Statistical Learning Theory

3. Vapnik-Chervonenkis (VC) theory

Yongdai Kim

Definition (Bounded Differences Condition)
Let $g : \mathcal{X} \to \mathbb{R}$ and constants c_i be given. Then g is said to satisfy the bounded differences condition (with constants c_i) if

$$\sup_{x_1,\ldots,x_n,x_i'} \left| g\left(x_1,\ldots,x_n\right) - g\left(x_1,\ldots,x_i',\ldots,x_n\right) \right| \leq c_i$$

for every i.

Theorem (Bounded Differences Inequality)
Suppose that $X_1, \ldots, X_n$ are indepent random variables. If $g : \mathcal{X} \to \mathbb{R}$ satisfies the bounded differences condition, then

$$\mathbb{P}\left[\mid g\left(X_1, \ldots, X_n\right) - \mathbb{E}\left[g\left(X_1, \ldots, X_n\right) \mid > t\right] \leq 2 \exp\left(-\frac{2t^2}{\sum_i c_i^2}\right)$$

Empirical measure

- The upper bounds proved so far are meaningful only for a finite dictionary $\mathcal{H}$, because if $M = |\mathcal{H}|$ is infinite all of the bounds we have will simply be infinity.
- To extend previous results to the infinite case, we essentially need the condition that only a finite number of elements in an infinite dictionary $\mathcal{H}$ really matter.
- This is the objective of the VapnikChervonenkis (VC) theory which was developed in 1971 .

Empirical measure

- Recall that the key quantity we need to control is

$$2 \sup_{h \in \mathcal{H}} \left(\hat{R}_n(h) - R(h) \right).$$

- Instead of the union bound which would not work in the infinite case, we seek some bound that potentially depends on n and the complexity of the set $\mathcal{H}$.
- One approach is to consider some metric structure on $\mathcal{H}$ and hope that if two elements in $\mathcal{H}$ are close, then the quantity evaluated at these two elements are also close.
- On the other hand, the VC theory is more combinatorial and does not involve any metric space structure as we will see.

Empirical measure

- By definition

$$\hat{R}_n(h) - R(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{I}\left(h\left(X_i\right) \neq Y_i\right) - \mathbb{E}\left[\mathbb{I}\left(h\left(X_i\right) \neq Y_i\right)\right]\right)$$

- Let $Z = (X, Y)$ and $Z_i = (X_i, Y_i)$, and let $\mathcal{A}$ denote the class of measurable sets in the sample space $\mathcal{X} \times \{0, 1\}$.
- For a classifier h, define $A_h \in \mathcal{A}$ by

$$\{Z_i \in A_h\} = \{h\left(X_i\right) \neq Y_i\}$$

Empirical measure

- Moreover, define measures μ_n and μ on $\mathcal{A}$ by

$$\mu_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Z_i \in A\right) \text{ and } \mu(A) = \mathbb{P}\left[Z_i \in A\right]$$

  for $A \in \mathcal{A}$.

- With this notation, we have proved that

$$\sup_{h\in\mathcal{H}} \hat{R}_n(h) - R(h) = \sup_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)| \leq \sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{2n}}$$

Empirical measure

- Since this is not accessible in the infinite case, we will derive an upper bound by use of bounded differences inequality.
- If we change the value of only one z_i in the function

$$z_1, \ldots, z_n \mapsto \sup_{A \in A} |\mu_n(A) - \mu(A)|,$$

  the value of the function will differ by at most $1/n$.
- Hence it satisfies the bounded difference assumption with $c_i = 1/n$ for all $1 \leq i \leq n$.
- Applying the bounded difference inequality, we get that

$$\left| \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| - \mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|] \right| \leq \sqrt{\frac{\log(2/\delta)}{2n}}.$$

  with probability at least $1 - \delta$.

Symmetrization and Rademacher complexity

- We will drive an upper bound of $\mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|]$, and symmetrization is a frequently used technique for this purpose.
- Let $\mathcal{D} = \{Z_1, \ldots, Z_n\}$ be the sample set.
- To employ symmetrization, we take another independent copy of the sample set $\mathcal{D}' = \{Z_1', \ldots, Z_n'\}$.
- This sample only exists for the proof, so it is sometimes referred to as a ghost sample.

Symmetrization and Rademacher complexity

- Then we have

$$\begin{aligned}
\mu(A) &= \mathbb{P}[Z \in A] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Z_i' \in A\right)\right] \\
&= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Z_i' \in A\right) \mid \mathcal{D}\right] \\
&= \mathbb{E}\left[\mu_n'(A) \mid \mathcal{D}\right]
\end{aligned}$$

where $\mu_n' := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left(Z_i' \in A\right)$.

