Machine Learning Algorithms (CS260) Cheat Sheet

By Patricia Xiao

Notations

Term	Notation
Scalars	$a, b, c, \dots$
Vectors	$\mathbf{u}, \mathbf{v}, \mathbf{w}, \mathbf{x}, \dots$
Matrices	$\mathbf{A}, \mathbf{X}, \mathbf{W}, \dots$
Inner Product	$\langle \mathbf{w}, \mathbf{x} \rangle = \sum_i w_i x_i = \mathbf{w}^T \mathbf{x}$; $\langle \mathbf{W}, \mathbf{X} \rangle = tr(\mathbf{W}^T \mathbf{X})$
* tr(A)	trace of matrix A, $tr(A) = \sum_{i=1}^n a_{ii}$
norm	$\|\cdot\|$, $\|\mathbf{x}\|_2 = \sqrt{\sum_i x_i^2}$
$[a]_+$	$max(a, 0)$
Set	$\mathcal{H}$
Domain Set / Input Space	An arbitrary set $\mathcal{X}$, usually vector of features, $\mathcal{X} \subseteq \mathbb{R}^d$.
Instance	$x \in \mathcal{X}$
Label Set / Target Space	$\mathcal{Y}$, usually $\{0, 1\}$ or $\{-1, +1\}$
Target / Label	$y \in \mathcal{Y}$
Instance-Label Pair	$(x, y) \in \mathcal{X} \times \mathcal{Y}$
Training data	$S = ((x_1, y_1), \dots (x_m, y_m)) \in (\mathcal{X} \times \mathcal{Y})^n$, a **finite** sequence of pairs in $\mathcal{X} \times \mathcal{Y}$
Hypothesis	$h : \mathcal{X} \to \mathcal{Y}$ (prediction rule / learner's output / classifier / function / predictor)
Data Generation Model	$f : \mathcal{X} \to \mathcal{Y}$, $y_i = f(x_i)$
Probability Distribution	$\mathcal{D}$, over $\mathcal{X}$, learner don't know
Probability	$\mathbb{P}(\cdot)$
Expectation	$\mathbb{E}[\cdot]$
Indicator Function	$\mathbb{1}(\epsilon) = 1$ if ϵ is true, otherwise $= 0$
Learning Goal	Find a hypothesis h from $\mathcal{H}$ with (mostly) correct predictions on future **unseen** examples
Correct Classifier	f
Accuracy	ϵ
Confidence	δ
Sample size	Sample complexity: $m_{\mathcal{H}}$, lower bound of learnability
inf	Infimum, $\approx$ lower bound
sup	Supremum, $\approx$ upper bound
exp	Exponential, $exp(x) = e^x$

Learnability Theorems Overview

For binary classification, where $\mathcal{Y} = \{-1, +1\}$, error of h with respect to f is (PAC):

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{x \sim \mathcal{D}}[h(x) \neq f(x)] = \mathcal{D}(\{x \in \mathcal{X} : h(x) \neq f(x)\})$$

That of agnostic PAC:

$$L_{\mathcal{D},f}(h) = \mathbb{P}_{(x,y) \sim \mathcal{D}}[h(x) \neq y] = \mathcal{D}(\{(x, y) \in \mathcal{X} \times \mathcal{Y} : h(x) \neq y\})$$

Important background knowledge include:

1. **i.i.d.** - Independently Identically Distributed, Each x_i is sampled independently according to $\mathcal{D}$.
2. **Empirical Risk** - $L_S(h) = \frac{|i \in [m] : h(x_i) \neq y_i|}{m}$ (m is the training set size), finding a predictor h that minimizes ER is called ERM (Empirical Risk Minimization).

