FoSML: Lecture 6

Last time: a generalization bound of the form $R(h) \le \hat{R}_S(h) + \mathcal{R}_m(\mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{2m}}$.

- It would be nice to know $\mathcal{R}_m(\mathcal{H})$ exactly, but this may be beyond reach of simple computation. We can instead consider surrogates that are easier to get our hands on.

Defn: Let $\mathcal{H}$ be a hypothesis class. The associated **growth function** $\Pi_{\mathcal{H}}: \mathbb{Z}_{\ge 0} \to \mathbb{Z}_{\ge 0}$ is

$$\Pi_{\mathcal{H}}(m) = \max_{\{x_1,\dots,x_m\} \subseteq \mathcal{X}} \left| \{ (h(x_1),\dots,h(x_m)) \}_{h \in \mathcal{H}} \right|.$$

- Intuitively, $\Pi_{\mathcal{H}}(m)$ measures complexity by counting the maximum number of distinct ways an m-element subset of $\mathcal{X}$ can be classified.

- We now relate the growth function to Rademacher complexity via Massart's lemma.

Theorem (Massart's Lemma): Let $\mathcal{A} \subseteq \mathbb{R}^m$ be finite and let $r = \max_{x \in \mathcal{A}} \|x\|_2$.

Then:

$$\mathbb{E}_z\left(\frac{1}{m} \sup_{x \in \mathcal{A}} \sum_{i=1}^{m} z_i x_i \right) \le \frac{r\sqrt{2\log|\mathcal{A}|}}{m}.$$

Corollary (Growth Function Bounds $\mathcal{R}_m(\cdot)$): Let G be a family of functions taking values in $\{-1, 1\}$. Then $\mathcal{R}_m(G) \le \sqrt{\frac{2\log(\Pi_G(m))}{m}}$.

Proof: Let S be a fixed sample $S = \{x_1,\dots,x_m\}$. Denote by $G|_S$ the set

of vectors of function values $(g(x_1),\dots,g(x_m))$, where $g \in G$. Since g maps into $\{-1,1\}$, $\|(g(x_1),\dots,g(x_m))\|_2 \le \sqrt{m}$. We can then apply Massart's lemma as follows:

$$\mathcal{R}_m(G) = \mathop{\mathbb{E}}_{S}\left(\mathop{\mathbb{E}}_{\sigma}\left(\sup_{u \in G|_S} \frac{1}{m}\sum_{i=1}^{m}\sigma_i u_i\right)\right)$$

$$\le \mathop{\mathbb{E}}_{S}\left(\frac{\sqrt{m}\sqrt{2\log|G|_S|}}{m}\right)$$

$$\le \mathop{\mathbb{E}}_{S}\left(\frac{\sqrt{m}\sqrt{2\log(\Pi_G(m))}}{m}\right)$$

$$= \frac{\sqrt{m}\sqrt{2\log(\Pi_G(m))}}{m}$$

$$= \sqrt{\frac{2\log(\Pi_G(m))}{m}} \quad \blacksquare$$

This immediately yields a generalization bound in terms of the growth function, simply by replacing $\mathcal{R}_m(G)$ with this estimate in the Rademacher complexity bound:

Corollary (Growth function generalization bound): Let $\mathcal{H}$ be a family of functions taking values in $\{-1,1\}$. Then $\forall \delta > 0$, with probability exceeding $1-\delta$,

$$\forall h \in \mathcal{H},\quad R(h) \le \hat{R}_S(h) + \sqrt{\frac{2\log(\Pi_{\mathcal{H}}(m))}{m}} + \sqrt{\frac{\log(1/\delta)}{2m}}.$$

- It remains to count the number of ways to classify m points... this could be hard.
- This suggests something else: Vapnik-Chervonenkis (VC) notion of complexity

In order to introduce VC dimension, we need the notion of shattering.

Defn: Let $\mathcal{H}$ be a hypothesis class with associated growth function $\Pi_{\mathcal{H}}$. A set of points with cardinality m is shattered by $\mathcal{H}$ if $\Pi_{\mathcal{H}}(m) = 2^m$.

In other words, a set S is shattered if all possible labels of it (of which there are 2^m) can be realized by $\mathcal{H}$.

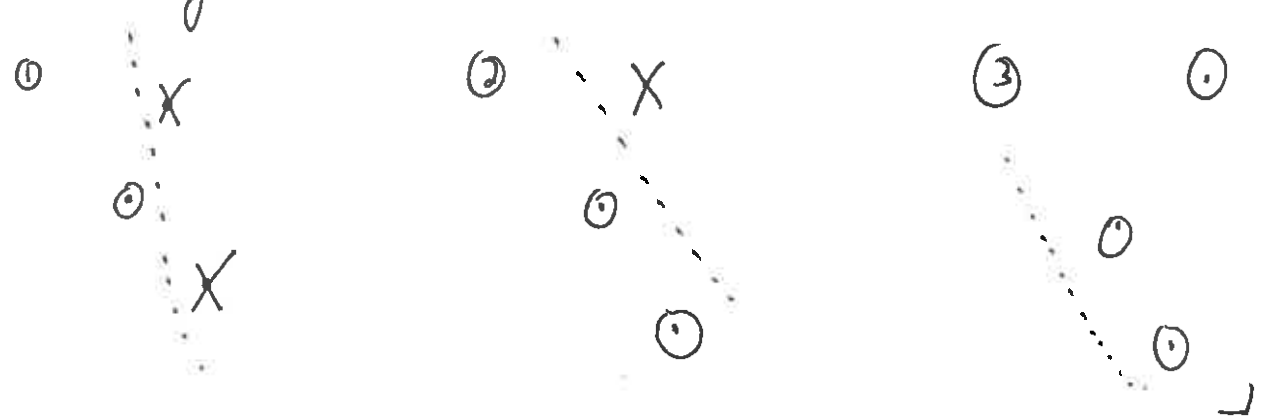
Defn (VC dimension): The VC-dimension of a hypothesis class $\mathcal{H}$ is the size of the largest set that can be shattered by $\mathcal{H}$:

