DDA5001 Machine Learning

Training versus Testing (Part II)

Xiao Li

School of Data Science
The Chinese University of Hong Kong, Shenzhen

Recap: In-sample Error versus Out-of-sample Error

- In-sample Error: Given a set of training samples $\{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$,

$$\boxed{\mathrm{Er}_{\mathrm{in}} = \frac{1}{n} \sum_{i=1}^{n} e(f_{\boldsymbol{\theta}}(\boldsymbol{x}_i), g(\boldsymbol{x}_i))}$$

- Out-of-sample Error: Suppose data $\boldsymbol{x}$ follows a certain distribution $\mathcal{D}$ in an i.i.d. manner,

$$\boxed{\mathrm{Er}_{\mathrm{out}} = \mathbb{E}_{\boldsymbol{x} \sim \mathcal{D}} \left[e(f_{\boldsymbol{\theta}}(\boldsymbol{x}), g(\boldsymbol{x})) \right]}$$

Remarks:

- The In-sample error $\mathrm{Er}_{\mathrm{in}}$ is also known as the training error.
- The out-of-sample error $\mathrm{Er}_{\mathrm{out}}$ is more general than the test error. Fortunately, we can use the test error to approximate $\mathrm{Er}_{\mathrm{out}}$ very well when the test dataset is large enough.

Recap: Concept of Training versus Testing / Generalization

Recall that learning is all about to infer g outside of the seen training dataset, i.e.,

Make the out-of-sample error small

- But $\mathrm{Er}_{\mathrm{out}}$ is not computable at all.

Here is a simple decomposition:

$$\mathrm{Er}_{\mathrm{out}} = \underbrace{\mathrm{Er}_{\mathrm{out}} - \mathrm{Er}_{\mathrm{in}}}_{generalization\ error} + \underbrace{\mathrm{Er}_{\mathrm{in}}}_{traning\ error}$$

- We need to simultaneously make generalization error and training error small, in order to make $\mathrm{Er}_{\mathrm{out}}$ small.

The goal of generalization is to

explore how out-of-sample error is related to in-sample error, i.e., bound the generalization error

- The reason to explore this relationship is that $\mathrm{Er}_{\mathrm{in}}$ is computable, checkable, and even amenable.

Recap: The Starting Point

Given training samples $\mathcal{S} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

In expectation for any $f \in \mathcal{H}$: (we omit $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ for simplicity)

$$\boxed{\mathbb{E}_{\mathcal{S} \sim_{i.i.d.} \mathcal{D}} [\mathrm{Er}_{\mathrm{in}}(f)] = \mathrm{Er}_{\mathrm{out}}(f).}$$

- Law of large numbers: When $n \to \infty$, we have in-sample error estimates the out-of-sample error accurately.
- However, this is true only when $n \to \infty$.

$\rightsquigarrow$ We will derive a non-asymptotic (finite n) result for $f \approx g$, which is analogous to the one derived for $\nu \approx \mu$.

Finite Hypothesis Space Generalization

VC Dimension

Additional Assumption: Finite Hypothesis Space

Assumption: Finite hypothesis space

$$\text{The cardinality } |\mathcal{H}| < +\infty$$

where $|\cdot|$ denotes the cardinality (number of elements) of a set. Namely, $|\mathcal{H}|$ measures the number of all possible $f_{\boldsymbol{\theta}} \in \mathcal{H}$.

- It means that the number of possible $f_{\boldsymbol{\theta}}$ in $\mathcal{H}$ is finite. Is it a practical assumption?
- We will omit $\boldsymbol{\theta}$ in $f_{\boldsymbol{\theta}}$ in the sequel to ease notation. Remind yourself f is almost always parameterized by some parameter $\boldsymbol{\theta}$.