Theorem	assumption	statement (of prob $\geq (1 - \delta)$)
PAC Learnable	$\mathcal{D} \sim \mathcal{X}$; $\mathcal{H}$; Realizability; $m \geq m(\epsilon, \delta)$; ERM	$L_{\mathcal{D},f}(A(S)) \leq \epsilon$
Agnostic PAC Learnable	$\mathcal{D} \sim \mathcal{X} \times \mathcal{Y}$; the rest the same	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$
Uniform Convergence	S is ϵ - representative; $m \geq m_{\mathcal{H}}^{UC}(\epsilon, \delta)$ the rest the same	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$
Nonuniform Learnability	Major change is in expression of prior knowledge; $m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$	$L_{\mathcal{D}}(A(S)) \leq \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + \epsilon$

Formula

1. $(1 - \epsilon)^m \approx e^{-\epsilon m}$, $(1 - x) \leq e^{-x}$
2. (*Union Bound*) $\mathcal{D}(A \cup B) \leq \mathcal{D}(A) + \mathcal{D}(B)$
3. PAC & agnostic PAC: $m_{\mathcal{H}}(\epsilon, \delta) = \frac{1}{\epsilon} log(\frac{|\mathcal{H}|}{\delta})$
4. $\forall$ finite $\mathcal{H}$ is agnostically PAC learnable with: $m_{\mathcal{H}}(\epsilon, \delta) \leq \lceil \frac{2}{\epsilon^2} log(\frac{2|\mathcal{H}|}{\delta}) \rceil$
5. $m_{\mathcal{H}}(\epsilon, \delta) \leq m_{\mathcal{H}}^{UC}(\epsilon/2, \delta)$; **every** $\mathcal{H}$ with uniform convergence property is agnostic PAC learnable.
6. $m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \lceil \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \rceil$
7. $m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h) \leq \frac{-\log(w(h)) + \log(2/\delta)}{2\epsilon^2}$ where $\mathcal{H}$ is the class of all computable functions, not PAC learnable but NU learnable, MDL.
8. In SRM settings, $m_{\mathcal{H}}^{NUL}(\epsilon, \delta, h)$ is upper-bounded by $\min_{n : h \in \mathcal{H}_n} C \frac{d_n - \log(w(h)) + \log(1/\delta)}{\epsilon^2}$.
9. Validation set V and $\ell \in [0, 1]$, $|L_V(h) - L_{\mathcal{D}}(h)| \leq \sqrt{\frac{\log(2/\delta)}{2m_v}}$, $2|\mathcal{H}|$ if optimized $\widehat{h}$.
10. For a finite class $\mathcal{H}$, $VCdim(\mathcal{H}) \leq \log_2(|\mathcal{H}|)$
11. $\sqrt{a^2 - b^2} \leq \sqrt{(a + b)^2} = a + b$

Realizability Assumption

There exists $h^* \in \mathcal{H}$ such that $L_{(\mathcal{D},f)}(h^*) = 0$. It implies that, with probability of 1 over random sample S, $L_S(h^*) = 0$.

No Free Lunch Theorem

Fix $\delta \in (0, 1)$, $\epsilon \in (0, 1/2)$. $\forall$ learner A and training set size m, $\exists \mathcal{D}, f$ such that:

$$\mathbb{P}(L_{\mathcal{D},f}(A(S)) \geq \epsilon) \geq \delta$$

not better than a random guess at 1/2

ϵ - representative sample

A training set S is ϵ - representative when it holds that:

$$\forall h \in \mathcal{H}, \ |L_S(h) - L_{\mathcal{D}}(h)| \leq \epsilon$$

Hints on Proofs

1. PAC: Consider the bad hypothesis class $\mathcal{H}_B$ and misleading samples M.

2. Uniform Convergence:

$$\begin{aligned} &\forall h \in \mathcal{H}, \\ &L_{\mathcal{D}}(h_S) \leq L_S(h_S) + \epsilon^* \\ &\qquad\leq L_S(h) + \epsilon^* \\ &\qquad\leq L_{\mathcal{D}}(h) + \epsilon^* + \epsilon^* \end{aligned}$$

3. Finite class sample complexity upper bound: same with prove that $m_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \lceil \frac{log(2|\mathcal{H}|/\delta)}{2\epsilon^2} \rceil$, and then use the Hoeffding's inequality to bound $\mathcal{D}^m(\cdots > \epsilon) \leq 2exp(-2m\epsilon^2)$; Union bound and that's it.

