Machine Learning - B
Lectures 5
Week 5: VC Analysis of SVM

Amartya Sanyal

University of Copenhagen

Last Time: VC Generalization Bound

$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \widehat{L}(h, S) + \varepsilon\right) \leq 2 \cdot \mathbb{P}\left(\exists h \in \mathcal{H} : \widehat{L}(h, S') - \widehat{L}(h, S) \geq \frac{\varepsilon}{2}\right)$$
$$\leq 2 \cdot m_{\mathcal{H}}(2n) \cdot e^{-\frac{n\varepsilon^2}{8}} \; = \; \delta$$

Breakdown:

Selection: Union bound over $m_{\mathcal{H}}(2n)$ "bags of errors"

Concentration: Hoeffding bound for deviation in one bag

Theorem (Restated):

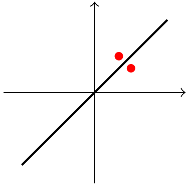
$$\mathbb{P}\left(\exists h \in \mathcal{H} : L(h) \geq \widehat{L}(h, S) + \sqrt{\frac{8 \ln\left(\frac{2 m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}\right) \leq \delta$$

Bounding $m_{\mathcal{H}}(n)$ — Shattering

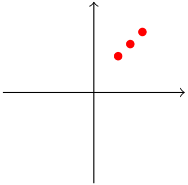
- **Definition:** $x_1, \ldots, x_n$ are *shattered* by $\mathcal{H}$ if

$$|\Pi_{\mathcal{H}}(x_1, \ldots, x_n)| = 2^n$$

- That is, $\mathcal{H}$ can realize **all** 2^n possible binary labelings of $x_1, \ldots, x_n$.
- **Example:** Homogeneous linear classifiers in $\mathbb{R}^2$



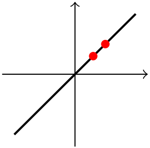
Shattered



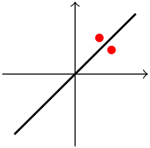
Not shattered

Definition: VC-dimension

$$d_{\mathrm{VC}}(\mathcal{H}) = \max\{n : m_{\mathcal{H}}(n) = 2^n\}$$



Not shattered



Shattered

To show $d_{\mathrm{VC}}(\mathcal{H}) = d$, we need:

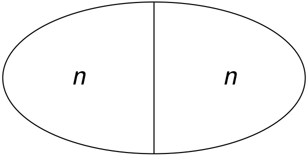
$$d_{\mathrm{VC}}(\mathcal{H}) \geq d \quad \text{(construct a shattered set of size } d\text{)}$$
$$d_{\mathrm{VC}}(\mathcal{H}) \leq d \quad \text{(prove no set of size } d+1 \text{ can be shattered)}$$

Always require *both* directions to prove equality. **Example: Linear separators in $\mathbb{R}^d$**

$$d_{\mathrm{VC}} = \begin{cases} d & \text{Homogeneous Linear Classifier} \\ d+1 & \text{Non-homogeneous Linear Classifier} \end{cases}$$

VC lower bound

- If $d_{\mathrm{VC}}(\mathcal{H}) = \infty$, then for any n there exists a distribution $p(x, y)$, such that
$$\mathbb{P}\left(\exists\, h \in \mathcal{H} : L(h) - \widehat{L}(h, S) \geq \tfrac{1}{8}\right) \;\geq\; \tfrac{1}{8}.$$
- Proof idea: construct the same lower bound as for a finite $\mathcal{H}$ in ML-A based on $2n$ points shattered by $\mathcal{H}$.



Message: if the VC-dimension of $\mathcal{H}$ is infinite, we cannot learn with $\mathcal{H}$.

VC Generalization Bound

- Last time: $\mathbb{P}\left(\exists\, h \in \mathcal{H} : L(h) - \widehat{L}(h,S) \;\geq\; \sqrt{\frac{8\,\ln\left(\frac{2\,m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}\right) \;\leq\; \delta.$

- Theorem (Today): $m_{\mathcal{H}}(n) \underbrace{\leq}_{Step 1} \sum_{i=0}^{d_{\mathrm{VC}}(\mathcal{H})} \binom{n}{i} \underbrace{\leq}_{Step 2} n^{d_{\mathrm{VC}}(\mathcal{H})} + 1.$

- Corollary:

$$\mathbb{P}\left(\exists\, h \in \mathcal{H} : L(h) - \widehat{L}(h,S) \;\geq\; \sqrt{\frac{8\,\ln\left(\frac{(2n)^{d_{\mathrm{VC}}(\mathcal{H})}+1}{\delta}\right)}{n}}\right) \;\leq\; \delta.$$

- If $d_{\mathrm{VC}}(\mathcal{H}) \ll n$, then with high probability $\widehat{L}(h,S)$ is close to $L(h)$ for all $h \in \mathcal{H}$ and we can learn.

