Chapter 2: Training versus Testing, Theory of generalization:

1. Ein from in sample points; Eout from fresh points. Ein not always generalize to Eout.; Generalization error: discrepancy between Ein & Eout
2. Error bar depends on infinite M, so is meaningless.
   But it is a union bound and overestimated.

$\mathbb{P}[|E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon] \le 2Me^{-2\epsilon^2 N}$, a tolerance level δ, $\delta = 2Me^{-2N\epsilon^2}$ $E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{1}{2N}\ln\frac{2M}{\delta}}$.

$E_{\text{out}} \ge E_{\text{in}} - \epsilon$ for all $h \in \mathcal{H}$.

3. Effective number of hypotheses
   - Growth function def:

Definition 2.1. *Let $\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}$. The dichotomies generated by $\mathcal{H}$ on these points are defined by*

$$\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N) = \{ (h(\mathbf{x}_1), \cdots, h(\mathbf{x}_N)) \mid h \in \mathcal{H} \}. \tag{2.3}$$

Definition 2.2. *The growth function is defined for a hypothesis set $\mathcal{H}$ by*

$$m_{\mathcal{H}}(N) = \max_{\mathbf{x}_1, \cdots, \mathbf{x}_N \in \mathcal{X}} |\mathcal{H}(\mathbf{x}_1, \cdots, \mathbf{x}_N)|,$$

where $|\cdot|$ denotes the cardinality (number of elements) of a set.

pick the N points that give the most dichotomy

Growth function: a measure of number of hypotheses in a set, considered on N points rather than entire input field. $m_{\mathcal{H}}(N) \le 2^N$.

   - Growth function Examples:

Illustrate Example 2.1: growth function for 2D perceptron

Illustrate Example 2.2: cases: positive ray; positive intervals; convest set; hypothesis set complexity↑, growth function↑grows with N faster

   - Break point definition:

Definition 2.3. *If no data set of size k can be shattered by $\mathcal{H}$, then k is said to be a break point for $\mathcal{H}$.*

If k is a break point, then $m_{\mathcal{H}}(k) < 2^k$. Example 2.1 shows that $k = 4$ is a

Exercise 2.1: find break point of the examples above

[Learning-From-Data-A-Short-Course/Solutions to Chapter 2 Training versus Testing.ipynb at master · niuers/Learning-From-Data-A-Short-Course (github.com)]

   - Bound the growth function:

if no break point, growth func = 2^N, replace M by 2^N in the bound of generalization error (discrepancy between Ein and Eout), as N↑, the bound will not ↓ (converge to a constant, verifiable)

if growth func bound by a polynomial, replace it in the generalization error, bound of generalization error: numerator: ln(N); denominator: N. N tends to infinity, bound of error tends to 0. ->generalize well

Mathematical induction proves the polynomial bound of growth func:

Definition 2.4. *$B(N, k)$ is the maximum number of dichotomies on N points such that no subset of size k of the N points can be shattered by these dichotomies.*

Lemma 2.3 (Sauer's Lemma).

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

Theorem 2.4. If $m_{\mathcal{H}}(k) < 2^k$ for some value k, then

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

for all N. The RHS is polynomial in N of degree $k-1$.

Homework： exercise 2.2 verification of theorem 2.4

$$\sum_{0}^{n} \binom{n}{i} = 2^n \qquad (1+1)^n = \sum_{i=0}^{n} \binom{n}{i} 1^i 1^{n-i}.$$

$m_H(N) = N + 2^{\lfloor \frac{N}{2} \rfloor} \le \sum_{i=0}^{k-1} \binom{N}{i}$; k: break point | or $k = N+1$ (valid, max)

$\Downarrow$ N at exp

$\Downarrow$ $\binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{k-1}$

$\Downarrow$ $1 + N + \cdots + \frac{N(N-1)\cdots(N-k+2)}{(k+1)!}$

$\Downarrow$ polynomial in $N^{(k-1)}$

$\not\le$ not true for all N

4. The VC dimension:

Definition 2.5. *The Vapnik-Chervonenkis dimension of a hypothesis set $\mathcal{H}$, denoted by $d_{\text{VC}}(\mathcal{H})$ or simply d_{VC}, is the largest value of N for which $m_{\mathcal{H}}(N) = 2^N$. If $m_{\mathcal{H}}(N) = 2^N$ for all N, then $d_{\text{VC}}(\mathcal{H}) = \infty$.*