4. Proof of $|L_V(h) - L_{\mathcal{D}}(h)|$: use Hoeffding's Inequality.

5. Kraft's Inequality: consider generating an expression as flipping coins or other random process, then $\mathbb{P}(\sigma) = \frac{1}{2^{|\sigma|}}$

6. Minimum Description Length (MDL) bound proof: Make $\delta_h = w(h) \cdot \delta$ for each h; Apply Hoeffding to show that for each h, $\mathcal{D}^m(\{S : L_{\mathcal{D}}(h) > L_S(h) + \sqrt{\frac{\log(2/\delta_h)}{2m}}\}) \leq \delta_h$; apply union bound to get altogether they are $\sum \delta_h \leq \delta$.

Hoeffding's Inequality

Let $\theta_1, \ldots \theta_m$ be a sequence of i.i.d. random variables that satisfies:

1. $\forall i, \mathbb{E}[\theta_i] = \mu$

2. $\forall i, \mathbb{P}[a \leq \theta_i \leq b] = 1$

Then $\forall \epsilon > 0$,

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

where $exp(x) = e^x$.

Markov's Inequality

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

specifically, when $a \in (0,1)$ and $Z \sim [0,1]$, assume that $\mathbb{E}[Z] = \mu$, we have:

$$\mathbb{P}[Z > a] \geq \frac{\mu - a}{1 - a} \geq \mu - a$$

k-fold cross validation

- divide the training dataset into k folds, use one fold as validation set, the rest for training
- a method of selecting the best parameters before going testing
- use the average of all the selections of $i \in \{1, \ldots k\}$'s error to be the estimated error of a parameter set

Error Decomposition

$$L_{\mathcal{D}}(h_S) = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h) + (L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h))$$

1. The approximation error: $\epsilon_{app} = \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$
   - bring in by restriction of $\mathcal{H}$
   - independent from S
   - decreases with complexity of $\mathcal{H}$ (denoted by size or VCdim)
2. The estimation error: $\epsilon_{est} = L_{\mathcal{D}}(h_S) - \min_{h \in \mathcal{H}} L_{\mathcal{D}}(h)$
   - Result of L_S being only an estimation of L_D
   - Decreases with the size of S
   - Might increase with the complexity of $\mathcal{H}$.

VC Dimension

$\mathcal{H}$ **Shatters** C means that all possible value of a given set C could be explained by a hypothesis from class $\mathcal{H}$, $|\mathcal{H}_C| \leq 2^{|C|}$, where $|\mathcal{H}_C|$ is the restriction of $\mathcal{H}$ to C.

$$VCdim(H) = sup\{|C| : \mathcal{H} \ shatters \ C\}$$

Bias-Variance Decomposition for Regression

Given that $(\mathbf{x}, y) \sim \mathcal{D}$, regression loss-function $\ell(h, (\mathbf{x}, y)) = (h(\mathbf{x}) - y)^2$.
The expected loss is:

$$\begin{aligned} L_{\mathcal{D}}(h) &= \mathbb{E}_{\mathcal{D}}[\ell(h, (\mathbf{x}, y))] \\ &= \int\int (h(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy \\ &= \int (h(\mathbf{x}) - h^*(\mathbf{x}))^2 p(\mathbf{x}, y) d\mathbf{x} \\ &\quad + \int\int (h^*(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy \end{aligned}$$

The expectation

$$\begin{aligned} \mathbb{E}_S[L_{\mathcal{D}}(h)] &= \mathbb{E}_S[\mathbb{E}_{\mathcal{D}}[\ell(h, (\mathbf{x}, y))]] \\ &= \mathbb{E}_S[\int (h_S(\mathbf{x}) - h^*(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}] \\ &\quad + \int\int (h^*(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy \end{aligned}$$

where $\int\int (h^*(\mathbf{x}) - y)^2 p(\mathbf{x}, y) d\mathbf{x} dy$ is the noise and:

$$\begin{aligned} &\mathbb{E}_S[\int (h_S(\mathbf{x}) - h^*(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x}] \\ &= \int (\mathbb{E}_S[h_S] - h^*(\mathbf{x}))^2 p(\mathbf{x}) d\mathbf{x} \\ &\quad + \int \mathbb{E}_S[(h_S(\mathbf{x}) - E_S[h_S(\mathbf{x})])^2] p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

where the first part is $bias^2$ and the second part is the *variance*.

