Learning Theory

PAC framework

- Train a model $\Leftrightarrow$ Take a course
- Training examples $\Leftrightarrow$ Take exercises/homework
- Testing $\Leftrightarrow$ Take exam
- Can we estimate our performance in exam by performance on exercise?
  - training error $\rightarrow$ testing error
- E_{in} := the in-sample error i.e. **Error in Training data**
  - Let $h \in \mathcal{H}$
    - e.g. $h(x) = \text{sign}(w^T x + b)$
  - model function hypothesis $\rightarrow$ hypothesis space
  - $E_{in}(h) = \frac{1}{n}\sum_{i=1}^{n} 1(h(x_i) \neq y_i) \rightarrow$ **Error Rate**
  - $\{(x_1, y_1), \ldots, (x_n, y_,)\}$ are training data, $(x_i, y_i) \sim P_{xy}$
- E_{out} := the out-of-sample error
  - measures how well a model **generalizes**
  - $E_{out}(h) := P(h(x) \neq y) = E_{(x,y)\sim P_{xy}}[1(h(x) \neq y)]$
- $E_{out}(h) - E_{in}(h)$ is the **Generalization Error**
- We can say with a large probability $1-\delta$ (δ is small), $E_{out}(h) - E_{in}(h) < \delta$. This is called **Probably Approximately Correct (PAC) Learning Framework**

Hoeffding Inequality

- $x_1, \ldots, x_n$ are independent random variables, $x_i \in [a_i, b_i]$, $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$. Then $\forall \epsilon > 0$, we have
  - $P(\bar{x} - E[\bar{x}] \geq \delta) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2})$
  - $P(E[\bar{x}] - \bar{x} \geq \delta) \leq \exp(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2})$

Growth Function

- Now for a given fixed h, we have:
  $P(E_{out}(h) - E_{in}(h) \geq \epsilon) \leq \exp(-2n\epsilon^2)$
  Because

$$\begin{aligned}
E_{in}(h) &= \frac{1}{n}\sum_{i=1}^{n} 1(h(x_i) \neq y_i) = \bar{x} \quad\quad \Rightarrow x_i \in [0,1] \\
E[\bar{x}] &= E[\frac{1}{n}\sum_{i=1}^{n} 1(h(x_i) \neq y_i)] \\
&= \frac{1}{n}\sum_{i=1}^{n} E_{(x,y)\sim P_{xy}}[1(h(x_i) \neq y_i)] \\
&= \frac{1}{n} n E_{out}(h) \\
&= E_{out}(h) \\
P(E_{out}(h) - E_{in}(h) < \epsilon) &> 1 - \exp(2n\epsilon^2)
\end{aligned}$$

  - with probability at least $1-\delta$, $\exists h \in \mathcal{H}$, $E_{out}(h) - E_{in}(h) < \epsilon$
  - But this bound doesn't consider training by assuming h is given before *seeing* the training data. It's not meaningful in practice.
- Since we cannot know which $h \in \mathcal{H}$ to use before seeing training data, we can bound $\mathcal{H}$ instead, thus independent of particular h.
  Let's first assume $\mathcal{H}$ is finite, $\mathcal{H} = \{h_1, \ldots, h_M\}$.

$$\begin{aligned}
P(\exists h \in \mathcal{H}, E_{out}(h) - E_{in}(h) \geq \epsilon) &\leq \sum_{i=1}^{n} P(E_{out}(h_i) - E_{in}(h_i) \geq \epsilon) \\
&\leq M \exp(-2n\epsilon^2)
\end{aligned}$$

  This is the first practical **PAC learning bound**.
  Let

$$\delta = M \exp(-2n\epsilon^2)$$
$$\epsilon = \sqrt{\frac{1}{2n} \log \frac{M}{\delta}}$$