Generalization for Fixed f: A Lemma

Lemma: High probability bounds for fixed f

Fix any model $f : \mathcal{X} \mapsto \{-1, 1\}$ ($f \in \mathcal{H}$ is fixed). The following inequalities hold for any $t > 0$:

$$\Pr\left[\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f) \geq t\right] \leq e^{-2nt^2},$$

and

$$\Pr\left[\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f) \leq -t\right] \leq e^{-2nt^2}.$$

Thus, we have the two side tail probability bound

$$\boxed{\Pr\left[|\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f)| \geq t\right] \leq 2e^{-2nt^2}.}$$

- Non-asymptotic bounds valid for any n.
- Equivalently, $\Pr\left[|\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f)| \leq t\right] \geq 1 - 2e^{-2nt^2}$, which is a high probability bound.

Proof

Recall the Hoeffding's inequality:

Hoeffding's inequality for bounded random variables

Suppose X_i are independent random variables with mean μ_i and bounded on $[a_i, b_i]$ for $i = 1, \cdots, n$, then for any $t > 0$, we have

$$\boxed{\Pr\left[\sum_{i=1}^{n}(X_i - \mu_i) \geq t\right] \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}}}$$

Note that $e(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i))$ equals to either 0 or 1, we have

$$\begin{aligned}
&\Pr\Big[\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f) \geq t\Big] \\
&= \Pr\left[\frac{1}{n}\sum_{i=1}^{n}(e(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} e(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i))\right] \geq t\right] \\
&= \Pr\left[\sum_{i=1}^{n}(e(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i)) - \mathbb{E}[e(f(\boldsymbol{x}_i), g(\boldsymbol{x}_i))]) \geq nt\right] \\
&\leq e^{-2nt^2}.
\end{aligned}$$

Generalization for Fixed Model f

Proposition: Generalization for fixed f

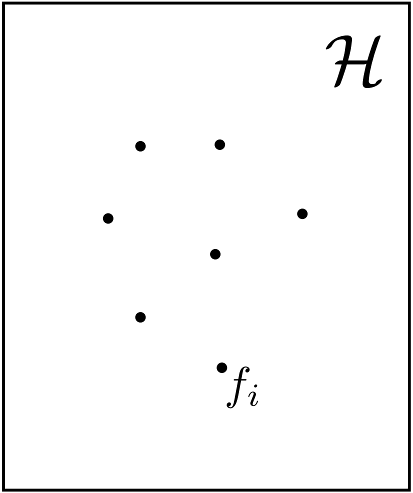
Fix a model $f : \mathcal{X} \mapsto \{-1, 1\}$. For any $\delta > 0$, the following generalization bound holds with probability at least $1 - \delta$

$$\boxed{\mathrm{Er}_{\mathrm{out}}(f) \le \mathrm{Er}_{\mathrm{in}}(f) + \sqrt{\frac{\log\left(\frac{2}{\delta}\right)}{2n}}}$$

Proof: Letting $\delta = 2e^{-2nt^2}$ and solving for t yields the desired result.

Are we done?

What We Need is Uniform Bound for All Possible $f \in \mathcal{H}$



Generalization for Finite Hypothesis Space

Theorem: Generalization for finite hypothesis space

Let $\mathcal{H}$ be a finite hypothesis space, i.e., $|\mathcal{H}| < \infty$. For any $\delta > 0$, the following generalization bound holds with probability at least $1 - \delta$

$$\boxed{\forall f \in \mathcal{H} \qquad \mathrm{Er}_{\mathrm{out}}(f) \leq \mathrm{Er}_{\mathrm{in}}(f) + \sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2n}}} \tag{1}$$

- The dependence on δ is only logarithmically.
- The generalization error increase when $|\mathcal{H}|$ grows, but only logarithmically.
- More samples (larger n) lead to better generalization (always true in practice).