Growth Function

The growth function of $\tau_{\mathcal{H}}(m)$ is defined as:

$$\tau_{\mathcal{H}}(m) = \max_{C \subset \mathcal{X}, |C| = m} |\mathcal{H}_C|$$

$\tau_{\mathcal{H}}(m)$ the number of different functions from a set $\mathcal{C}$ of size m to 0, 1 that can be obtained by restricting $\mathcal{H}$ to C.
If $VCdim(\mathcal{H}) = d$ then for any $m \leq d$ we have $\tau_{\mathcal{H}}(m) = 2^m$, $\mathcal{H}$ induces all possible functions from C to 0,1.

Sauer-Shelah-Perles-Vapnik-Chervonenkis Lemma

Given $VCdim(\mathcal{H}) \leq d \leq, \infty$, then for all $C \subset \mathcal{X}$ s.t. $|C| = m > d+1$, we have:

$$|\mathcal{H}_C| \leq (\frac{em}{d})^d$$

Fundamental Theorem of Learning

$\mathcal{H}$ is a class of binary classifiers with $VCdim(\mathcal{H}) = d$. Then there are absolute constants C_1 and V_2 such that the sample complexity of PAC learning $\mathcal{H}$ is:

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

And this sample complexity is achieved using ERM rule.

Prior Knowledge

Described as hypothesis class $\mathcal{H}$ in PAC learning and uniform learning. However there are other ways of expressing it, such as bias to shorter expressions. Generally, bias could be denoted as a weight $w(h)$ assigned to each hypothesis in a **countable** hypothesis class $\mathcal{H}$. The weight reflects prior knowledge on the importance of each h.

$$\sum_{h \in \mathcal{H}} w(h) \leq 1$$

An example is the description length.

Description Length

- Description language is denoted by $d(h)$
- The term **prefix-free** means that $\forall h \neq h'$, $d(h)$ is not a prefix of $d(d')$; could always be achieved by including "end-of-word" symbol.
- Let $|h|$ be the length of $d(h)$
- Then, set $w(h) = 2^{-|h|}$
- $\sum_h w(h) \leq 1$ according to **Kraft's inequality**.

Kraft's Inequality

If $\mathcal{S} \subset \{0,1\}$ is a prefix-free set of strings, then:

$$\sum_{\sigma \in S} \frac{1}{2^{|\sigma|}} \leq 1$$

Minimum Description Length (MDL) bound

Let $w : \mathcal{H} \to \mathbb{R}$ be such that $\sum_{h \in \mathcal{H}} w(h) \leq 1$. Then with prob $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$ we have:

$$\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{-\log(\mathbf{w(h)}) + \log(2/\delta)}{2m}}$$

Compared with VC bound:

$$\forall h \in \mathcal{H}, L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{\mathbf{VCdim}(\mathcal{H}) + \log(2/\delta)}{2m}}$$

Minimizing VC bound: ERM rule; Minimizing MDL bound: MDL rule.

Minimum Description Length (MDL) Guarantee

For every $h \in \mathcal{H}$, w.p. $\geq 1 - \delta$ over $S \sim \mathcal{D}^m$ we have:

$$L_{\mathcal{D}}(MDL(S)) \leq L_S(h) + \sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}}$$
$$\leq L_{\mathcal{D}}(h) + 2\sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}}$$

Note than VC dim could be infinite.

Condition of NU Learnable

A class $\mathcal{H} \subset \{0,1\}^{\mathcal{X}}$ is non-uniform learnable **if and only if** it is a **countable** union of **PAC learnable hypothesis classes**.

Structural Risk Minimization (SRM)

$$SRM(S) \in \underset{h \in \mathcal{H}}{\operatorname{argmin}}[L_S(h) + \min_{n: h \in \mathcal{H}_n} \sqrt{C \frac{d_n - \log(w(n)) + \log(1/\delta)}{m}}]$$

where $w(n) = w(\mathcal{H}_n)$

The Cost of Weaker Prior Knowledge

Suppose: $\mathcal{H} = \cup_n \mathcal{H}_n$, where $VCdim(\mathcal{H}_n) = n$.