$k = d_{\text{VC}} + 1$ **$\mathcal{H}$ can shatter d_{VC} points,**

Exercise 2.3: d = k – 1

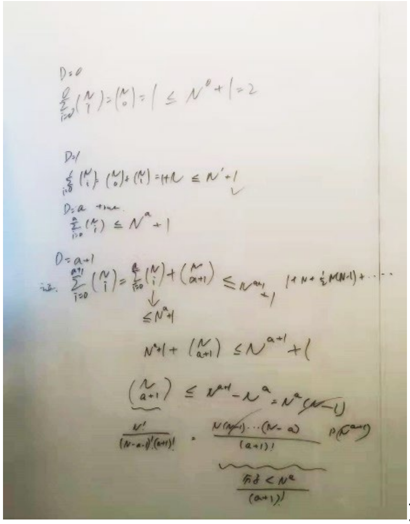
$$m_{\mathcal{H}}(N) \le \sum_{i=0}^{d_{\text{VC}}} \binom{N}{i}.$$

the VC dimension is the order of the polynomial bound on growth func.

$$m_{\mathcal{H}}(N) \le N^{d_{\text{VC}}} + 1.$$

Problem 2.5 Prove by induction that $\sum_{i=0}^{D} \binom{N}{i} \le N^D + 1$, hence

$$m_{\mathcal{H}}(N) \le N^{d_{\text{vc}}} + 1.$$

或者用 github 的证明

5. New inequality -> VC Bound
   $m_H(2N)$:
   Hoeffding inequality: entire input space
   Growth function: only data set
   ∴ replace Eout with E'in (depend on another N)
   ∴ total # of hypotheses depends on 2N.

Not quite:

$$\mathbb{P}[\, |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \,] \;\le\; 2 \;\; m_{\mathcal{H}}(\; N \;) \;\; e^{-\,2\,\epsilon^2 N}$$

but rather:

$$\mathbb{P}[\, |E_{\text{in}}(g) - E_{\text{out}}(g)| > \epsilon \,] \;\le\; 4 \;\; m_{\mathcal{H}}(2N) \;\; e^{-\,\frac{1}{8}\,\epsilon^2 N}$$

The Vapnik-Chervonenkis Inequality

Theorem 2.5 (VC generalization bound)**.** For any tolerance $\delta > 0$,

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}} \qquad (2.12)$$

with probability $\ge 1 - \delta$.

$$P[|E_{in}(g) - E_{out}(g)| > \varepsilon] \leq 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\varepsilon^2 N}$$

$$\Downarrow$$

$$\delta$$

$E_{out}(g) - E_{in} \leq \varepsilon$ with Probability $1-\delta$

$$E_{out}(g) \leq E_{in}(g) + \varepsilon$$

$$\delta = 4 m_{\mathcal{H}}(2N) e^{-\frac{1}{8}\varepsilon^2 N}$$

$$\frac{\delta}{4 m_{\mathcal{H}}(2N)} = e^{-\frac{1}{8}\varepsilon^2 N}$$

$$-\frac{1}{8}\varepsilon^2 N = \ln \frac{\delta}{4 m_{\mathcal{H}}(2N)}$$

$$\varepsilon^2 = \frac{-8}{N} \ln \frac{\delta}{4 m_{\mathcal{H}}(2N)} = \frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}$$

$$\varepsilon = \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$$

$$E_{out}(g) \leq E_{in}(g) + \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$$

6. Why VC bound so loose:

   Exercise 2.5: VC bound very loose

   The slack is from:

   a) Hoeffding inequality: different Eout -> different variance in Ein, -> different probability, however, all the cases covered by one bound ∴hoeffding inequality loose
   b) Growth function $m_H(N)$: did NOT consider probability distribution on input space (if consider, pick specific data set or use expectation). $m_H(N)$ is an upper bound
   c) Use the bound of $m_H(N)$ (polynomial in order d_{vc}), rather than its value, further slack

   VC analysis (bound & dimension)-> generalization performance

   Both useful in practice

7. Sample complexity:

how many training examples N are needed to achieve a certain generalization performance
error tolerance: ε the allowed generalization error
confidence parameter: δ how often the error tolerance ε is violated

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}}$$

with probability $\geq 1 - \delta$.

