Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Lecture 11: VC Dimension & SRM

- Capacity (Vapnik 3.13)
- VC Dimension (Vapnik 4.9.1-4.9.2, 4.11)
- Structural Risk Minimization (SRM)

Formal Statement (finite case)

• With probability (1-eta), simultaneously for all functions in the set {k=1,…N}, the inequality below holds true

$$R(\alpha_k) < R_{emp}(\alpha_k) + \frac{\varepsilon^2}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_k)}{\varepsilon^2}}\right), \ \varepsilon^2 = 2\frac{\ln N - \ln\eta}{\ell}$$

$$R(\alpha_k) < R_{emp}(\alpha_k) + \frac{\ln N - \ln\eta}{\ell}\left(1 + \sqrt{1 + 2\frac{R_{emp}(\alpha_k)\ell}{\ln N - \ln\eta}}\right)$$

• Since it holds for all functions in the set, it holds in particular for the function that minimizes ERM. In other words we get a bound on “the value of achieved risk (for the rule selected by ERM)”

• The second bound (difference) follows easily from the first, we do not discuss it here

$$(2) \quad \Delta(\alpha_\ell) = R(\alpha_\ell) - R(\alpha_0)$$

Formal Statement (infinite case)

• With probability (1-eta), simultaneously for all functions in the set, the inequality below holds true

$$R(\alpha_k) < R_{emp}(\alpha_k) + \frac{\mathrm{E}(\ell)}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_k)}{\mathrm{E}(\ell)}}\right)$$

• Same two comments from the previous slide apply

• Note $\mathrm{E}(\ell)$ is a quantity expressed in terms of some capacity concept (not necessarily entropy)

Recap

- We showed that capacity concepts completely define the quantitative theory (bounds) as well
- However the bounds we obtained are *non-constructive*!
- For a given set of functions, how do you compute entropy? (You can't!)
- Moreover, bounds in terms of entropy are *distribution-dependent*
- To evaluate entropy must plug in a specific pdf (it can be any pdf)
- This motivates a structure of capacity concepts.
- Goal: *distribution-independent* and *constructive* bounds

Structure of Capacity Concepts

- Number of clusters induced by the sample & function set:

$$N^\wedge(z_1,\ldots,z_\ell) \leq 2^\ell$$

- *Random Entropy* (of the set of indicator functions on the given sample):

$$H^\wedge(z_1,\ldots,z_\ell) = \ln N^\wedge(z_1,\ldots,z_\ell)$$

- *Entropy* (of the set of indicator functions on samples of size ℓ):

$$H^\wedge(\ell) = E\left[\ln N^\wedge(z_1,\ldots,z_\ell)\right] = \int \ln N^\wedge(z_1,\ldots,z_\ell)\, dF(z_1,\ldots,z_\ell)$$

- *Annealed Entropy* (…):

$$H_{ann}^\wedge(\ell) = \ln E\left[N^\wedge(z_1,\ldots,z_\ell)\right]$$

- *Growth function* (…):

$$G^\wedge(\ell) = \ln\left[\sup_{z_1,\ldots,z_\ell} N^\wedge(z_1,\ldots,z_\ell)\right]$$

Structure of Capacity Concepts

• What's the point? Growth function is distribution independent and upper-bounds entropy (due to Jensen's inequality). Anywhere we have entropy, we can always substitute growth and get a dist-ind bound!

$$H^{\wedge}(\ell) \leq H^{\wedge}_{ann}(\ell) \leq G^{\wedge}(\ell)$$

$$E\left[\ln N^{\wedge}(z_1,\ldots,z_\ell)\right] \leq \ln E\left[N^{\wedge}(z_1,\ldots,z_\ell)\right] \leq \ln\left[\sup_{z_1,\ldots,z_\ell} N^{\wedge}(z_1,\ldots,z_\ell)\right]$$

• *Jensen's inequality*: assuming we have a convex function f, and a random variable X,

$$f\left(E[X]\right) \leq E\left[f(X)\right]$$

• But logarithm is a concave function, hence the inequality is reversed when we consider number of clusters (our random variable)

VC Dimension (idea)

• Growth function is distribution independent but is not constructive (hard to evaluate for a given set of functions)

• Introduce a new capacity concept (function) which bounds the growth function but is easier to evaluate

$$H^{\wedge}(\ell) \leq H^{\wedge}_{ann}(\ell) \leq G^{\wedge}(\ell) \leq J(h,\ell)$$