Trade-off

$$\boxed{\forall f \in \mathcal{H} \qquad \mathrm{Er}_{\mathrm{out}}(f) \leq \mathrm{Er}_{\mathrm{in}}(f) + \sqrt{\frac{\log\left(\frac{2|\mathcal{H}|}{\delta}\right)}{2n}}}$$

On the training side, we need

more complex hypothesis $\mathcal{H}$ (larger $|\mathcal{H}|$)

On the generalization side, we need

less complex hypothesis $\mathcal{H}$ (smaller $|\mathcal{H}|$)

Proof

The proof is done by applying union bound. Let $f_1, \cdots, f_{|\mathcal{H}|}$ be the elements of $\mathcal{H}$. We have

$$\begin{aligned}
&\Pr\left[\exists f \in \mathcal{H} \ s.t. \ |\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f)| \geq t\right] \\
&\quad = \Pr\Big[|\,\mathrm{Er}_{\mathrm{in}}(f_1) - \mathrm{Er}_{\mathrm{out}}(f_1)| \geq t, \textbf{or} \cdots \\
&\qquad\qquad \textbf{or } |\,\mathrm{Er}_{\mathrm{in}}(f_{|\mathcal{H}|}) - \mathrm{Er}_{\mathrm{out}}(f_{|\mathcal{H}|})| \geq t\Big] \\
&\quad \overset{\text{union bound}}{\leq} \sum_{i=1}^{|\mathcal{H}|} \Pr\left[|\,\mathrm{Er}_{\mathrm{in}}(f_i) - \mathrm{Er}_{\mathrm{out}}(f_i)| \geq t\right] \\
&\quad \leq 2|\mathcal{H}|e^{-2nt^2}
\end{aligned}$$

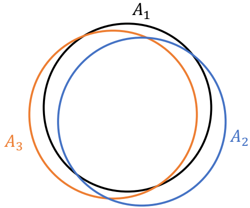
Setting the RHS probability to δ yields the desired result.

Issue with Finite Hypothesis Space

However, $|\mathcal{H}|$ is usually infinite in practice

- Think about the the simplest (one-dimensional) linear model $f_\theta(x) = \theta x$, $\theta \in \mathbb{R}$. We have infinitely many choices for θ.
- If we let $|\mathcal{H}|$ tends to ∞, we will have a trivial bound: Some finite thing $< \infty$.

Issue: Union bound can be very coarse.



In this case, $\Pr[A_1 \text{ or } A_2 \text{ or } A_3]$ is much less than $\sum_{i=1}^{3} \Pr[A_i]$.

Solution: Find a way to measure the complexity of $\mathcal{H}$ more smartly.

Finite Hypothesis Space Generalization

VC Dimension

Road Map

- Instead of using $|\mathcal{H}|$ directly to count the complexity of $\mathcal{H}$, we have to properly account for the overlaps of different hypotheses.
- In this way, our goal is to replace the number of hypotheses $|\mathcal{H}|$ with an effective number which is finite even when $|\mathcal{H}|$ is infinite.

$\rightsquigarrow$ This quantity will be the so-called VC dimension, which is of combinatorial nature.

- The VC dimension captures how different the hypotheses in $\mathcal{H}$ are, and hence how much overlap the different hypotheses have.
- Using this new notion, we will show that we can replace $|\mathcal{H}|$ in the obtained generalization bound with VC dimension.

Before that, let us introduce the related notion called dichotomy and growth function.

Dichotomy

- If $f \in \mathcal{H}$ is applied to a finite sample set $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, we get n-tuple $\{f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)\}$ of ± 1's.
- Such a n tuple is called a dichotomy since it splits $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$ into two groups: The points for which f is $+1$ and those for which f is -1.
- Each $f \in \mathcal{H}$ generates a dichotomy on $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, but two different f's may generate the same dichotomy.

We can now define the dichotomies of the whole hypothesis space $\mathcal{H}$.