- If, for some $h^* \in \mathcal{H}_n$ has $L_{\mathcal{D}}(h^*) = 0$, we can apply ERM so the sample complexity is $C\frac{n + \log(1/\delta)}{\epsilon^2}$
- Without the prior knowledge, sample complexity will be $C\frac{n + \log(\pi^2 n^2/6) + \log(1/\delta)}{\epsilon^2}$

Condition of NU Learnable: Proof

Assume that $\mathcal{H}$ is non-uniform learnable with sample complexity $m_{\mathcal{H}}^{NUL}$

- For every $n \in \mathbb{N}$ let $\mathcal{H}_n = \{h \in \mathcal{H} : m_{\mathcal{H}}^{NUL}(\frac{1}{8}, \frac{1}{7}, h) \leq n\}$, then clearly $\mathcal{H} = \cup_{n \in \mathbb{N}} \mathcal{H}_n$
- For every $\mathcal{D}$ s.t. $\exists h \in \mathcal{H}_n$ with $L_{\mathcal{D}}(h) = 0$, we have that $\mathcal{D}^n(\{S : L_{\mathcal{D}}(S(S)) \leq \frac{1}{8}\}) \geq \frac{6}{7}$
- The fundamental theorem of statistical learning implies that each $\mathcal{H}_n$ has finite VC dimension d_n, each of them is agnostic PAC learnable.
- Choose a proper weight so that $\sum_n w(n) \leq 1$ and apply it to $w(n) = w(\mathcal{H}_n)$. One example is $w(n) = \frac{6}{\pi^3 n^2}$ since sum up from 1 to ∞ it adds up to 1.
- Choose $\delta_n = \delta \cdot w(n)$ and $\epsilon_n = \sqrt{C\frac{d_n + \log(1/\delta_n)}{m}}$.
- By the fundamental theorem, for every n, $\mathcal{D}^m(\{S : \exists h \in \mathcal{H}_n, L_{\mathcal{D}}(h) > L_S(h) + \epsilon_n\}) \leq \delta_n$.
- Apply union bound, $\mathcal{D}^m(\{S : \exists n, h \in \mathcal{H}_n, L_{\mathcal{D}}(h) > L_S(h) + \epsilon_n\}) \leq \sum_n \delta_n \leq \delta$.

SRM Guarantee Proof

By NUL, we have:

$$L_{\mathcal{D}}(SRM(S)) \leq L_S(SRM(S)) + \min_{n: h \in \mathcal{H}_n} \sqrt{\frac{-\log(w(SRM(S))) + \log(2/\delta)}{2m}}$$

By the optimality of SRM, we have:

above right hand side

$$\leq L_S(h) + \min_{n: h \in \mathcal{H}_n} \sqrt{\frac{-\log(w(h)) + \log(2/\delta)}{2m}}$$

Claim: For any infinite domain set $\mathcal{X}$, $\mathcal{H} = \{0,1\}^{\mathcal{X}}$ is not a countable union of classes of finite VC-dimension, hence such $\mathcal{H}$ are **not** non-uniformly learnable.

Sample Midterm Conclusion

1. PAC learnable problem

   a. PAC learnable $\rightarrow$ ERM Algorithm, specify a loss function and describe it using math language

   b. Describe the occasions of making mistakes, using math language; in other word, describe the misleading data that leads to *bad hypothesis class*, why and how. Union Bound infers that we need to find $(1-\epsilon/2)^m \leq \delta/2$, for we have 2 bounds to decide.

   c. Let a margin be the distribution of $\mathbb{P}[\cdot] = \epsilon$, falling into that region means that single data point's error will be no more than ϵ; then use something like $\mathbb{P}[no\ misled] = (1-\epsilon)^m \leq e^{-\epsilon m} \leq \delta$ to get the bound of m by ϵ, δ.

