Lecture 2 Part A: VC Dimension

Pranabendu Misra
Chennai Mathematical Institute

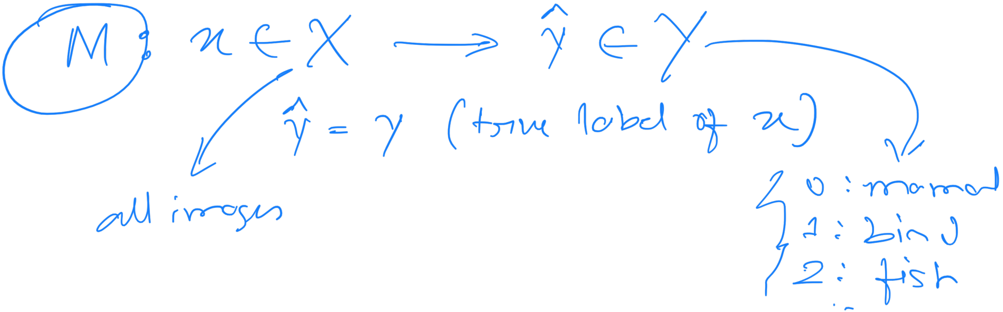
Advanced Machine Learning 2022

(based on slides by Madhavan Mukund)

Representational capacity

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.



- Training Set $S = \{(x, y) \mid x \in X\}$

$$S \longrightarrow M$$

- M "works well"

  given some $x \in X$ (st $x \notin S$)

  $M(x) = y$ (true label of x)

  with high probability

$S \longrightarrow$ Representation of the real data

Representational capacity

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln |\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\text{err}_D > \epsilon$ has training error $\text{err}_S > 0$.

Uniform convergence

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$. If a training set S of size $n \geq \frac{1}{2\epsilon^2}(\ln |\mathcal{H}| + \ln(2/\delta))$ is drawn using D, then with probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ satisfies $|\text{err}_S(h) - \text{err}_D(h)| \leq \epsilon$.

Representational capacity

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\text{err}_D > \epsilon$ has training error $\text{err}_S > 0$.

Uniform convergence

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$. If a training set S of size $n \geq \frac{1}{2\epsilon^2}(\ln|\mathcal{H}| + \ln(2/\delta))$ is drawn using D, then with probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ satisfies $|\text{err}_S(h) - \text{err}_D(h)| \leq \epsilon$.

- $|\mathcal{H}|$ is representational capacity, when $\mathcal{H}$ is finite

Representational capacity

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1-\delta$, every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.

Uniform convergence

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$. If a training set S of size $n \geq \frac{1}{2\epsilon^2}(\ln|\mathcal{H}| + \ln(2/\delta))$ is drawn using D, then with probability $\geq 1-\delta$, every $h \in \mathcal{H}$ satisfies $|\mathrm{err}_S(h) - \mathrm{err}_D(h)| \leq \epsilon$.

- $|\mathcal{H}|$ is representational capacity, when $\mathcal{H}$ is finite
- How do we adapt and apply these bounds when $\mathcal{H}$ is infinite?

Shattering

- Set system: $(X, \mathcal{H})$
  - X is a set — instance space
  - $\mathcal{H}$, set of subsets of X — set of possible classifiers / hypotheses

Shattering

- Set system: $(X, \mathcal{H})$
  - X is a set — instance space
  - $\mathcal{H}$, set of subsets of X — set of possible classifiers / hypotheses
- $A \subseteq X$ is shattered by $\mathcal{H}$ if every subset of A is given by $A \cap h$ for some $h \in \mathcal{H}$

Shattering

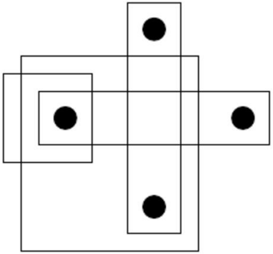
- Set system: $(X, \mathcal{H})$
  - X is a set — instance space
  - $\mathcal{H}$, set of subsets of X — set of possible classifiers / hypotheses
- $A \subseteq X$ is shattered by $\mathcal{H}$ if every subset of A is given by $A \cap h$ for some $h \in \mathcal{H}$
  - Every way of splitting A is captured by a hypothesis in $\mathcal{H}$
  - $2^{|A|}$ different subsets of A

