COL866: Foundations of Data Science

Ragesh Jaiswal, IITD

Machine Learning: Generalization

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size $n \geq \frac{2}{\varepsilon}\left[\log_2(2\mathcal{H}[2n]) + \log_2(1/\delta)\right]$. then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with error $err_D(h) \geq \varepsilon$ has $err_S(h) > 0$. Equivalently, every $h \in \mathcal{H}$ with $err_S(h) = 0$ has $err_D(h) < \varepsilon$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, if a training sample S is drawn from D of size $n \geq \frac{8}{\varepsilon^2}\left[\log_2(2\mathcal{H}[2n]) + \log_2(2/\delta)\right]$. then with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ will have $|err_D(h) - err_S(h)| \leq \varepsilon$.

Theorem (Sauer's Lemma)

If $VCdim(\mathcal{H}) = d$, then $\mathcal{H}[n] \leq \sum_{i=0}^{d} \binom{n}{i} \leq \left(\frac{en}{d}\right)^d$.

Theorem

For any hypothesis class $\mathcal{H}$ and distribution D, a training sample S of size

$$O\left(\frac{1}{\varepsilon}\left[VCdim(\mathcal{H})\log(1/\varepsilon) + \log 1/\delta\right]\right)$$

is sufficient to ensure that with probability at least $(1-\delta)$, every $h \in \mathcal{H}$ with $err_D(h) \geq \varepsilon$ has $err_S(h) > 0$.

- The VC-dimension of intervals on a real line is ____?
- For intervals on the real line, $H[n] =$ ____?
- The VC-dimension of convex polygons in d dimensional space is ____?
- For convex polygons in d dimensional space, $H[n] =$ ___?
- The VC-dimension of halfspaces in d dimensional space is ____?

- The VC-dimension of intervals on a real line is 2?
- For intervals on the real line, $H[n] = \underline{O(n^2)}$?
- The VC-dimension of convex polygons in d dimensional space is $\underline{\infty}$?
- For convex polygons in d dimensional space, $H[n] = \underline{2^n}$?
- The VC-dimension of halfspaces in d dimensional space is ____?

- The VC-dimension of intervals on a real line is 2?
- For intervals on the real line, $H[n] = O(n^2)$?
- The VC-dimension of convex polygons in d dimensional space is ∞?
- For convex polygons in d dimensional space, $H[n] = 2^n$?
- The VC-dimension of halfspaces in d dimensional space is $d+1$?

- The VC-dimension of halfspaces in d dimensional space is $d + 1$?
  - Claim 1: There exists a set of $d + 1$ points in $\mathbb{R}^d$ that is shattered by halfspaces.
  - Claim 2: No set of $d + 2$ points in $\mathbb{R}^d$ can be shattered by halfspaces.

- The VC-dimension of halfspaces in d dimensional space is $d+1$?
  - Claim 1: There exists a set of $d+1$ points in $\mathbb{R}^d$ that is shattered by halfspaces.
  - Claim 2: No set of $d+2$ points in $\mathbb{R}^d$ can be shattered by halfspaces.

Theorem (Radon)

Any set $S \subseteq \mathbb{R}^d$ with $|S| \geq d+2$, can be partitioned into disjoint subsets A and B such that $CV(A) \cap CV(B) \neq \emptyset$. Here $CV(.)$ denotes the convex hull *of the points.*

Machine Learning: Online learning and Perceptron

- The learning scenario that we have seen until now is called the batch learning scenario.
- We now discuss the online learning scenario where we remove the assumption that data is sampled from a fixed probability distribution (or from any probabilistic process at all).
- Here are some main ideas of online learning:
  - At each time $t = 1, 2, 3...$, the algorithm is presented with an arbitrary example $x_t \in \mathcal{X}$.
  - The algorithm is told the true label $c^{\star}(x_t)$ and is charged for a mistake, i.e., when $c^{\star}(x_t) \neq \ell_t$.
  - The goal of the algorithm is to make as few mistakes as possible.
- Online learning model is harder than the batch learning model. (In fact, we will show that an online algorithm can be converted to a batch learning algorithm)

- Here are some main ideas of online learning:
  - At each time $t = 1, 2, 3...$, the algorithm is presented with an arbitrary example $x_t \in \mathcal{X}$.
  - The algorithm is told the true label $c^\star(x_t)$ and is charged for a mistake, i.e., when $c^\star(x_t) \neq \ell_t$.
  - The goal of the algorithm is to make as few mistakes as possible.
- Case study:
  - Let $\mathcal{X} = \{0, 1\}^d$ and let the target hypothesis be a disjunction.
  - Question: Can you give an online algorithm that makes bounded number of mistakes?
  - Question: Argue that for any deterministic algorithm A there exists a sequence of examples σ and disjunction c^* such that A makes at least d mistakes on sequence σ labeled by c^*.

- Here are some main ideas of online learning:
  - At each time $t = 1, 2, 3...$, the algorithm is presented with an arbitrary example $x_t \in \mathcal{X}$.
  - The algorithm is told the true label $c^\star(x_t)$ and is charged for a mistake, i.e., when $c^\star(x_t) \neq \ell_t$.
  - The goal of the algorithm is to make as few mistakes as possible.
- Case study:
  - Let $\mathcal{X} = \{0, 1\}^d$ and let the target hypothesis be a disjunction.
  - Question: Can you give an online algorithm that makes bounded number of mistakes?
  - Question: Argue that for any deterministic algorithm A there exists a sequence of examples σ and disjunction c^* such that A makes at least d mistakes on sequence σ labeled by c^*.
  - Question: Show that there always exists an online algorithm that makes at most $\log_2 |\mathcal{H}|$ mistakes.

End