- In particular, if $d_{\mathrm{VC}} \ll \frac{n}{\log n}$, then we can learn linear classifiers in $\mathbb{R}^d$.

Bounding the growth function

Sauer–Shelah lemma

Lemma

If $d_{\mathrm{VC}}(\mathcal{H}) = d < \infty$, then for all n,

$$m_{\mathcal{H}}(n) \;\le\; \sum_{i=0}^{d} \binom{n}{i} \;\le\; n^d + 1.$$

Otherwise, if $d_{\mathrm{VC}}(\mathcal{H}) = \infty$, then $m_{\mathcal{H}}(n) = 2^n$.

- This shows a *phase transition*: infinite VC $\implies$ exponential growth; finite VC $\implies$ polynomial growth.
- Plug this into the VC-generalization bound

$$\mathbb{P}\big(\exists h : L(h) - \widehat{L}(h) \ge \varepsilon\big) \;\le\; 2\, m_{\mathcal{H}}(2n)\, e^{-n\varepsilon^2/8}$$

Achieves convergence whenever $d_{\mathrm{VC}} < \infty$.

Proof of Step 1: Recap Pascal's identity

$$\binom{n}{k} = \binom{n-1}{k-1} + \binom{n-1}{k}.$$

$$\underbrace{\circ \circ \cdots \circ}_{n \text{ choose } k} = \underbrace{\circ\ \circ\ \ldots\ \circ}_{n-1 \text{ choose } k-1} + \underbrace{\circ\ \ldots\ \circ}_{n-1 \text{ choose } k}$$

- Red circle: the distinguished "last" element.
- Left term: *include* that element, then choose $k-1$ of the remaining $n-1$.
- Right term: *exclude* it, choose all k from the remaining $n-1$.

Proof of Step 1: Defining $B(n, d)$ and its recursion

- $B(n, d)$: number of labelings of n points that avoid shattering any $(d + 1)$-subset.
- Splitting on whether the forbidden shattered set includes point a specific point or not gives

$$B(n, d) \leq B(n - 1, d - 1) + B(n - 1, d).$$

$B(n, d)$

$\bullet\ \underbrace{\circ\ \dots\ \circ}_{n-1}$

include

$\Rightarrow B(n - 1, d - 1)$

$\underbrace{\circ\ \dots\ \circ}_{n-1}\ \bullet$

exclude

$\Rightarrow B(n - 1, d)$

Proof of Step 1: Completing by induction

Claim

For all $n \geq 0$ and $d \geq -1$,

$$B(n,d) \;\leq\; \sum_{i=0}^{d} \binom{n}{i},$$

- **Base cases:**
  - $n = 0$: $B(0,d) = 1$, and $\sum_{i=0}^{d} \binom{0}{i} = \binom{0}{0} = 1$.
  - $d = -1$: $B(n,-1) = 0$, and $\sum_{i=0}^{-1} \binom{n}{i} = 0$.
- We know $B(n,d) \leq B(n-1,d-1) + B(n-1,d)$
- **Inductive step:** assume true for $(n-1,d-1)$ and $(n-1,d)$. Then

$$B(n,d) \leq \sum_{i=0}^{d-1} \binom{n-1}{i} + \sum_{i=0}^{d} \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n}{i} \leq n^d + 1$$

Second last equality follows from Pascal's identity. Last inequality is exercise.

Mid-Summary

- Last time: $\mathbb{P}\left(\exists\, h \in \mathcal{H} : L(h) - \widehat{L}(h,S) \;\geq\; \sqrt{\frac{8\,\ln\left(\frac{2\,m_{\mathcal{H}}(2n)}{\delta}\right)}{n}}\right) \;\leq\; \delta.$
- Today: $m_{\mathcal{H}}(n) \;\leq\; \sum_{i=0}^{d_{\mathrm{VC}}(\mathcal{H})} \binom{n}{i} \;\leq\; n^{d_{\mathrm{VC}}(\mathcal{H})} + 1.$
- Together:

$$\mathbb{P}\left(\exists\, h \in \mathcal{H} : L(h) - \widehat{L}(h,S) \;\geq\; \sqrt{\frac{8\,\ln\left(\frac{(2n)^{d_{\mathrm{VC}}(\mathcal{H})}+1}{\delta}\right)}{n}}\right) \;\leq\; \delta.$$

- For linear classifiers in $\mathbb{R}^d$:
  - Homogeneous: $d_{\mathrm{VC}}(\mathcal{H}) = d$.
  - Non-homogeneous: $d_{\mathrm{VC}}(\mathcal{H}) = d + 1$.
  - If $d \ll \frac{n}{\ln n}$, then $\widehat{L}(h,S)$ reliably estimates $L(h)$ for all $h \in \mathcal{H}$.
- If $d_{\mathrm{VC}}(\mathcal{H}) = \infty$, we can build a lower bound showing $\widehat{L}(h,S)$ does not converge to $L(h)$ for every h.