Shattering

- Set system: $(X, \mathcal{H})$
  - X is a set — instance space
  - $\mathcal{H}$, set of subsets of X — set of possible classifiers / hypotheses
- $A \subseteq X$ is shattered by $\mathcal{H}$ if every subset of A is given by $A \cap h$ for some $h \in \mathcal{H}$
  - Every way of splitting A is captured by a hypothesis in $\mathcal{H}$
  - $2^{|A|}$ different subsets of A
- Example:
  - $X = \mathbb{R} \times \mathbb{R}$
  - $\mathcal{H}$: Axis-parallel rectangles
  - A : Four points forming a diamond

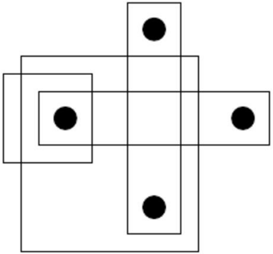
Shattering

- Set system: $(X, \mathcal{H})$
  - X is a set — instance space
  - $\mathcal{H}$, set of subsets of X — set of possible classifiers / hypotheses
- $A \subseteq X$ is shattered by $\mathcal{H}$ if every subset of A is given by $A \cap h$ for some $h \in \mathcal{H}$
  - Every way of splitting A is captured by a hypothesis in $\mathcal{H}$
  - $2^{|A|}$ different subsets of A
- Example:
  - $X = \mathbb{R} \times \mathbb{R}$
  - $\mathcal{H}$: Axis-parallel rectangles
  - A : Four points forming a diamond
  - $\mathcal{H}$ shatters A



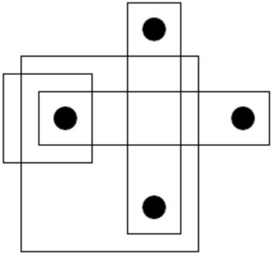
VC-Dimension [Vapnik-Chervonenkis]

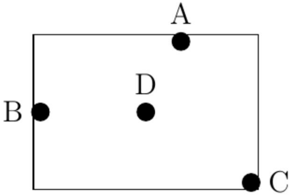
- VC-Dimension of $\mathcal{H}$ — size of the largest subset of X shattered by $\mathcal{H}$
  - For axis-parallel rectangles, VC-dimension is at least 4



VC-Dimension [Vapnik-Chervonenkis]

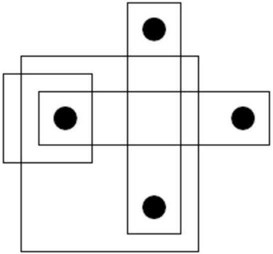
- VC-Dimension of $\mathcal{H}$ — size of the largest subset of X shattered by $\mathcal{H}$
  - For axis-parallel rectangles, VC-dimension is at least 4
  - Not a universal requirement — some sets of size 4 may not be shattered

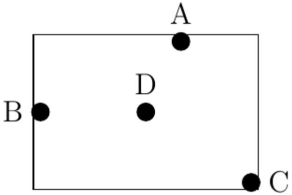




VC-Dimension [Vapnik-Chervonenkis]

- VC-Dimension of $\mathcal{H}$ — size of the largest subset of X shattered by $\mathcal{H}$
  - For axis-parallel rectangles, VC-dimension is at least 4
  - Not a universal requirement — some sets of size 4 may not be shattered
- No set of size 5 can be shattered by axis-parallel rectangles
  - Draw a bounding box rectangle — each edge touches a boundary point
  - At least one point lies inside the bounding box
  - Any set that includes the boundary points also includes the interior point





VC-Dimension, Examples

- Intervals of reals have VC-dimension 2
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \mid a \leq b \in \mathbb{R}\}$
  - Cannot shatter 3 points: consider subset with first and third point