Symmetrization and Rademacher complexity

Thus by Jensen's inequality,

$$\begin{aligned}
\mathbb{E}\left[\sup_{A\in\mathcal{A}} \mid \mu_n(A) - \mu(A)|\right] &= \mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \mathbb{E}\left[\mu'_n(A) \mid \mathcal{D}\right]\right|\right] \\
&\le \mathbb{E}\left[\sup_{A\in\mathcal{A}} \mathbb{E}\left[\left|\mu_n(A) - \mu'_n(A)\right| \mid \mathcal{D}\right]\right] \\
&\le \mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\mu_n(A) - \mu'_n(A)\right|\right] \\
&= \mathbb{E}\left[\sup_{A\in\mathcal{A}} \left|\frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{I}\left(Z_i \in A\right) - \mathbb{I}\left(Z'_i \in A\right)\right)\right|\right].
\end{aligned}$$

Symmetrization and Rademacher complexity

- Since $\mathcal{D}'$ has the same distribution of $\mathcal{D}$, by symmetry $\mathbb{I}(Z_i \in A) - \mathbb{I}(Z_i' \in A)$ has the same distribution as $\sigma_i(\mathbb{I}(Z_i \in A) - \mathbb{I}(Z_i' \in A))$ where $\sigma_1, \ldots, \sigma_n$ are i.i.d. $\text{Rad}\left(\frac{1}{2}\right)$, i.e.

$$\mathbb{P}[\sigma_i = 1] = \mathbb{P}[\sigma_i = -1] = \frac{1}{2}$$

  and σ_i 's are taken to be independent of both samples.
- Therefore,

$$\begin{aligned}
&\mathbb{E}[\sup_{A \in \mathcal{A}} \mid \mu_n(A) - \mu(A)|] \\
&\leq \mathbb{E}\left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \left(\mathbb{I}(Z_i \in A) - \mathbb{I}\left(Z_i' \in A\right)\right)\right|\right] \\
&\leq 2\mathbb{E}\left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{I}(Z_i \in A)\right|\right] \cdots (*)
\end{aligned}$$

Symmetrization and Rademacher complexity

- Using symmetrization we have bounded $\mathbb{E}\left[\sup_{A \in A} |\mu_n(A) - \mu(A)|\right]$. by a much nicer quantity.
- Yet we still need an upper bound of the last quantity that depends only on the structure of $\mathcal{A}$ but not on the random sample $\{Z_i\}$.
- This is achieved by taking the supremum over all $z_i \in \mathcal{X} \times \{0, 1\} =: \mathcal{Y}$.

Symmetrization and Rademacher complexity

Definition
The Rademacher complexity of a family of sets $\mathcal{A}$ in a space $\mathcal{Y}$ is defined to be the quantity

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\ldots,\bar{z}_n \in \mathcal{Y}} \mathbb{E}\left[\sup_{A \in \mathcal{A}} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i \mathbb{I}\left(z_i \in A\right) \right| \right].$$

The Rademacher complexity of a set $B \subset \mathbb{R}^n$ is defined to be

$$\mathcal{R}_n(B) = \mathbb{E}\left[\sup_{b \in B} \left| \frac{1}{n} \sum_{i=1}^{n} \sigma_i b_i \right| \right].$$

- We conclude from (*) and the definition that

$$\mathbb{E}[\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|] \leq 2\mathcal{R}_n(\mathcal{A}).$$

Symmetrization and Rademacher complexity

- In the definition of Rademacher complexity of a set, the quantity $\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i b_i\right|$ measures how well a vector $b \in B$ correlates with a random sign pattern $\{\sigma_i\}$.
- The more complex B is, the better some vector in B can replicate a sign pattern.
- In particular, if B is the full hypercube $[-1, 1]^n$, then $\mathcal{R}_n(B) = 1$.
- However, if $B \subset [-1, 1]^n$ contains only k -sparse vectors, then $\mathcal{R}_n(B) = k/n$.
- Hence $\mathcal{R}_n(B)$ is indeed a measurement of the complexity of the set B.