Margin-based analysis of SVMs

SVM (hard margin):

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad \forall i: \; y_i(w^\top x_i + b) \geq 1.$$

$$\mathrm{dist}(h_{w,b}, x) = \frac{y(w^\top x + b)}{\|w\|} \;\geq\; \frac{1}{\|w\|},$$

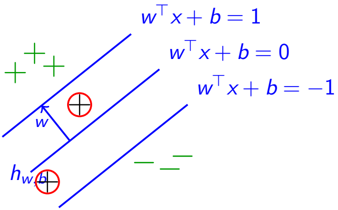
$$\gamma = \frac{1}{\|w\|} \text{ (margin).}$$

Fat loss:

$$\ell_{\mathrm{FAT}}(h_{w,b}(x), y) = \begin{cases} 0, & y(w^\top x + b) \geq 1, \\ 1, & \text{otherwise.} \end{cases}$$

$$\widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\mathrm{FAT}}(h_{w,b}(X_i), Y_i),$$

$$L_{\mathrm{FAT}}(h_{w,b}) = \mathbb{E}\left[\ell_{\mathrm{FAT}}(h_{w,b}(X), Y)\right].$$



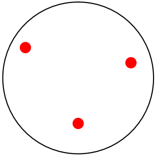
Fat-shattering

- $\mathcal{H}_\gamma = \{h_{w,b} : \|w\| \leq 1/\gamma\}$
- **Definition:** A set $\{x_1, \ldots, x_n\}$ is *fat-shattered* by $\mathcal{H}_\gamma$ if for every labeling $(y_1, \ldots, y_n) \in \{0,1\}^n$ there exists $h_{w,b} \in \mathcal{H}_\gamma$ such that

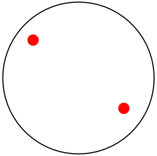
$$y_i \left(w^\top x_i + b\right) \geq 1 \quad \forall i.$$

- **Fat-shattering dimension** $d_{\mathrm{FAT}}(\mathcal{H}_\gamma)$ is the largest n for which some set of n points in the unit ball ($\|x\| \leq 1$) can be fat-shattered.
- Minimizing $\frac{1}{2}\|w\|^2$ pushes for large γ, hence small d_{FAT}.

small γ: $d_{\mathrm{FAT}} = 3$



large γ: $d_{\mathrm{FAT}} = 2$



Fat generalization bound

Theorem (Similar to VC & Occam's razor):

$$\mathbb{P}\left(\exists\, h_{w,b} \in \mathcal{H}_\gamma : L_{\mathrm{FAT}}(h_{w,b}) - \widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) \;\geq\; \sqrt{\frac{8}{n} \ln\left(\frac{2\left((2n)^{d_{\mathrm{FAT}}(\mathcal{H}_\gamma)} + 1\right)}{\delta}\right)}\right)$$

$$\leq\; \delta.$$

Theorem (Will not prove in class): If $\mathcal{X} = \{x : \|x\| \leq 1\}$, then

$$d_{\mathrm{FAT}}(\mathcal{H}_\gamma) \;\leq\; \left\lfloor \frac{1}{\gamma^2} \right\rfloor + 1.$$

Note that d_{FAT} depends on the "margin" γ of the resulting hypothesis
Change Indexes to reflect sequence of classes of decreasing margin

$$j = \left\lfloor \frac{1}{\gamma^2} \right\rfloor, \quad \mathcal{H}_j = \{h_{w,b} : \|w\| \leq \sqrt{j}\} = \mathcal{H}_{\gamma = 1/\sqrt{j}}, \quad d_{\mathrm{FAT}}(\mathcal{H}_j) = j + 1.$$

Margin based generalisation bound

Theorem (FAT Generalisation Bound)

With probability at least $1-\delta$*, every* $h_{w,b} \in \mathcal{H}$ *satisfies*

$$L(h_{w,b}) \;\leq\; \widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) \;+\; \sqrt{\frac{8\,\ln\left(\frac{2\left((2n)^{d_{\mathrm{FAT}}(\mathcal{H}_\gamma)}+1\right)}{\delta}\right)}{n}}.$$

Substitute $d_{\mathrm{FAT}}(\mathcal{H}_\gamma) \;\leq\; \left\lfloor \frac{1}{\gamma^2} \right\rfloor + 1 \leq \|w\|^2 + 1$ and choosing δ appropriately