VC-Dimension, Examples

- Intervals of reals have VC-dimension 2
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \mid a \leq b \in \mathbb{R}\}$
  - Cannot shatter 3 points: consider subset with first and third point
- Pairs of intervals of reals have VC-dimension 4
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \cup [c, d] \mid a \leq b, c \leq d \in \mathbb{R}\}$
  - Cannot shatter 5 points: consider subset with first, third and fifth point

VC-Dimension, Examples

- Intervals of reals have VC-dimension 2
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \mid a \leq b \in \mathbb{R}\}$
  - Cannot shatter 3 points: consider subset with first and third point
- Pairs of intervals of reals have VC-dimension 4
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \cup [c, d] \mid a \leq b, c \leq d \in \mathbb{R}\}$
  - Cannot shatter 5 points: consider subset with first, third and fifth point
- Finite sets of real numbers
  - $X = \mathbb{R}$, $\mathcal{H} = \{Z \mid Z \subseteq \mathbb{R}, |Z| < \infty\}$
  - Can shatter any finite set of reals — VC-dimension is infinite

VC-Dimension, Examples

- Intervals of reals have VC-dimension 2
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \mid a \leq b \in \mathbb{R}\}$
  - Cannot shatter 3 points: consider subset with first and third point
- Pairs of intervals of reals have VC-dimension 4
  - $X = \mathbb{R}$, $\mathcal{H} = \{[a, b] \cup [c, d] \mid a \leq b, c \leq d \in \mathbb{R}\}$
  - Cannot shatter 5 points: consider subset with first, third and fifth point
- Finite sets of real numbers
  - $X = \mathbb{R}$, $\mathcal{H} = \{Z \mid Z \subseteq \mathbb{R}, |Z| < \infty\}$
  - Can shatter any finite set of reals — VC-dimension is infinite
- Convex polygons, $X = \mathbb{R} \times \mathbb{R}$
  - For any n, place n points on unit circle
  - Each subset of these points is a convex polygon — VC-dimension is infinite

VC-dimension and machine learning

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.

VC-dimension and machine learning

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.

- We can rewrite this using VC-dimension.

Sample bound using VC-dimension

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn using D of size $O\left(\frac{1}{\epsilon}\left[\text{VC-dim}(\mathcal{H})\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right]\right)$, then with probability $\geq 1 - \delta$,

- every $h \in \mathcal{H}$ with true error $\mathrm{err}_D(h) \geq \epsilon$ has training error $\mathrm{err}_S(h) > 0$,
- i.e., every $h \in \mathcal{H}$ with training error $\mathrm{err}_S(h) = 0$. has true error $\mathrm{err}_D(h) < \epsilon$

VC-dimension and machine learning

PAC learning guarantee

Let $\mathcal{H}$ be a hypothesis class, $\delta, \epsilon > 0$ and S a training set of size $n \geq \frac{1}{\epsilon}(\ln|\mathcal{H}| + \ln(1/\delta))$ drawn using D. With probability $\geq 1 - \delta$, every $h \in \mathcal{H}$ with true error $\mathrm{err}_D > \epsilon$ has training error $\mathrm{err}_S > 0$.

- We can rewrite this using VC-dimension. Can similarly restate uniform convergence.

Sample bound using VC-dimension

For any class $\mathcal{H}$ and distribution D, if a training sample S is drawn using D of size $O\left(\frac{1}{\epsilon}\left[\text{VC-dim}(\mathcal{H}) \ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right]\right)$, then with probability $\geq 1 - \delta$,

- every $h \in \mathcal{H}$ with true error $\mathrm{err}_D(h) \geq \epsilon$ has training error $\mathrm{err}_S(h) > 0$,
- i.e., every $h \in \mathcal{H}$ with training error $\mathrm{err}_S(h) = 0$. has true error $\mathrm{err}_D(h) < \epsilon$

Summary

- PAC learning and uniform convergence use size of finite hypothesis set as measure of representational capacity
- VC-dimension provides a way of measuring capacity for infinite hypothesis sets
- VC-dimension may be finite or infinite
- For finite VC-dimension, we have analogues of PAC learning guarantee and uniform convergence
- Note that these theoretical bounds are hard to use in practice
- Difficult, if not impossible, to compute VC-dimension for complex models