$$VCdim(\mathcal{H}) = \max_m \{ m \mid \Pi_{\mathcal{H}}(m) = 2^m \}.$$

Remark: If $VCdim(\mathcal{H}) = d$, there is at least one set of cardinality d in $\mathcal{X}$ which is shattered by $\mathcal{H}$. It does not mean all sets with d elements are shattered by $\mathcal{H}$.

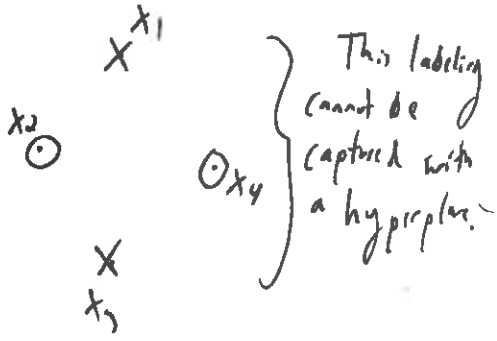
ex: Let $\mathcal{H}$ consist of lines in $\mathbb{R}^2$, where points on one side are given one label, points on the other another label. This is what linear SVM do.

Claim 1: $VCdim(\mathcal{H}) \geq 3$. Let x_1, x_2, x_3 be any set of non-collinear points. Then the following configurations of labels are all we need to consider:

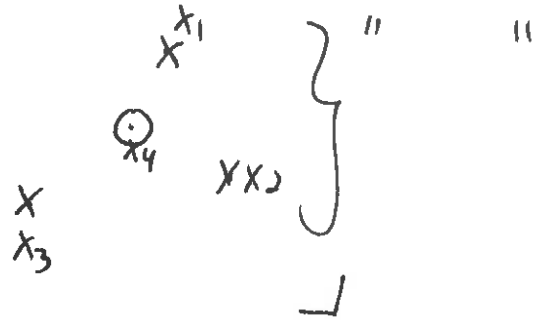


Claim 2: $VCdim(\mathcal{H}) = 3$. To show this, we must first show no set of 4 points is shattered by $\mathcal{H}$. It suffices to consider two cases, defined in terms of the convex hull of $\{x_1, x_2, x_3, x_4\}$:

① All points are on the ~~exterior~~ boundary of the convex hull:



② Three points are on the boundary of the convex hull, and one is in the interior:



Ex: Let $\mathcal{H}$ consist of the set of hyperplanes in $\mathbb{R}^d$. Let $x_0 = (0, \dots, 0)$

$x_i = e_i$, $i = 1, \dots, d$.

Let $y_0, y_1, \dots, y_d \in \{-1, 1\}$ be an associated labeling for $x_0, x_1, \dots x_d$. We will show $VCdim(\mathcal{H}) \geq d+1$ by constructing a hyperplane that achieves the labeling $\{(x_i, y_i)\}_{i=0}^{d}$, for any choice of labeling $\{y_i\}_{i=0}^{d}$. Let $w \in \mathbb{R}^d$ be defined as $w_i = y_i$. Consider the hyperplane $\{x \mid w \cdot x + \frac{y_0}{2} = 0\}$. Notice the label this hyperplane gives to x_i is

$$sgn\left(w \cdot x_i + \frac{y_0}{2}\right) =$$

$$= sgn\left(y_i + \frac{y_0}{2}\right)$$

← since $x_i = e_i$, and $w_i = y_i$ (a base of notation...)

$= \text{sgn}(y_i)$

$= y_i$. ⌋

Harder is to show $VC\dim(\mathcal{H}) = d+1$, by showing $VC\dim(\mathcal{H}) < d+2$. To show this, we use Radon's Theorem:

Theorem (Radon): Let $X \subset \mathbb{R}^d$ be s.t. $|X| = d+2$. Then $\exists$ a partition $X_1 \cup X_2 = X$ such that the convex hulls of X_1, X_2 intersect.

Returning to our argument that $VC\dim(\mathcal{H}) \leq d+2$, let X be a set of $d+2$ points. By Radon's Theorem, partition $X_1 \cup X_2 = X$ s.t. the convex hulls of X_1, X_2 intersect. Then X_1, X_2 cannot be separated by a hyperplane! So, give labels 1 to X_1, -1 to X_2; no hyperplane can achieve this labeling. Thus, $\mathcal{H}$ doesn't shatter X.

/

. There is a natural connection between growth functions and VC-dimension:

Thm (Sauer's Lemma): Let $\mathcal{H}$ be a hypothesis set with $VC\dim(\mathcal{H}) = d$. Then

$$\forall m \in \mathbb{Z}_{\geq 0}, \quad \Pi_{\mathcal{H}}(m) \leq \sum_{i=0}^{d} \binom{m}{i}.$$

- After a bit of combinatorics, this gives an upper bound of the form $\Pi_{\mathcal{H}}(m) \leq m^{VC\dim(\mathcal{H})}$.

- Proofs next time!