Symmetrization and Rademacher complexity

- The set of vectors to our interest in the definition of Rademacher complexity of $\mathcal{A}$ is

$$T(z) := \left\{ \left(\mathbb{I}\left(z_1 \in A\right), \ldots, \mathbb{I}\left(z_n \in A\right) \right)^T, A \in \mathcal{A} \right\}.$$

- Thus the key quantity here is the cardinality of $T(z)$, i.e., the number of sign patterns these vectors can replicate as A ranges over $\mathcal{A}$.
- Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is bounded by 2^n.

Shattering

- We will complete the analysis of the performance of the empirical risk minimizer under a constraint on the VC dimension of the family of classifiers.
- To that end, we will see how to control Rademacher complexities using shatter coefficients.
- Moreover, we will see how the problem of controlling uniform deviations of the empirical measure μ_n from the true measure μ as done by Vapnik and Chervonenkis relates to our original classification problem.

Shattering

- Recall from the previous slide that we are interested in sets of the form

$$T(z) := \{(\mathbb{I}(z_1 \in A), \ldots, \mathbb{I}(z_n \in A)), A \in \mathcal{A}\}, z = (z_1, \ldots, z_n) \cdots (**)$$

- In particular, the cardinality of $T(z)$, i.e., the number of binary patterns these vectors can replicate as A ranges over $\mathcal{A}$, will be of critical importance, as it will arise when controlling the Rademacher complexity.
- Although the cardinality of $\mathcal{A}$ may be infinite, the cardinality of $T(z)$ is always at most 2^n.
- When it is of the size 2^n, we say that $\mathcal{A}$ shatters the set $z_1, \ldots, z_n$. Formally, we have the following definition.

Shattering

Definition
A collection of sets $\mathcal{A}$ shatters the set of points $\{z_1, z_2, \ldots, z_n\}$

$$\text{card}\left\{\left(\mathbb{I}\left(z_1 \in A\right), \ldots, \mathbb{I}\left(z_n \in A\right)\right), A \in \mathcal{A}\right\} = 2^n$$

- The sets of points $\{z_1, z_2, \ldots, z_n\}$ that we are interested are realizations of the pairs $Z_1 = \left(X_1, Y_1\right), \ldots, Z_n = \left(X_n, Y_n\right)$ and may, in principle take any value over the sample space.
- Therefore, we define the shatter coefficient to be the largest cardinality that we may obtain.

Shattering

Definition
The shatter coefficients of a class of sets $\mathcal{A}$ is the sequence of numbers $\{\mathcal{S}_{\mathcal{A}}(n)\}_{n\geq 1}$, where for any $n \geq 1$

$$\mathcal{S}_{\mathcal{A}}(n) = \sup_{z_1,\dots,z_n} \operatorname{card}\{(\mathbb{I}(z_1 \in A), \dots, \mathbb{I}(z_n \in A)), A \in \mathcal{A}\}$$

and the suprema are taken over the whole sample space.

Shattering

- By definition, the n th shatter coefficient $\mathcal{S}_{\mathcal{A}}(n)$ is equal to 2^n if there exists a set $\{z_1, z_2, \ldots, z_n\}$ that $\mathcal{A}$ shatters.
- The largest of such sets is precisely the Vapnik-Chervonenkis or VC dimension.

Definition
The Vapnik-Chervonenkis dimension, or *VC* -dimension of $\mathcal{A}$ is the largest integer d such that $\mathcal{S}_{\mathcal{A}}(d) = 2^d$. We write $\mathrm{VC}(\mathcal{A}) = d$.
If $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for all positive integers n, then $\mathrm{VC}(\mathcal{A}) := \infty$

Shattering

- In other words, $\mathcal{A}$ shatters some set of points of cardinality d but shatters no set of points of cardinality $d + 1$.
- In particular, $\mathcal{A}$ also shatters no set of points of cardinality $d' \geq d$ so that the VC dimension is well defined.
- In the sequel, we will see that the VC dimension will play the role similar to of cardinality, but on an exponential scale.
- For interesting classes $\mathcal{A}$ such that $\text{card}(\mathcal{A}) = \infty$, we also may have $\text{VC}(\mathcal{A}) < \infty$.

Shattering

- For example, assume that $\mathcal{A}$ is the class of half-lines, $\mathcal{A} = \{(-\infty, a], a \in \mathbb{R}\} \cup \{[a, \infty), a \in \mathbb{R}\}$, which is clearly infinite.
- Then, we can clearly shatter a set of size 2 but we for three points $z_1, z_2, z_3, \in \mathbb{R}$, if for example $z_1 < z_2 < z_3$, we cannot create the pattern $(0, 1, 0)$ (see Figure 1 in the next slide).
- Indeed, half lines can can only create patterns with zeros followed by ones or with ones followed by zeros but not an alternating pattern like $(0, 1, 0)$.

