Goal: understand generalization. How much training data is needed?

\[ \text{Goal: understand generalization. How much training data is needed?} \]

Lecture 36

Combinatorics (VC dimension)  ↓  Machine learning

VC theory

Discrete math

Combinatorial ingredient:

Lemma [Papya 85] A finite class of boolean functions \( \mathcal{H} \) on \( X \),
\[ |\mathcal{H}| \leq \#(\text{subsets of } X \text{ shattered by } \mathcal{H}) \]

Convention: \( \emptyset \) is shattered by \( \mathcal{H} \) whenever \( \mathcal{H} \) is nonempty.

Proof WLOG, \( X = \{1, \ldots, n\} \). Denote by \( \text{sh}(\mathcal{H}) \) the family of all subsets of \( X \) shattered by \( \mathcal{H} \). To prove
\[ |\mathcal{H}| \leq |\text{sh}(\mathcal{H})|, \]

- partition \( \mathcal{H} \) according to the value at point \( n \), i.e.
\[ \mathcal{H} = \mathcal{H}_0 \cup \mathcal{H}_1 \]

where \( \mathcal{H}_0 = \{ h \in \mathcal{H} : h(n) = 0 \} \) and \( \mathcal{H}_1 = \{ h \in \mathcal{H} : h(n) = 1 \} \).

A subset \( \{i_1, \ldots, i_d\} \subset X \) shattered by \( \mathcal{H}_0 \) or \( \mathcal{H}_1 \) is also shattered by \( \mathcal{H} \). Thus
\[ |\text{sh}(\mathcal{H})| \geq |\text{sh}(\mathcal{H}_0)| + |\text{sh}(\mathcal{H}_1)| \quad (\star) \]

\( \star \) domain = \{1, \ldots, n-1\}

- Iterate: partition \( \mathcal{H}_0 \) and \( \mathcal{H}_1 \) according to the value \( h(n-1) \):
\[ \geq |\text{sh}(\mathcal{H}_{00})| + |\text{sh}(\mathcal{H}_{01})| + |\text{sh}(\mathcal{H}_{10})| + |\text{sh}(\mathcal{H}_{11})| \quad (\star) \]

... down to single-point classes, each of which shatters one set \( \emptyset \ldots \geq |\mathcal{H}| \)
\textbf{MISTAKE:} we double counted in (*)

He sets that are shattered
by both $H_0$ and $H_1$

\textbf{FIX:} Suppose \{i_1, \ldots, i_d\} is shattered by both $H_0$ and $H_1 \Rightarrow$

\begin{enumerate}
  \item A label assignment $y_1, \ldots, y_d \in \{0, 1\}$
  \item \exists h \in H_0: \quad h(i_1) = y_{i_1}, \ldots, h(i_d) = y_{i_d}, \quad h(n) = 0
  \item \exists g \in H_1: \quad g(i_1) = y_{i_1}, \ldots, g(i_d) = y_{i_d}, \quad g(n) = 1

  \Rightarrow \{i_1, \ldots, i_d, n\} is shattered by \mathcal{H} = H_0 \cup H_1 \quad \text{(choose either } h \text{)}
\end{enumerate}

This set is NOT shattered by either $H_0$ or $H_1 \Rightarrow$ it was \textbf{NOT} counted before

\Rightarrow \quad \text{a set that we double counted, we find a set we never counted}

\Rightarrow (\ast) is true. Proceed as before. QED
By def of vc dimension, a subset shattered by \( \mathcal{H} \) has cardinality \( \leq \text{vc}(\mathcal{H}) =: d \). So Pajor’s lemma yields

\[
|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k}
\]

Cor (Sauer-Shelah lemma) let \( \mathcal{H} \) be a class of Boolean functions on an \( n \)-point domain. Then

\[
|\mathcal{F}| \leq \sum_{k=0}^{d} \binom{n}{k} \quad \text{where} \quad d = \text{vc}(\mathcal{H})
\]

Examples

(a) Integer intervals: \( \mathcal{F} = \{ ([a, b]) : 1 \leq a \leq b \leq n \} \)

\[\text{vc}(\mathcal{F}) = 2 \quad (\text{as in the previous lecture for real intervals})\]

\[|\mathcal{F}| = 1 + n + \binom{n}{2} = \sum_{k=0}^{2} \binom{n}{k} \quad \Rightarrow \quad \text{Pajor lemma is sharp}\]

(b) \( \mathcal{F} = \{ \text{all functions on an } n \text{-point domain supported by } \leq d \text{ pts} \} \)

\[\text{vc}(\mathcal{F}) = d \quad (\text{HW?}) \quad \text{and} \quad |\mathcal{F}| = \sum_{k=0}^{d} \binom{n}{k} \quad \Rightarrow \quad \text{sharp again!}\]

Remarks

1. \( \sum_{k=1}^{d} \binom{n}{k} \leq \left( \frac{en}{d} \right)^{d} \) (HW 3, Problem 3)

\[d \leq \log |\mathcal{F}| = d \log \left( \frac{en}{d} \right) \quad \text{where} \quad d = \text{vc}(\mathcal{F}).\]

2. Heuristically, \( \log |\mathcal{F}| = \# \text{bits to specify a function in } \mathcal{F} \)

\[d = \text{vc}(\mathcal{F}) \sim \# \text{parameters that describe functions in } \mathcal{F} \]

\[\Rightarrow \log |\mathcal{F}| \approx d \quad \text{is expected.}\]