$$\sqrt{\frac{8}{N} \ln \frac{4 m_{\mathcal{H}}(2N)}{\delta}} \leq \epsilon. \qquad N \geq \frac{8}{\epsilon^2} \ln \left(\frac{4 m_{\mathcal{H}}(2N)}{\delta} \right) \qquad N \geq \frac{8}{\epsilon^2} \ln \left(\frac{4((2N)^{d_{\text{vc}}} + 1)}{\delta} \right)$$

Get N in an iterative way

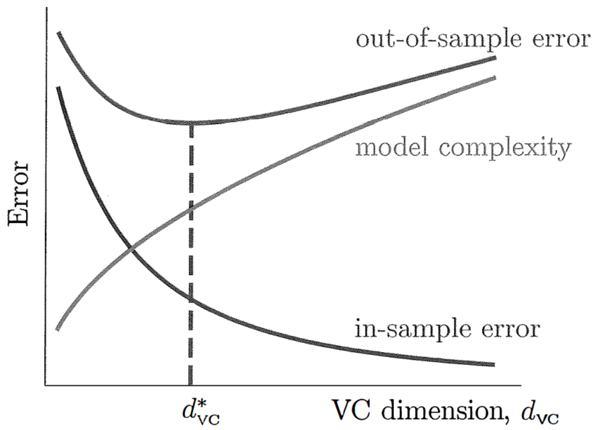
8. penalty for model complexity

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \Omega(N, \mathcal{H}, \delta),$$

where

$$\begin{aligned} \Omega(N,\mathcal{H},\delta) &= \sqrt{\frac{8}{N}\ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)} \\ &\le \sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{vc}}}+1)}{\delta}\right)}. \end{aligned}$$

a) Higher confidence (更加保证偏离不超过某个值，error 大于某个值的可能性非常小的（Hoeffding inequality）)， lower δ，这个 error，或者说 penalty Ω↑。因为要更 confident，只能放宽条件
b) More training examples, N↑， fit better, penalty Ω↓
c) More complex hypothesis set, higher VC dimension, more choices to fit the data Ein↓, penalty Ω↑



9. Etest estimates Eout
Estimate Eout using a test set which is no involved in the training process
Etest generalize to Eout by hoeffding inequality since only one hypothesis g applies to Etest
Exercise 2.6
(a) M is given directly, so do not need to bother with growth func or dvc (applies to binary target function), just use the initial hoeffding multiply with # of hypothesis
(b) more examples on training, more complex hypothesis set, may result in greater penalty
Less examples on training, fewer choices to fit the data, maybe less good g, even generalization (to Eout) bound on Etest is small (Eout Etest close), Etest may be very large due to bad g.
[Learning-From-Data-A-Short-Course/Solutions to Chapter 2 Training versus Testing.ipynb at master · niuers/Learning-From-Data-A-Short-Course (github.com)]
Training set: bias in estimate of Eout, VC bound takes the bias
Test set: no bias, only variance due to finite sample size, just tell how well we did

10. Other Target Types

Real-valued function -> square error

$$E_{\text{out}}(h) = \mathbb{E}\left[(h(\mathbf{x}) - f(\mathbf{x}))^2\right] \quad E_{\text{in}}(h) = \frac{1}{N}\sum_{n=1}^{N}(h(\mathbf{x}_n) - f(\mathbf{x}_n))^2$$

Exercise 2.7 hint: Exctation[x] = ΣxP(x)

Law of Large Number: Ein converge to Eout. Hoeffding inequality is one form

11. Approximation – generalization tradeoff

a) Eout decomposition: bias and variance

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right],$$

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[E_{\text{out}}(g^{(\mathcal{D})})\right] &= \mathbb{E}_{\mathcal{D}}\left[\mathbb{E}_{\mathbf{x}}[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2]\right] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\,\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})]f(\mathbf{x}) + f(\mathbf{x})^2\right]
\end{aligned}$$

