DDA5001 Machine Learning

Missing Proofs for VC Dimension

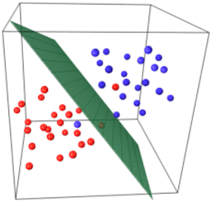
Xiao Li

School of Data Science
The Chinese University of Hong Kong, Shenzhen

VC Dimension of Linear Classifier

For a linear classifier, we can derive its VC dimension in a general sense.

This can be generalized to the following general result:



Theorem

For d-dimensional (binary) linear classifier, we have

$$\boxed{d_{\mathsf{VC}} = d + 1.}$$

Proof: VC Dimension of Binary Linear Classifier

Our general proof idea is divided into two parts: 1) We show that $d_{\mathsf{VC}} \geq d+1$. 2) We show that $d_{\mathsf{VC}} \leq d+1$. The only possibility will be $d_{\mathsf{VC}} = d+1$.

We prove the first direction.

- We consider any invertible data matrix with $d+1$ data points, i.e., $\boldsymbol{X} \in \mathbb{R}^{(d+1)\times(d+1)}$ (why also $d+1$ columns?).
- We can choose a $\mathcal{H} \ni f_{\boldsymbol{\theta}}(\boldsymbol{x}) = \operatorname{sign}(\boldsymbol{\theta}^\top \boldsymbol{x})$ by $\boldsymbol{\theta} = \boldsymbol{X}^{-1}\boldsymbol{y}$ for arbitrary $\boldsymbol{y} \in \{-1,+1\}^{d+1}$.
- Then, we will have $\operatorname{sign}(\boldsymbol{X}\boldsymbol{\theta}) = \boldsymbol{y}$. Since $\boldsymbol{y} \in \{-1,+1\}^{d+1}$ is arbitrary. We has shown that $d_{\mathsf{VC}} \geq d+1$.

Proof: VC Dimension of Binary Linear Classifier

We now prove the second direction by showing: We cannot shatter any set of $d+2$ data points.

- Consider any $d+2$ data points $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{d+2}\}$.
- We have more points than dimension. Through the basic linear algebra, there must be some j such that $\boldsymbol{x}_j = \sum_{i \neq j} \alpha_i \boldsymbol{x}_i$ and not all α_i's are zero.
- Consider the following dichotomy: All $\boldsymbol{x}_i$'s with $\alpha_i \neq 0$ are labeled as $y_i = \mathrm{sign}(\alpha_i)$, and $y_j = -1$.
- $\boldsymbol{x}_j = \sum_{i \neq j} \alpha_i \boldsymbol{x}_i$ implies that $\boldsymbol{\theta}^\top \boldsymbol{x}_j = \sum_{i \neq j} \alpha_i \boldsymbol{\theta}^\top \boldsymbol{x}_i$. For $\boldsymbol{x}_i$'s with $\alpha_i \neq 0$, by our construction, we force $y_i = \mathrm{sign}(\boldsymbol{\theta}^\top \boldsymbol{x}_i) = \mathrm{sign}(\alpha_i)$, which implies $\alpha_i \boldsymbol{\theta}^\top \boldsymbol{x}_i > 0$ whenever $\alpha_i \neq 0$.
- This implies $y_j = \mathrm{sign}(\boldsymbol{\theta}^\top \boldsymbol{x}_j) = \mathrm{sign}(\sum_{i \neq j} \alpha_i \boldsymbol{\theta}^\top \boldsymbol{x}_i) = +1$, which contradicts to our setting $y_j = -1$. Hence, our constructed dichotomy cannot be achieved by choosing any $f \in \mathcal{H}$ (more precisely, choosing $\boldsymbol{\theta}$). This means $\mathcal{G}_{\mathcal{H}}(d+2) < 2^{d+2}$.

We then have $d_{\mathsf{VC}} \leq d+1$ and complete the proof.

VC Dimension Generalization Result

After introducing all the related notions, we can now introduce the VC dimension generalization result.

VC generalization bound

For any $\delta > 0$, with probability at least $1 - \delta$, we have the following generalization bound:

$$\forall f \in \mathcal{H} \qquad \mathrm{Er}_{\mathrm{out}}(f) \leq \mathrm{Er}_{\mathrm{in}}(f) + \sqrt{\frac{8}{n} \log \left(\frac{4\mathcal{G}_{\mathcal{H}}(2n)}{\delta} \right)}$$

Upon invoking the upper bound on growth function using VC dimension, we have

$$\boxed{\forall f \in \mathcal{H} \qquad \mathrm{Er}_{\mathrm{out}}(f) \leq \mathrm{Er}_{\mathrm{in}}(f) + \sqrt{\frac{8}{n} \log \left(\frac{4((2n)^{d_{\mathrm{VC}}} + 1)}{\delta} \right)}}$$

- pp. 187 - 192 in the “Learning from data” book provides a full proof. We provide a sketch.

Proof Sketch

- Applying union bound by counting $|\mathcal{H}|$ leads to infinity. Nonetheless, $\mathcal{H}$ can only generate $\mathcal{G}_{\mathcal{H}}(n)$ (finite) dichotomies even if $\mathcal{H}$ has infinitely many f.
- Hence, $\mathrm{Er}_{\mathrm{in}}(f)$ can only take $\mathcal{G}_{\mathcal{H}}(n)$ different values. However, $\mathrm{Er}_{\mathrm{out}}(f)$ has the space $\mathcal{X}$ as input space, which can still take infinitely many values.
- The key idea in the proof is to consider a "ghost dataset" $\mathcal{S}'$ that are i.i.d. to $\mathcal{S}$. Then, one can show that

$$\Pr\left[|\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}_{\mathrm{out}}(f)| \geq t\right] \leq 2 \Pr\left[\left|\mathrm{Er}_{\mathrm{in}}(f) - \mathrm{Er}'_{\mathrm{in}}(f)\right| \geq t/2\right].$$

- Applying standard union bound and then Hoeffding's inequality to the right-hand side yields the result.

It is because of the introducing of the ghost dataset (which introduce the factors 2 highlighted in a purple color), bound changes to

$$\sqrt{\frac{1}{2n}\log\left(\frac{2|\mathcal{H}|}{\delta}\right)} \dashrightarrow \sqrt{\frac{1}{2n}\log\left(\frac{2\mathcal{G}_{\mathcal{H}}(n)}{\delta}\right)} \implies \sqrt{\frac{8}{n}\log\left(\frac{4\mathcal{G}_{\mathcal{H}}(2n)}{\delta}\right)}$$