Corollary (Choose δ_k)

$$\mathbb{P}\left(\exists\, h_{w,b} \in \mathcal{H} :\; L_{\mathrm{FAT}}(h_{w,b}) - \widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) \;\geq\; \sqrt{\frac{8\,\ln\left(2\left((2n)^{\|w\|^2+1}+1\right)\,\|w\|^2(\|w\|^2+1)/\delta\right)}{n}}\right) \;\leq\; \delta/\|w\|^2(\|w\|^2+1).$$

Margin based generalisation bound

Based on Occam's razor + VC, peel off a union bound over the "norm-layers" $\mathcal{H}_j$:

$$\mathbb{P}(\exists\, h_{w,b} \in \mathcal{H} : \ldots) \;\le\; \sum_{j=1}^{\infty} \mathbb{P}\left(\exists\, h_{w,b} \in \mathcal{H}_j : L_{\mathrm{FAT}}(h_{w,b}) - \widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) \ge \sqrt{\frac{8\,\ln\left(\frac{2\left((2n)^{j+1}+1\right)\, j(j+1)}{\delta}\right)}{n}} \right)$$

$$\le \sum_{j=1}^{\infty} \frac{\delta}{j(j+1)} \;=\; \delta.$$

Theorem

With probability at least $1-\delta$, for every linear separator $h_{w,b} \in \mathcal{H}$ we have

$$L(h_{w,b}) \;\le\; L_{\mathrm{FAT}}(h_{w,b}) \;\le\; \widehat{L}_{\mathrm{FAT}}(h_{w,b}, S) \;+\; \sqrt{\frac{8}{n}\,\ln\left(\frac{2\,(2n)^{\|w\|^2+1}+1}{\delta} \;\times\; \|w\|^2\,(\|w\|^2+1)\right)}.$$

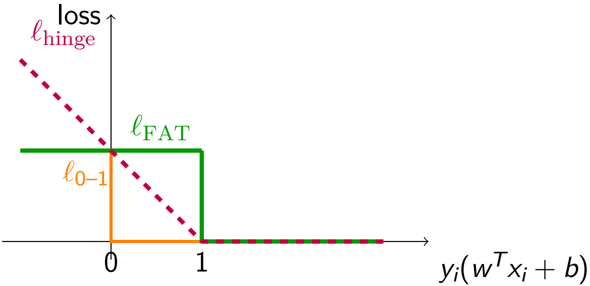
Hard-margin vs. soft-margin SVM

Hard-margin SVM

$$\min_{w,b} \frac{1}{2}\|w\|^2 \quad \text{s.t.} \quad y_i\big(w^\top x_i + b\big) \ \geq \ 1 \quad \forall i = 1, \ldots, n.$$

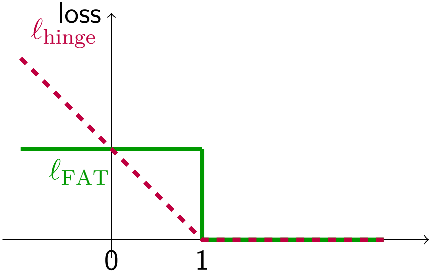
Soft-margin SVM

$$\min_{w,b,\,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \quad \text{s.t.} \quad \begin{cases} y_i\big(w^\top x_i + b\big) \ \geq \ 1 - \xi_i, \\ \xi_i \ \geq \ 0, \end{cases} \qquad i = 1, \ldots, n.$$



Discussion

- **Why optimize soft-margin SVM and not the bound?**
  - Finding the exact minimizer of the FAT-bound is NP-hard.
- We find the global minimum of the hinge-loss objective, but the gap between L_{FAT} and L_{hinge} can be arbitrarily large.
- We obtain a solid generalization guarantee, yet we don't necessarily achieve the optimal classifier.
- **Tuning the trade-off C:**
  - Use the bound to guide $\|w\|$ complexity.
  - Cross-validation to select C—only one parameter to tune!



Margins and Kernels

- We can use kernels to map data into higher-dimensional (even infinite-dimensional) feature spaces, e.g. the RBF kernel.
- For infinite-dimensional mappings, the "plain" VC bound

$$d_{\mathrm{VC}}(\widehat{\mathcal{H}}) \;\leq\; d+1$$

  is vacuous (since $d=\infty$).
- The margin-based (fat-shattering) bound still applies in infinite dimensions, because it scales with $\|w\|$ in feature space, provided
  - $\|w\|$ remains controlled, and
  - the input set $\mathcal{X}$ lies inside the unit ball of that feature space.
- Note: for true infinite-dimensional spaces one has

$$d_{\mathrm{FAT}}(\mathcal{H}) \;=\; \infty,$$

  so why isn't this in conflict with the lower bound?