$E_D(g^{(D)}(x))$: $D_1 \cdots\cdots D_k$

$\downarrow \qquad \downarrow$

$g_1 \cdots\cdots g_k$

$\frac{1}{K}\sum g_1(x) \cdots g_k(x) \Rightarrow \bar{g}(x)$ random variable

$$\begin{aligned}
&\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})] \\
&= \mathbb{E}_{\mathbf{x}}\left[\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2\right], \\
&= \mathbb{E}_{\mathbf{x}}\Big[\underbrace{\mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})^2] - \bar{g}(\mathbf{x})^2}_{\mathbb{E}_{\mathcal{D}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2\right] \;①} + \underbrace{\bar{g}(\mathbf{x})^2 - 2\bar{g}(\mathbf{x})f(\mathbf{x}) + f(\mathbf{x})^2}_{(\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \;②}\Big],
\end{aligned}$$

-> bias

$$\begin{aligned}
① &= E_D\left[g^{(D)}(x)^2 + \bar{g}(x)^2 - 2g^{(D)}(x)\bar{g}(x)\right] \\
&= E_D\left[g^{(D)}(x)^2\right] + \bar{g}(x)^2 - 2E_D\left[g^{(D)}(x)\right]\bar{g}(x) \\
&= \qquad\qquad - 2\bar{g}(x)^2 \\
&= E_D\left[g^{(D)}(x)^2\right] - \bar{g}(x)
\end{aligned}$$

$$\text{bias}(\mathbf{x}) = (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2$$

How much the learning model is biased away from target function
Limitation from learning model

$$\text{var}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$$

Variation in the final hypothesis, depending on data set

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})] \\ &= \text{bias} + \text{var},\end{aligned}$$

Variance: small variation in data -> vastly different hypothesis
Small model: large bias, small var
Large model: large var, small bias
If data noisy: additional noise term

b) Problem 2.22

Problem 2.22

$$E_{out}(g^{(D)}) = E_{xy}\left[\left(g^{(D)}(x) - y(x)\right)^2\right] \quad y(x) = f(x) + \varepsilon$$

ε zero mean random variable. $E_x(\varepsilon) = 0$ $\quad Var(\varepsilon) = 6^2$

$$E_D[E_{out}(g^{(D)})] = E_D\left[E_{xy}\left[\left(g^{(D)}(x) - f(x) - \varepsilon\right)^2\right]\right]$$

$$= E_x\left[E_D\left[\left(g^{(D)}(x) - f(x)\right)^2 + \varepsilon^2 - 2\left(g^{(D)}(x) - f(x)\right)\varepsilon\right]\right]$$

$$= E_x[\text{Var}(x) + \text{bias}(x)] + E_x(\varepsilon^2) - 2E_x\left[\varepsilon\left(g^{(D)}(x) - f(x)\right)\right]$$

$$= Var + bias + 6^2 - 2\underbrace{E_x(\varepsilon)}_{=0} E_x\left[g^{(D)}(x) - f(x)\right] \quad \text{independent.}$$

$$= var + bias + 6^2$$

c) Example 2.8
Hypothesis 0: bias & var 的推导
Hypothesis 1: 用代码算

$f(x) = \sin(\pi x)$

hypothesis: $g(x) = \frac{y_1 + y_2}{2}$

$y_1 = \sin(\pi x_1)$

$$\bar{g}(x) = E_D\left(g^{(D)}(x)\right)$$

$$= E_D\left(\frac{y_1+y_2}{2}\right)$$

$$= \frac{1}{2}E_D(y_1) + \frac{1}{2}E_D(y_2)$$

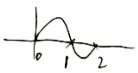
$$= \frac{1}{2}E_D(\sin(\pi x_1)) + \frac{1}{2}E_D(\sin(\pi x_2))$$

$$= 0 + 0 = 0$$

$$\text{bias}(x) = (\bar{g}(x) - f(x))^2$$

$$= f(x)^2 = [\sin(\pi x)]^2$$

$\sin(\pi x)$



$$\text{bias} = E_x(\text{bias}(x)) = E_x[\sin^2(\pi x)]$$

$$= \int_0^2 \sin^2(\pi x) \times \frac{1}{2} dx$$

$$= \frac{1}{2}\int_0^2 \frac{1-\cos 2\pi x}{2} dx$$

$$= \frac{1}{2}\int_0^2 \frac{1}{2} dx - \frac{1}{2}\int_0^2 \frac{1}{2}\cos 2\pi x \, dx$$

