

10-315 Recitation

VC-Dimension Practice

Misha

21 March 2019

Sample Complexity: Finite Hypothesis Class

- Fix hypothesis class H such that $|H| < \infty$
- Find $h \in H$ with smallest error on training set (ERM)
- In realizable case, if $m \geq \frac{1}{\varepsilon} \left(\log |H| + \log \frac{2}{\delta} \right)$ then we get error at most ε with probability at least $1 - \delta$.

Problem: Infinite Hypothesis Classes

Most classes of practical interest are infinite:

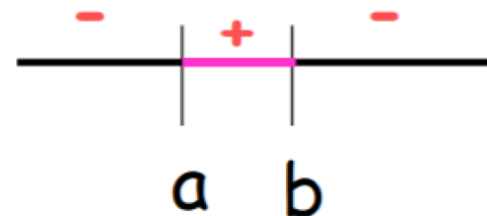
E.g., linear separators in \mathbb{R}^d



E.g., thresholds on the real line



E.g., intervals on the real line



Solution: VC-Dimension

- Want to consider an *effective* number of hypotheses – how many ways we can split the data using hypotheses from H
- A set of points S is **shattered** by hypothesis class H if there is a hypothesis in H that classifies S in all $2^{|S|}$ possible ways.
- The **VC-dimension** d_H is the cardinality $|S|$ of the largest set S that can be shattered by H

VC-Dimension Guarantee: Sauer's Lemma

- Fix hypothesis class H such that $d_H < \infty$
- Find $h \in H$ with smallest error on training set (ERM)
of size $|S| = d_H$ that is shattered by H .
- In realizable case, if $m \geq \frac{1}{2\varepsilon} \left(d_H \log \frac{1}{\varepsilon} + \log \frac{2}{\delta} \right)$ then we get error at most ε with probability at least $1 - \delta$.

How to Find VC-Dimension?

Given a hypothesis class H and sample space X :

How to Find VC-Dimension?

Given a hypothesis class H and sample space X :

- Find $S \subset X$ of size $|S| = d$ that is shattered by H .
- Show that any $S \subset X$ of size $|S| > d$ is not shattered by H .
- Then $d_H = d$ is the VC-dimension of H

Practice: Thresholds

Suppose $H = \{h(x) = 1_{x \geq a} : a \in \mathbb{R}\}$ for $X = \mathbb{R}$.

What is the VC-dimension?

Practice: Quadratic Separators

Suppose $H = \left\{ h(\mathbf{x}) = 1_{0 \leq a_{0,0} + \sum_{i,j}^d a_{i,j} x_i x_j} : a_{i,j} \in \mathbb{R} \right\}$ for $X = \mathbb{R}^d$.

Show that the VC-dimension is at most $O(d^2)$

Practice: Quadratic Separators

Suppose $H = \left\{ h(\mathbf{x}) = 1_{0 \leq a_{0,0} + \sum_{i,j}^d a_{i,j} x_i x_j} : a_{i,j} \in \mathbb{R} \right\}$ for $X = \mathbb{R}^d$.

Show that the VC-dimension is at most $O(d^2)$

Hint: recall that the VC-dimension of *linear* separators is $d + 1$

Practice: Convex Polygons

Suppose $H = \{1_{x \in \text{ConvexHull}(p_1, \dots, p_n)} : n \in \mathbb{Z}_+, p_i \in \mathbb{R}^2\}$ for $X = \mathbb{R}^2$.

What is the VC-dimension?

Does VC-Dimension Roughly Correspond to Number of Parameters?

- Common heuristic that often works:
 - Linear separators
 - Quadratic separators
 - Convex polygons
 - Intervals
 - Circles
- But also fails:
 - Consider $H = \{1_{\sin(\theta x) \geq 0} : \theta \in \mathbb{R}\}$ for $X = \mathbb{R}$. Single parameter, but for any $m > 0$ can shatter any set $S = \{2^{-k} : k \in [m]\}$, which has size m .