Dichotomies of $\mathcal{H}$

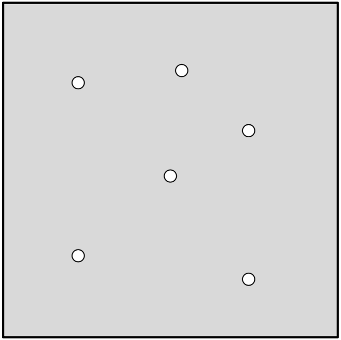
Given $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_n\}$, the dichotomies generated by $\mathcal{H}$ on these points are defined by

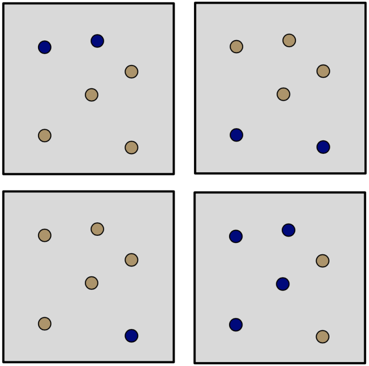
$$\mathcal{H}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n) = \{(f(\boldsymbol{x}_1), \dots, f(\boldsymbol{x}_n)) : f \in \mathcal{H}\}.$$

- One can think of the dichotomies $\mathcal{H}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ as a set of hypothesis just like $\mathcal{H}$ is, except that the hypotheses are only seen through the n data points.
- Larger $\mathcal{H}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)$ means that $\mathcal{H}$ is more diverse / rich.

Example of Dichotomies

We have five points. There are four different $f \in \mathcal{H}$ on the points.





Growth Function

Growth function is a number, which is defined based on the number of dichotomies.

Growth function

The growth function for the hypothesis set $\mathcal{H}$ is defined as:

$$\mathcal{G}_{\mathcal{H}}(n) = \max_{\{\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n\} \subseteq \mathcal{X}} \left|\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)\right|,$$

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

Instead of counting the size of $\mathcal{H}$ by $|\mathcal{H}|$, the idea of growth function is: Using $\mathcal{H}$, what is the maximum number of ways we can label a n-points dataset?

Properties of Growth Function

- $G_{\mathcal{H}}(n)$ counts the most dichotomies that can possibly be generated on any n points in $\mathcal{X}$.
- To compute $\mathcal{G}_{\mathcal{H}}(n)$, we consider all possible choice of n points, and pick the one that gives us the most dichotomies, which is of combinatorial nature.
- Similar to $|\mathcal{H}|$, $\mathcal{G}_{\mathcal{H}}(n)$ is a measure of the richness of the hypothesis set $\mathcal{H}$. The difference is that it is now considered on n points rather than the entire input space $\mathcal{X}$.
- Since $\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) \subset \{-1, +1\}^n$ (the set of all possible dichotomies on any n points). Clearly, we have

$$\mathcal{G}_{\mathcal{H}}(n) \leq 2^n.$$

- If $\mathcal{H}$ is capable of generating all possible dichotomies on $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, then $\mathcal{H}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n) = \{-1, +1\}^n$, i.e., $\mathcal{G}_{\mathcal{H}}(n) = 2^n$, and we say that $\mathcal{H}$ can shatter the data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$.

Vapnik-Chervonenkis (VC) Dimension

We now introduce a well known notion — Vapnik-Chervonenkis (VC) dimension.

VC dimension

The VC dimension of a hypothesis space $\mathcal{H}$, denoted by $d_{\mathsf{VC}}(\mathcal{H})$ or simply d_{VC}, is the largest n so that it can be shattered by $\mathcal{H}$, i.e.,

$$d_{\mathsf{VC}}(\mathcal{H}) := \max\{n : \mathcal{G}_{\mathcal{H}}(n) = 2^n\}.$$

If $\mathcal{G}_{\mathcal{H}}(n) = 2^n$ for all n, then $d_{\mathsf{VC}}(\mathcal{H}) = \infty$.

- By definition, VC dimension indicates the representation power of $\mathcal{H}$.
- $d_{\mathsf{VC}} + 1$ counts the number of data points n that $\mathcal{H}$ starts to not shatter.

Fact:

Bounding Growth Function using VC Dimension

$$\boxed{\mathcal{G}_{\mathcal{H}}(n) \leq n^{d_{\mathsf{VC}}} + 1.}$$

$\rightsquigarrow$ Next lecture: VC dimension generalization.