$$= \frac{1}{2} - \frac{1}{4}\frac{1}{2\pi}(\sin 2\pi x)\Big|_0^2$$

$$= \frac{1}{2} - \frac{1}{4}\frac{1}{2\pi}(0-0) = \frac{1}{2}$$

$\cos 2x = 1 - 2\sin^2 x$

$\sin^2 x = \frac{1-\cos 2x}{2}$

$\frac{1}{4}(x)\Big|_{-1}^{1} = \frac{1}{4}(1+1)$

$$\text{Var}(x) = E_D\left[(g^{(D)}(x) - \bar{g}(x))^2\right]$$

$$= E_D\left[g^{(D)}(x)^2\right]$$

$$= E_D\left[\left(\frac{y_1+y_2}{2}\right)^2\right]$$

$$= E_D\left[\frac{y_1^2 + y_2^2 + 2y_1y_2}{4}\right]$$

$$= \frac{1}{4}E_D\left[y_1^2 + y_2^2 + 2y_1y_2\right]$$

$$= \frac{1}{4}\left(\frac{1}{2} + \frac{1}{2}\right) = \frac{1}{4}$$

$$\text{Var} = E_x(\text{Var}(x)) = 0.25$$

x_1, x_2 independent

$\bar{g}(x) = 0$

$y_1 = \sin(\pi x_1)$

$E_D(\sin^2(\pi x_1)) = \frac{1}{2}$

$E_D(\sin(\pi x_1)\sin(\pi x_2))$

$= 0 \times 0 = 0$

For the Visionaries

d) Recap VC analysis:

Growth function: a measure of number of hypotheses in a set, considered on N points

the VC dimension is the order of the polynomial bound on growth func.

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

$$E_{\text{out}}(g) \leq E_{\text{in}}(g) + \Omega(N, \mathcal{H}, \delta),$$

where

$$\begin{aligned}\Omega(N, \mathcal{H}, \delta) &= \sqrt{\frac{8}{N} \ln\left(\frac{4 m_{\mathcal{H}}(2N)}{\delta}\right)} \\ &\leq \sqrt{\frac{8}{N} \ln\left(\frac{4((2N)^{d_{\text{vc}}} + 1)}{\delta}\right)}.\end{aligned}$$

More complex hypothesis set, higher VC dimension, more choices to fit the data Ein↓, penalty Ω↑

(Growth function & VC Dimension): Given N points (a set), how many hypothesis -> complexity of hypothesis

e) Example 2.8: given 2 points, the two models (with different learning algorithm), both have one possible hypothesis each (complexity of hypothesis same) -> same VC bound on out-of-sample error.

f) VC analysis: depend purely on Hypothesis Set, independent of learning algorithm
Bias-Variance analysis: out of sample error: bias and variance decomposition

$$E_{\text{out}}(g^{(\mathcal{D})}) = \mathbb{E}_{\mathbf{x}}\left[(g^{(\mathcal{D})}(\mathbf{x}) - f(\mathbf{x}))^2\right],$$

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})] &= \mathbb{E}_{\mathbf{x}}[\text{bias}(\mathbf{x}) + \text{var}(\mathbf{x})] \\ &= \text{bias} + \text{var},\end{aligned}$$

$$\text{bias}(\mathbf{x}) = (\bar{g}(\mathbf{x}) - f(\mathbf{x}))^2 \qquad \text{var}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[(g^{(\mathcal{D})}(\mathbf{x}) - \bar{g}(\mathbf{x}))^2]$$

$$\bar{g}(\mathbf{x}) = \mathbb{E}_{\mathcal{D}}[g^{(\mathcal{D})}(\mathbf{x})]$$

Different complexity of hypothesis set / learning algorithm -> different $g^{(\mathcal{D})}$, building block of bias-variance analysis -> different **bias** & **var** terms

(Bias-Variance analysis based on squared-error (to measure bias and variance), the learning algorithm not have to base on minimizing squared-error measure)

g) Application of Bias-variance analysis:

to compute: Target function + input probability distribution ∴just a conceptual tool to DEVELOP A MODEL

↓bias, -var: prior info regarding the target, to steer the Hypothesis Set in the direction of target func

↓var, -bias: general techniques

h) The learning curve of a model

In Sample Error: $\mathbb{E}_{\mathcal{D}}[E_{\text{in}}(g^{(\mathcal{D})})]$

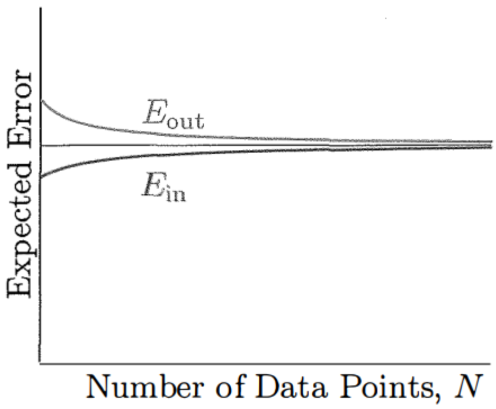
Out of Sample Error: $\mathbb{E}_{\mathcal{D}}[E_{\text{out}}(g^{(\mathcal{D})})]$.

