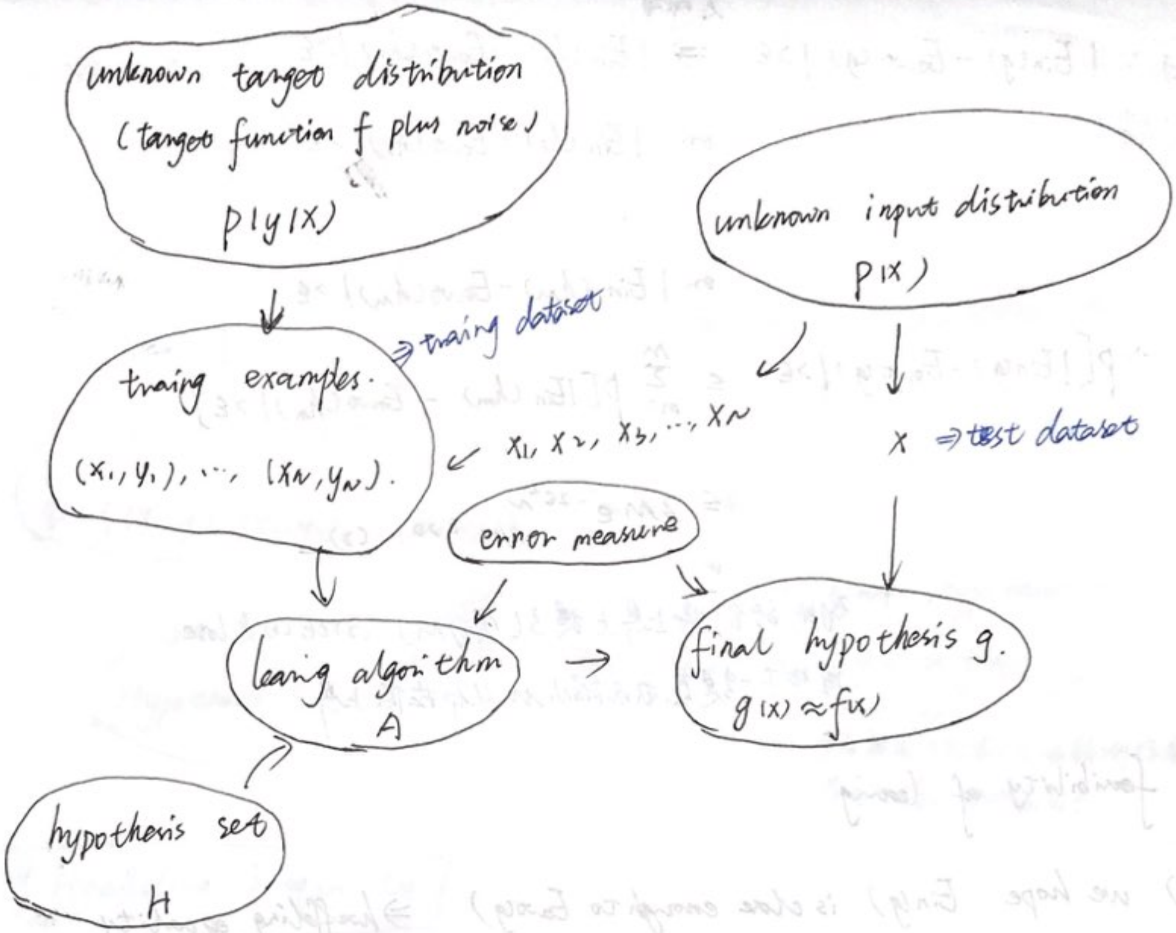
$$\le 2Me^{-2\varepsilon^2 N}, \quad \varepsilon > 0. \quad (2)$$

到此，我们将上界上提了(因为M). (2)比(1)更lose,
因此下一步是想办法缩小M, 从而拉低上界

⇓ feasibility of leaning

1) we hope $E_{in}(g)$ is close enough to $E_{out}(g)$ ⇒ hoeffding equality
2) we hope $E_{in}(g)$ small enough ⇒ test g with traing set

⇒ $E_{out}(g)$ is small enough.


unknown target distribution
(target function f plus noise)
p(y|x)
unknown input distribution
p(x)
⇒ traing dataset
traing examples.
(x1, y1), ..., (xN, yN).
x1, x2, x3, ..., xN
x ⇒ test dataset
error measure
leaing algorithm
A
final hypothesis g.
g(x) ≈ f(x)
hypothesis set
H

Traing vs Testing

由 Hoeffding Inequality,有:

$$P[\,|E_{in}(g) - E_{out}(g)| > \varepsilon\,] \le 2M e^{-2\varepsilon^2 N}, \quad \varepsilon > 0.$$

$\therefore$ 至少有 $1-2Me^{-2\varepsilon^2 N}$ 的概率, $E_{out}(g) \le E_{in}(g) + \varepsilon$.

假设 $\delta = 1-2Me^{-2\varepsilon^2 N}$. $\Rightarrow \varepsilon = \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$

$\therefore$ 至少有 $1-\delta$ 的概率,

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}} \qquad (3)$$

$\Downarrow$

generalization bound.