2. VC Dimension Steps to prove VC Dimension:

   - Form a sample set C
   - Prove that C could be shattered by $\mathcal{H}$
   - Prove that adding another sample point then $\mathcal{H}$ could no longer shatter C'

   In this problem we need to form the set C and prove that for all the possibilities of C it will have corresponding hypothesis in $\mathcal{H}$. In this specific case where it is a hypothesis class of axis aligned rectangles in $\mathbb{R}^d$, I suggest forming a dataset C where points are in pairs, located on the axis, paired like $(-a, a)$. To illustrate shattering, we could specify a rectangle using a and d, denoting it by assigning the range to each of its dimension. For example, $h = rect((-2a, 2a), (-2a, 2a), \ldots)$ and to exclude 1 positive point from the set, $(-2a, 2a) \rightarrow (-a/2, 2a)$, to exclude 2 we do $(-a/2, a/2)$. So we could make any number from 0 to 2d using those combinations. But if there's an additional point added, it must have $\cos\langle e, x\rangle \neq 0$ with one of the axis's base vectors. From slicing the triangle on a certain plane we know why it doesn't work.o
   Basically, it tests your ability of formally describing the proof in n-d space.

3. Uniform Convergence

   (a) Prove the uniform convergence accuracy, given a replacement of Hoeffding's inequality: the **Bernstein's Inequality**.

   $$\mathbb{P}[|\sum_{i=1}^{m}(\theta_i - \mu)| > \epsilon] \leq 2exp(-\frac{\epsilon^2/2}{\sum_{i=1}^{m}\mathbb{E}[X_i^2] + M\epsilon/3}) \tag{1}$$

   where for i.i.d. $\theta_1 \ldots \theta_m$, $\mathbb{E}[\theta_i] = \mu$ and $|\theta_i| \leq M$. To prove it:

   - Set the value of the right hand side of the inequation (1) to be δ.
   - $a\log(b) = \log(a^b)$, $-\log(\delta/2) = \log(2/\delta)$
   - For $ax^2 + bx + c = 0$, $x = \frac{-b\pm\sqrt{b^2-4ac}}{2a}$, then we get ϵ.
   - Triangular Inequality, $\sqrt{a^2-b^2} \leq a + b$, holds that probability is always $1-\delta$ and that the $\mathbb{P}$ part could be omitted, so it is proved.

   (b) Prove the upper bound of empirical risk. (w.p. $1-\delta$)

   - We have that: $\theta_i = \mathbb{1}(h(x_i) = y)$, $|\theta_i| \leq 1 = M$, $\mathbb{E}[x_i^2] = \mathbb{E}[x_i] = L_{\mathcal{D}}(h)$; $\frac{1}{m}\sum_i \theta_i = L_S(h)$.
   - According to part (a), $\mathbb{P}[|L_S(h) - L_{\mathcal{D}}(h)| \geq \epsilon'] \leq \delta'$, where coming from the right hand side of (a) conclusion, $\epsilon' = \frac{2M\log(2/\delta)}{3m} + \sqrt{\frac{2\mathbb{E}[\theta_1^2]\log(2/\delta)}{m}} = \frac{2\log(2/\delta)}{3m} + \sqrt{\frac{2L_{\mathcal{D}}(h)\log(2/\delta)}{m}}$
   - With Prob $\geq \frac{\delta'}{2}$, we get the lower bound of $|L_S(h) - L_{\mathcal{D}}(h)|$ and conclude that $\sum_{h\in\mathcal{H}}\mathcal{D}^m(\{\cdot\}) \leq \delta$, applying **union bound** we'll get the result. $(\frac{2\log(2/\delta)}{3m} + \sqrt{\frac{2L_{\mathcal{D}}(h)\log(2/\delta)}{m}}$ is the upper bound of $|L_S(h) - L_{\mathcal{D}}(h)|$. It holds uniformly on $\mathcal{H}$ w.p. at least $1 - |\mathcal{H}|\frac{\delta'}{2}$, set $\delta = |\mathcal{H}|\frac{\delta'}{2}$. Bridge $L_{\mathcal{D}}(\widehat{h_S}) \rightarrow L_S(\widehat{h_S}) \rightarrow L_S(h^*) \rightarrow L_{\mathcal{D}}(h^*)$, 2 times diff, thus proved; $C_1 = \frac{5\sqrt{2}}{2}, C_2 = \frac{13+2\sqrt{6}}{3}$ to be specific.)