Common：

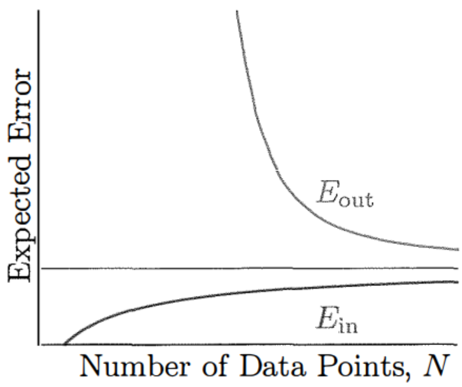
Small N: learning model has a easier task, regardless of outside N. Ein small, Eout large

N↑: hard to fit, Ein↑ -> smallest error the learning model can achieve （点差不多都出来了）

N↑: Eout↓to the best the learning model can achieve



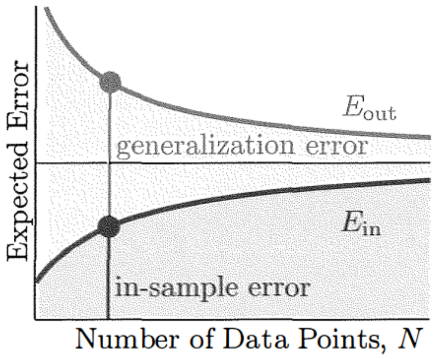
Simple Model



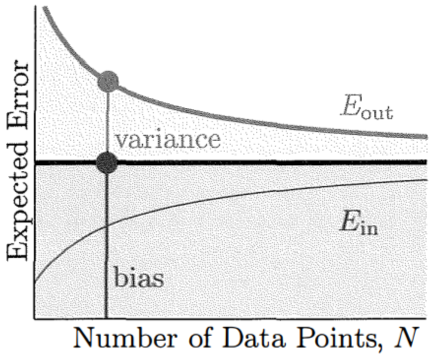
Complex Model

Differences:

Simple model: Converge more quickly but worse ultimate performance



VC Analysis



Bias-Variance Analysis

$$E_{\text{out}}(g) \le E_{\text{in}}(g) + \Omega(N, \mathcal{H}, \delta),$$

$$\begin{aligned}\Omega(N,\mathcal{H},\delta) &= \sqrt{\frac{8}{N}\ln\left(\frac{4m_{\mathcal{H}}(2N)}{\delta}\right)} \\ &\le \sqrt{\frac{8}{N}\ln\left(\frac{4((2N)^{d_{\text{vc}}}+1)}{\delta}\right)}.\end{aligned}$$

Eout=bias+var

Assumption: best approximation to f of the model

= bias = $\bar{g}(\mathbf{x})$ performance

Generalization error bounded by Ω (penalty for model complexity)

N↑, Ω↓, generalization error↓

Recap：

1. Given the hoeffding inequlity, how can it be modified to apply a bound to a hypothesis set with M terms?

2.what is the bound of Eout with respect to Ein, suppose the confidance that Eout is within the bound is 1-δ

3.replace the bound you derived above by VC generalisation bound.
What does each term represent?
growth function - formalise the effective number of hypothesis

4. sample size, complexity of hypothesis set, tolerence level δ.
how are the parameters influencing the VC generalisation bound?

5. how is the VC dimention (model complexity) affecting the Ein and Eout, draw a graph

6.based on square error measures, what is the out-of-sample errors?
(hint: make explicit the dependence of the final hypothesis on its particular data set used)

7.This out-of-sample error can be incomposed into "bias" and "variance",
what do bias and variance imply here?

8.plot and illustrate the "Learning Curves" of simple model and complex model respectively.

9. illustrate the VC analysis and Bias-Variance analysis with the learning curve