With probability at least $1-\delta$, we have $\forall h \in \mathcal{H}$, $E_{out}(h) - E_{in}(h) < \sqrt{\frac{1}{2n} \log \frac{M}{\delta}}$
  - $n \nearrow$, $M \searrow$, generalization error $\searrow$
  - $M \nearrow$, will be overfitting
- What if H is infinite?
  - These different hypotheses(hyperplane) give the same classification results on finite examples.
  - Union bound counts each h once, but if one h satisfies **PAC learning bound**, then all other h (with same classification results) will also satisfy the PAC bound.
- Growth Function
  - measure effective $\#$ of hypotheses in $\mathcal{H}$ on finite data
  - Assume binary classification $y \in \{-1, 1\}$, $x \in X$
    Given n training samples $x_1, \ldots, x_n \in X$, apply $h \in \mathcal{H}$ to them to get n-tople $(h(x_1), \ldots, h(x_n))$ of ± 1s. (called a *dichotomy*)
    Let $\mathcal{H}(x_1, \ldots, x_n) = \{(h(x_1), \ldots, h(x_n)) | h \in \mathcal{H}\}$. (set has no repeated elements)
    then $m_{\mathcal{H}}(h) := \underset{x_1, \ldots, x_n \in X}{max} |\mathcal{H}(x_1, \ldots, x_n)|$
  - e.g. $m_{\mathcal{H}}(3) = 2^3 = 8$
  - e.g. $m_{\mathcal{H}}(4) = 14 < 2^4$
- If $\mathcal{H}(x_1, \ldots, x_n)$ contains all possible ± 1s assignments to a subset of $\{x_1, \ldots, x_n\}$, denoted by S. We say $\mathcal{H}(x_1, \ldots, x_n)$ **shatters** S.
  - If $\mathcal{H}$ shatters S, then $\mathcal{H}$ shatters *all subsets* of S.

Vapnik-Chervonenkis (VC) dimension

- We call the maximum n s.t. $m_{\mathcal{H}}(n) = 2^n$ **the VC dimension of** $\mathcal{H}$, denoted by $d_{VC}(H)$ (d_{VC} for short)
  - $\#$ of parameters $\approx d_{VC}$
  - e.g. $\mathbb{R}^d$-dimension space, linear classifier $\mathcal{H} \to d_{VC}(\mathcal{H}) = d + 1$
  - $d_{VC}(\mathcal{H})$ **measures effective dimensions of** $\mathcal{H}$
  - small d_{VC} ⇔ small hypothesis space ⇔ less separating power ⇔ more generalizing power
  - VC维表明，只有这么多的数据能够被模型打散，如果数据更多，就不可能有一个模型能将它们在各种情况下都分类正确. 比如对于4个样本点而言，异或的情况使得任何一个线性分类器都不能对其正确分类. 可以说，VC维越大，对应空间的模型就越复杂
- **Saucer's Lemma**

$$m_{\mathcal{H}}(n) \leq \sum_{i=1}^{d_{VC}} \binom{n}{i} = O(n^{d_{VC}})$$

  - Proof:
    - First prove a stronger Lemma:
      On any points $x_1, \ldots, x_n \in X$, the $\#$ of subsets of $\{x_1, \ldots, x_n\}$ that can be shattered by $\mathcal{H}(x_1, \ldots, x_n)$ is at least $|\mathcal{H}(x_1, \ldots, x_n)|$
      - e.g. $\mathcal{H}(x_1, x_2, x_3) = \{(+1, -1, -1), (-1, +1, -1), (-1, +1, +1)\}$. So $|\mathcal{H}(x_1, x_2, x_3)| = 3$
        Let's check all subsets of $\{x_1, x_2, x_3\}$ to see whether it is shattered by $\mathcal{H}$:
        $\emptyset, \{x_1\}, \{x_2\}, \{x_3\}$ are shattered by $\mathcal{H}$, so $\# = 4 \geq 3$
      - e.g. $\mathcal{H}(x_1, x_2, x_3) = \{(-1, +1, -1), (-1, +1, +1)\}$.
        $\emptyset, \{x_3\}$ are shattered by $\mathcal{H}$, so $\# = 2 \geq 2$
      - Prove by **induction**.
        Base case is $|\mathcal{H}(x_1, \ldots, x_n)| = 1$, then $\emptyset$ can be shattered.
        Assume the lemma is true for all $\mathcal{H}'$ s.t. $|\mathcal{H}'(x_1, \ldots, x_n)| < |\mathcal{H}(x_1, \ldots, x_n)|, \quad |\mathcal{H}(x_1, \ldots, x_n)| \geq 2$
        Without loss of generality(W.L.O.U), let x_1 be a point that can take both $+1$ and -1 in $\mathcal{H}(x_1, \ldots, x_n)$. x_1 must exist, otherwise $|\mathcal{H}(x_1, \ldots, x_n)| = 1$
        Then divide $\mathcal{H}(x_1, \ldots, x_n)$ into $\mathcal{H}_1(x_1, \ldots, x_n)$ and $\mathcal{H}_2(x_1, \ldots, x_n)$, s.t. $\mathcal{H}_1(x_1, \ldots, x_n)$ only contains $x_1 : +1$ dichotomies and $\mathcal{H}_2(x_1, \ldots, x_n)$ only contains $x_1 : -1$ dichotomies.
        By induction hypothesis:

$$\begin{aligned} &\# \text{ of subsets shattered by } \mathcal{H}_1 + \# \text{ of subsets shattered by } \mathcal{H}_2 \\ &\geq |\mathcal{H}_1(x_1, \ldots, x_n)| + |\mathcal{H}_2(x_1, \ldots, x_n)| \end{aligned}$$

        Now consider a subset S of $\{x_1, \ldots, x_n\}$:
        - If S is only shattered by $\mathcal{H}_1$ or $\mathcal{H}_2$, then S is shattered by $\mathcal{H}$
        - If S is shattered by both $\mathcal{H}_1$ and $\mathcal{H}_2$, then S is shattered by $\mathcal{H}$ and $S \bigcup \{x_1\}$ is also shattered by $\mathcal{H}$
          Firstly, S doesn't contain x_1, because either $\mathcal{H}_1$ or $\mathcal{H}_2$ only contains one assignment to x_1 $\to$ cannot cover both ± 1 of x_1

Secondly, every dichotomy of $S\bigcup\{x_1\}$ must correspond to a dichotomy of $S\bigoplus\{x_1:+1\}$ or a dichotomy of $S\bigoplus\{x_1:-1\}$, where the former appears in $\mathcal{H}_1$ and the latter appears in $\mathcal{H}_2$.

So

$$\begin{aligned}&\#\text{ of subsets shattered by }\mathcal{H}\\ \geq&\#\text{ of subsets shattered by }\mathcal{H}_1+\#\text{ of subsets shattered by }\mathcal{H}_2\\ \geq&|\mathcal{H}_1(x_1,\ldots,x_n)|+|\mathcal{H}_2(x_1,\ldots,x_n)|\\ =&|\mathcal{H}(x_1,\ldots,x_n)|\end{aligned}$$

- Finally, if $\mathcal{H}$ has **a finite VC dimension** d_{VC} (no subsets of size $\geq d_{VC}+1$ can be shattered by $\mathcal{H}$), then we have on any $x_1,\ldots,x_n\in X$:

$$\begin{aligned}|\mathcal{H}(x_1,\ldots,x_n)|&\leq\#\text{ of subsets shattered by }\mathcal{H}(x_1,\ldots,x_n)\\ &\leq\sum_{i=1}^{d_{VC}}\binom{n}{i}\end{aligned}$$

So

$$m_{\mathcal{H}}(n)=\underset{x_1,\ldots,x_n}{max}|\mathcal{H}(x_1,\ldots,x_n)|\leq\sum_{i=1}^{d_{VC}}\binom{n}{i}$$

- **The VC Generalization bound**
  With probability at least $1-\delta$, $\forall h\in\mathcal{H}$,

$$E_{out}(h)-E_{in}(h)<\sqrt{\frac{8}{n}\log\frac{2m_{\mathcal{H}(2n)}}{\delta}}=O(\sqrt{d_{VC}\frac{\log n}{n}-\frac{\log\delta}{n}})$$

  - $n\nearrow\quad d_{VC}\searrow\quad\Rightarrow\quad\text{generalization error}\searrow$
  - This is a very loose bound, because we always consider **the worst cases**.