   (c) Use (b), and make $L_{\mathcal{D}}(\widehat{h_S}) - L_{\mathcal{D}}(h^*) \leq \epsilon$

4. Validation and model selection
   Clarification: $h^* \in \operatorname{argmin}_{h\in\mathcal{H}} L_{\mathcal{D}}(h)$ is contained in $\mathcal{H}_j$, while at the same time it could belong to multiple classes.

   a. ERM rule on $\mathcal{H}_j$, $VCdim(\mathcal{H}_j) = d_j$, according to the **VC bound** (mentioned in MDL), $\epsilon' = \sqrt{C\frac{d_j+\log(2/\delta)}{(1-\alpha)m}}$ and $L_D(\widehat{h_j}) \leq L_S(\widehat{h_j}) + \epsilon' \leq L_S(\widehat{h^*}) + \epsilon' \leq L_D(h^*) + 2\epsilon'$. ($2/\delta$: because $1-\delta/2$)

   b. Note that $\widehat{h} is ERM of \{\widehat{h_1}\ldots\widehat{h_k}.\}$ $|h_V - h_D|$ (denoted as $L_D(\widehat{h})$ and $L_D(\widehat{h_j})$ in this question) upper bound, given $\delta/2$, and that $|\mathcal{H}| = k$, result in $\sqrt{\frac{\log(4k/\delta)}{2\alpha m}}$ (twice to be the answer of (b)). Similar with (a) but use $L_D(\widehat{h}) \rightarrow L_V(\widehat{h}) \rightarrow L_V(\widehat{h_j}) \rightarrow L_D(\widehat{h_j})$. (According to **the fundamental theorem of learning**, agnostic PAC / UC sample complexity $\in [C_1\frac{d+\log(1/\delta)}{\epsilon^2}, C_2\frac{d+\log(1/\delta)}{\epsilon^2}]$, PAC sample complexity $\in [C_1\frac{d+\log(1/\delta)}{\epsilon}, C_2\frac{d\log(1/\epsilon)+\log(1/\delta)}{\epsilon}]$.)

   c. Known $L_D(\widehat{h_j}) - L_D(h^*)$ and $L_D(\widehat{h}) - L_D(\widehat{h_j})$, using union bound, $a + b = c$.

5. Nonuniform learnability

   a. Assign weight to δ_i, $\forall h \in \mathcal{H}, \epsilon, \delta$, let $m \geq m^{UC}_{\mathcal{H}_{n(h)}}(\epsilon, w(n(h))\delta)$, since w add up to 1, w.p. $\geq 1-\delta$ over $S \sim \mathcal{D}^m$, $\forall h' \in \mathcal{H}$, $L_D(h') \leq L_S(h') + \epsilon_{n(h')}(m, w(n(h'))\delta)$. (Holds particular for SRM, A(S).)
   By definition of SRM $L_D(A(S)) \leq \min_{h'}[L_S(h') + \epsilon_n(h')(m, w(n(h'))\delta)] \leq L_S(h) + \epsilon_n(h)(m, w(n(h))\delta)$. If $m \geq m_{\mathcal{H}_{n(h)}}(\epsilon/2, w(n(h))\delta)$, then clearly $\epsilon_n(h)(m, w(n(h))\delta) \leq \epsilon/2$. Each $\mathcal{H}_n$ is UC so w.p. $\geq 1-\delta$, $L_S(h) \leq L_D(h) + \frac{\epsilon}{2}$, proved by using $L_D(\widehat{h}) \rightarrow L_V(\widehat{h}) \rightarrow L_V(\widehat{h_j}) \rightarrow L_D(\widehat{h_j})$.

   b. The cost of weaker prior knowledge. Make $VCdim(\mathcal{H}_n) = n$ and then apply w to δ. The version in textbook uses that $m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h)$ is bounded by $m^{UC}_{\mathcal{H}_n}(\frac{\epsilon}{2}, w(n)\delta)$, and conclude that $m^{NUL}_{\mathcal{H}}(\epsilon, \delta, h) - m^{UC}_{\mathcal{H}_n}(\frac{\epsilon}{2}, w(n)\delta) \leq 4C\frac{2\log(2n)}{\epsilon^2}$, where $m^{UC}_{\mathcal{H}_n}(\epsilon, \delta) = C\frac{n+\log(1/\delta)}{\epsilon^2}$.