Shattering

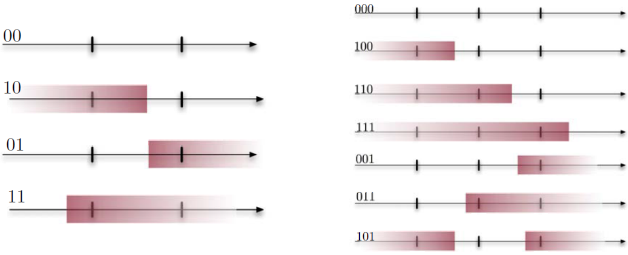

Figure 1 : If $\mathcal{A} = \{$ halflines $\}$, then any set of size $n = 2$ is shattered because we can create all $2^n = 40/1$ patterns (left); if $n = 3$ the pattern $(0, 1, 0)$ cannot be reconstructed: $\mathcal{S}_A(3) = 7 < 2^3$ (right). Therefore, $\mathrm{VC}(\mathcal{A}) = 2$

The VC inequality

- We have now introduced all the ingredients necessary to state the main result of this section: the VC inequality.

Theorem (VC inequality)
For any family of sets $\mathcal{A}$ *with* VC *dimension* $\mathrm{VC}(\mathcal{A}) = d$*, it holds*

$$\mathbb{E} \sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2d \log(2en/d)}{n}}$$

- Note that this result holds even if $\mathcal{A}$ is infinite as long as its VC dimension is finite.
- Moreover, observe that $\log(|\mathcal{A}|)$ has been replaced by a term of order $d \log(2en/d)$.
- To prove the VC inequality, we proceed in three steps:

The VC inequality

1. Symmetrization, to bound the quantity of interest by the Rademacher complexity:

$$\mathbb{E}[\sup_{A \in \mathcal{A}} \mid \mu_n(A) - \mu(A)|] \leq 2\mathcal{R}_n(\mathcal{A}).$$

We have already done this step in the previous lecture.

2. Control of the Rademacher complexity using shatter coefficients. We are going to show that

$$\mathcal{R}_n(\mathcal{A}) \leq \sqrt{\frac{2 \log (2\mathcal{S}_{\mathcal{A}}(n))}{n}}$$

3. We are going to need the Sauer-Shelah lemma to bound the shatter coefficients by the VC dimension. It will yield

$$\mathcal{S}_{\mathcal{A}}(n) \leq \left(\frac{en}{d}\right)^d, \quad d = \mathrm{VC}(\mathcal{A})$$

Put together, these three steps yield the VC inequality.

STEP 2: CONTROL OF THE RADEMACHER COMPLEXITY

- We need the following Lemma whose proof is HW.

Lemma
For any $B \subset \mathbb{R}^n$, such that $|B| < \infty$:, it holds

$$\mathcal{R}_n(B) = \mathbb{E}\left[\max_{b \in B}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i b_i\right|\right] \leq \max_{b \in B}|b|_2 \frac{\sqrt{2\log(2|B|)}}{n}$$

where $|\cdot|_2$ denotes the Euclidean norm.

STEP 2: CONTROL OF THE RADEMACHER COMPLEXITY

- We apply the above Lemma to our problem by observing that

$$\mathcal{R}_n(\mathcal{A}) = \sup_{z_1,\ldots,z_n} \mathcal{R}_n(T(z)).$$

- In particular, since $T(z) \subset \{0,1\}^n$, we have $|b|_2 \leq \sqrt{n}$ for all $b \in T(z)$.
- Moreover, by definition of the shatter coefficients, $|T(z)| \leq \mathcal{S}_{\mathcal{A}}(n)$.
- Together with the above lemma, it yields the desired inequality:

$$\mathcal{R}_n(\mathcal{A}) \leq \sqrt{\frac{2\log\left(2S_{\mathcal{A}}(n)\right)}{n}}.$$

STEP 3: SAUER-SHELAH LEMMA

- We need to use a lemma from combinatorics to relate the shatter coefficients to the VC dimension.
- A priori, it is not clear from its definition that the VC dimension may be at all useful to get better bounds.
- Recall that steps 1 and 2 put together yield the following bound

$$\mathbb{E}[\sup_{A \in A} | \mu_n(A) - \mu(A)|] \leq 2\sqrt{\frac{2\log(2\mathcal{S}_{\mathcal{A}}(n))}{n}} \cdots (***)$$

- In particular, if $\mathcal{S}_{\mathcal{A}}(n)$ is exponential in n, the bound (***) is not informative, i.e., it does not imply that the uniform deviations go to zero as the sample size n goes to infinity.