$\Rightarrow$ 目标: 使用一个更小的值来替换 M, 从而缩小 $E_{out}(g)$ 的上界.

1) 定义 growth function. $\Rightarrow$ 表征假设空间中的有效假设个数.

2) 找到 growth function 的上界

3) 使用 growth function 替换 M 得到更 tight 的 generalization bound.

1) Define Growth Function.

假设 H 将输入空间 X 映射为 $\{-1,+1\}$，那么有：

$$(x_1, x_2, \dots, x_n) \xrightarrow{h_1} (h_1(x_1), h_1(x_2), \dots, h_1(x_n)) \Rightarrow \text{1个 dichotomy}$$

$$(x_1, x_2, \dots, x_n) \xrightarrow{h_2} (h_2(x_1), h_2(x_2), \dots, h_2(x_n)) \Rightarrow \text{1个 dichotomy}$$

1° 由一个假设 h 生成的判别结果(1个 N-tuple)是一个 dichotomy，因为它将 n 个点一分为二(-1 类 +1 类)。

2° h_1 和 h_2 不是 dichotomy，它们的预测结果才是。2个假设可能生成相同的 dichotomy，因此 dichotomy 的数量小于 M.

Definition 由假设空间 H 在 $x_1, \dots, x_n \in X$ n 个点生成的 dichotomies 定义为：

$$H(x_1, x_2, \dots, x_n) = \{(h(x_1), h(x_2), \dots, h(x_n)) \mid h \in H\}$$

Definition growth function

$$m_H(n) = \max_{x_1, \dots, x_n \in X} |H(x_1, \dots, x_n)|$$

growth function 定义为假设集 H 在给定的 n 个样本上可以生成的最多个 dichotomy 的数量.

因为 H 将 X 映射为 2 个类别，因此：$m_H(n) \le 2^n < M$

⇓

从而使用 $m_H(n)$ 替换 M 得到更 tight 的上界.

我们说 H 可以 shatter $x_1, \dots, x_n$ 意味着 H 可以生成所有的 dichotomies，并且 $m_H(n) = 2^n$.

Definition 如果H不能够shatter任何~~大小~~包含k个点的数据集，那么k就是H的一个break point。并且 $m_H(k) < 2^k$.

Definition 给定N个点，定义 $B(N,k)$ 为这N个点上的最大dichotomies个数，且这些dichotomies无法shatter任意包含k个点的子集。因此有 $m_H(N) \le B(N,k)$，如果k是H的break point.

Bounding Growth Function

定义

$$\begin{cases} B(N,1) = 1 \\ B(1,k) = 2 & , \text{for } k>1 \end{cases}$$

$\because\ B(N,k) = \alpha + 2\beta.$

& $\alpha + \beta \le B(N-1,k)$

$\beta \le B(N-1,k-1).$

$\therefore\ B(N,k) = \alpha + 2\beta \le B(N-1,k) + B(N-1,k-1)$

Lemma

$$B(N,k) \le \sum_{i=0}^{k-1} \binom{N}{i}$$

Proof. ~~已知 $B(N_0,k)$~~ 设 $N = N_0+1$，有:

$$\begin{aligned} B(N_0+1,k) &\le B(N_0,k) + B(N_0,k-1) \\ &\le \sum_{i=0}^{k-1}\binom{N_0}{i} + \sum_{i=0}^{k-2}\binom{N_0}{i} = 1 + \sum_{i=1}^{k-1}\binom{N_0}{i} + \sum_{i=1}^{k-1}\binom{N_0}{i-1} \\ &= 1 + \sum_{i=1}^{k-1}\left[\binom{N_0}{i} + \binom{N_0}{i-1}\right] = 1 + \sum_{i=1}^{k-1}\binom{N_0+1}{i} = \sum_{i=0}^{k-1}\binom{N}{i} \end{aligned}$$

$\boxed{\text{Theorem}}$

$$m_H(N) \le \sum_{i=0}^{k-1} \binom{N}{i}$$

$\boxed{\text{Definition}}$ VC dimension. 一个假设集的 VC dimension 定义为~~冊~~H所能最多shatter的数据点数，即 $m_H(d_{vc}) = 2^{d_{vc}}$.

因此 $k = d_{vc}+1$ 是一个 break point，有：$\boxed{m_H(N) \le \sum_{i=0}^{d_{vc}} \binom{N}{i}}$.

$$\boxed{m_H(N) \le N^{d_{vc}} + 1}$$

$\boxed{\text{Definition}}$ (VC generalization bound). For any tolerance $\delta > 0$,

$$E_{out}(g) \le E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4 m_H(2N)}{\delta}}$$

with probability $\ge 1 - \delta$.

$\boxed{\text{Sample Complexity}}$ 假设给定 $\delta > 0$，我们希望泛化错误最多只有 ε，则.

$$N \ge \frac{8}{\varepsilon^2} \ln\left(\frac{4((2N)^{d_{vc}}+1)}{\delta}\right).$$