STEP 3: SAUER-SHELAH LEMMA

- The VC inequality suggest that this is not the case as soon as $\mathrm{VC}(\mathcal{A}) < \infty$ but it is not clear a priori.
- Indeed, it may be the case that $\mathcal{S}_{\mathcal{A}}(n) = 2^n$ for $n \leq d$ and $\mathcal{S}_{\mathcal{A}}(n) = 2^n - 1$ for $n > d$, which would imply that $\mathrm{VC}(\mathcal{A}) = d < \infty$ but that the right-hand side in (***) is larger than 2 for all n.
- It turns our that this can never be the case: if the VC dimension is finite, then the shatter coefficients are at most polynomial in n, which is stated in the Sauer-Shelah lemma.

STEP 3: SAUER-SHELAH LEMMA

Lemma (Sauer-Shelah)
If $\mathrm{VC}(\mathcal{A}) = d$*, then* $\forall n \geq 1$,

$$\mathcal{S}_{\mathcal{A}}(n) \leq \sum_{k=0}^{d} \binom{n}{k} \leq \left(\frac{en}{d}\right)^{d}$$

The VC inequality

To sum up everything, we have the following corollary.

Corollary (VC inequality)
For any family of sets $\mathcal{A}$ such that $\mathrm{VC}(\mathcal{A}) = d$ *and any* $\delta \in (0,1)$*,* *it holds with probability at least* $1-\delta$*,*

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| \leq 2\sqrt{\frac{2d \log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

Application to ERM

- The VC inequality provides an upper bound for $\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|$ in terms of the VC dimension of the class of sets $\mathcal{A}$.
- This result translates directly to our quantity of interest:

$$\sup_{h \in H} \left| \hat{R}_n(h) - R(h) \right| \leq 2\sqrt{\frac{2\,\mathrm{VC}(\mathcal{A}) \log\left(\frac{2en}{\mathrm{VC}(\mathcal{A})}\right)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

  where $\mathcal{A} = \{A_h : h \in \mathcal{H}\}$ and $A_h = \{(x, y) \in \mathcal{X} \times \{0, 1\} : h(x) \neq y\}$.
- Unfortunately, the VC dimension of this class of subsets of $\mathcal{X} \times \{0, 1\}$ is not very natural.
- Since, a classifier h is a $\{0, 1\}$ valued function, it is more natural to consider the VC dimension of the family

$$\overline{\mathcal{A}} = \{\{h = 1\} : h \in \mathcal{H}\} = \{A : \exists h \in \mathcal{H}, h(\cdot) = \mathbb{I}(\cdot \in A)\}.$$

Application to ERM

Definition
We define the VC dimension $\mathrm{VC}(\mathcal{H})$ of $\mathcal{H}$ to be the VC dimension of $\overline{\mathcal{A}}$.

- It is not clear how $\mathsf{VC}(\overline{\mathcal{A}})$ relates to the quantity $\mathsf{VC}(\mathcal{A})$.
- Fortunately, these two are actually equal as indicated in the following lemma.

Lemma
Define the two families for sets: $\mathcal{A} = \{A_h : h \in \mathcal{H}\} \in 2^{\mathcal{X}\times\{0,1\}}$ *where* $A_h = \{(x, y) \in \mathcal{X} \times \{0, 1\} : h(x) \neq y\}$ *and* $\bar{A} = \{\{h = 1\} : h \in \mathcal{H}\} \in 2^{\mathcal{X}}$ *Then,* $\mathcal{S}_{\mathcal{A}}(n) = \mathcal{S}_{\overline{\mathcal{A}}}(n)$ *for all* $n \geq 1$. *It implies* $\mathsf{VC}(\mathcal{A}) = \mathsf{VC}(\overline{\mathcal{A}})$.

Application to ERM

It yields the following corollary to the VC inequality.

Corollary
Let $\mathcal{H}$ be a family of classifiers with VC dimension d. Then the empirical risk classifier $\hat{h}^{erm}$ over $\mathcal{H}$ satisfies

$$R\left(\hat{h}^{\mathrm{erm}}\right) \leq \min_{h \in \mathcal{H}} R(h) + 4\sqrt{\frac{2d\log(2en/d)}{n}} + \sqrt{\frac{\log(2/\delta)}{2n}}$$

with probability $1-\delta$.