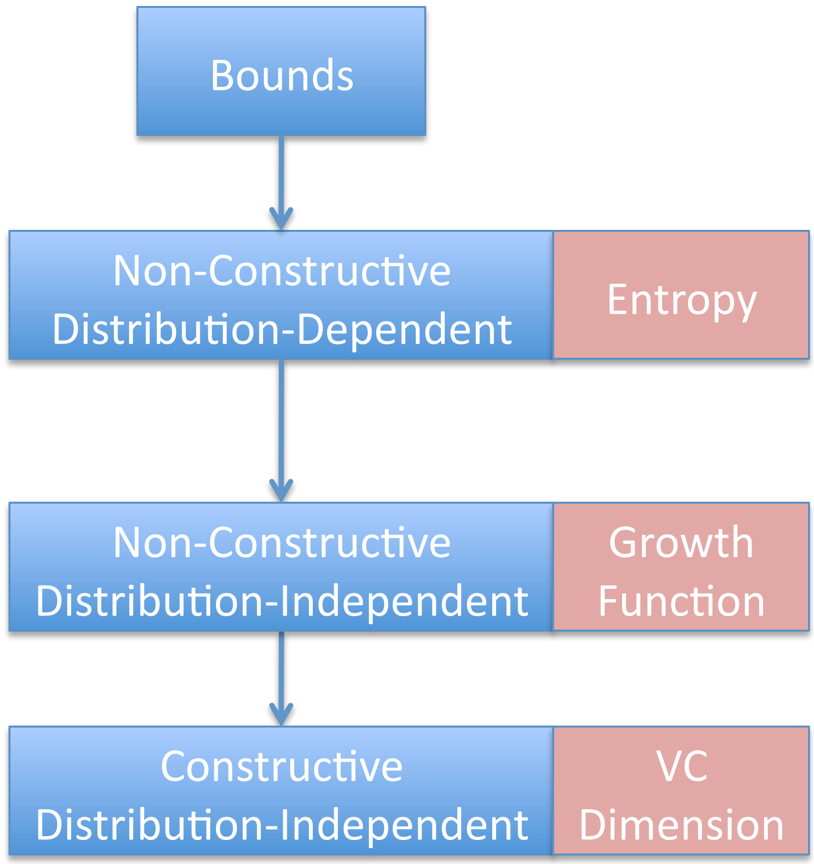
• "J" is some function of {coefficient, # examples}

• The coefficient h is called the Vapnik-Chervonenkis (VC) dimension of a set of indicator functions

• If the VC dimension for an admissible set of functions is finite, we know that ERM is consistent on this set (for indicator loss functions)

• Actually, we can show necessity as well (not discussed, see Vapnik 4.9.3)

Road Map (Capacity)



Binomial Coefficient

- In order to bound the growth function, we need to bound the following sum of binomial coefficients:

$$\sum_{i=0}^{h} \binom{m}{i}, \; h \leq m \quad easy: \sum_{i=0}^{m} \binom{m}{i} = 2^m$$

- We also need the following identity: $\exp \equiv e = \lim_{n \to \infty} \left(1 + \frac{1}{n}\right)^n$
- Derivation:

$$(1) \quad \sum_{i=0}^{h} \binom{m}{i} \leq \sum_{i=0}^{h} \binom{m}{i} \cdot \left(\frac{h}{m}\right)^i \left(\frac{m}{h}\right)^h \quad \leq \quad \left(\frac{m}{h}\right)^h \sum_{i=0}^{m} \binom{m}{i} \cdot \left(\frac{h}{m}\right)^i$$

$$(2) \quad \left(\frac{m}{h}\right)^h \sum_{i=0}^{m} \binom{m}{i} \cdot \left(\frac{h}{m}\right)^i 1^{m-i} = \left(\frac{m}{h}\right)^h \left(1 + \frac{h}{m}\right)^m$$

$$(1) \left(\frac{h}{m}\right) \leq 1$$

$$(2) \; Binomial \;\; formula$$

Binomial Coefficient

- Given: $\sum_{i=0}^{h}\binom{m}{i},\ h \le m \quad \exp \equiv e = \lim_{n\to\infty}\left(1+\frac{1}{n}\right)^{n}$

- Derivation:

$$(1)\quad \sum_{i=0}^{h}\binom{m}{i} \le \sum_{i=0}^{h}\binom{m}{i}\cdot\left(\frac{h}{m}\right)^{i}\left(\frac{m}{h}\right)^{h} \le \left(\frac{m}{h}\right)^{h}\sum_{i=0}^{m}\binom{m}{i}\cdot\left(\frac{h}{m}\right)^{i}$$

$$(2)\quad \left(\frac{m}{h}\right)^{h}\sum_{i=0}^{m}\binom{m}{i}\cdot\left(\frac{h}{m}\right)^{i}1^{m-i} = \left(\frac{m}{h}\right)^{h}\left(1+\frac{h}{m}\right)^{m}$$

$$(3)\quad \left(\frac{m}{h}\right)^{h}\left(1+\frac{h}{m}\right)^{m} \le \left(\frac{m}{h}\right)^{h} e^{h}$$

$$\Rightarrow \sum_{i=0}^{h}\binom{m}{i} \le \left(\frac{em}{h}\right)^{h}$$

$(1)\ \left(\frac{h}{m}\right) \le 1$

(2) *Binomial Formula*

(3) *Identity*

Growth Function

$$G^{\wedge}(\ell) = \ln\left[\sup_{z_1,\ldots,z_\ell} N^{\wedge}(z_1,\ldots,z_\ell)\right]$$

- The growth function for a set of indicator functions satisfies one of two conditions:

$$(a) \quad G^{\wedge}(\ell) = \ell \ln 2$$

$$(b) \quad G^{\wedge}(\ell) \equiv \begin{cases} \ell \ln 2 & \text{if } \ell \le h \\ \le \ln\left(\sum_{i=0}^{h} \binom{\ell}{i}\right) & \text{if } \ell > h \end{cases}$$

where h is the largest integer for which $G^{\wedge}(h) = h \ln 2$

- Using the bound from the previous slide:

$$\ln\left(\sum_{i=0}^{h} \binom{\ell}{i}\right) \le \ln\left(\frac{e\ell}{h}\right)^{h} = h\left(1 + \ln\frac{\ell}{h}\right)$$

Growth Function $G^{\wedge}(\ell) = \ln\left[\sup_{z_1,\ldots,z_\ell} N^{\wedge}(z_1,\ldots,z_\ell)\right]$

- The growth function for a set of indicator functions satisfies one of two conditions:

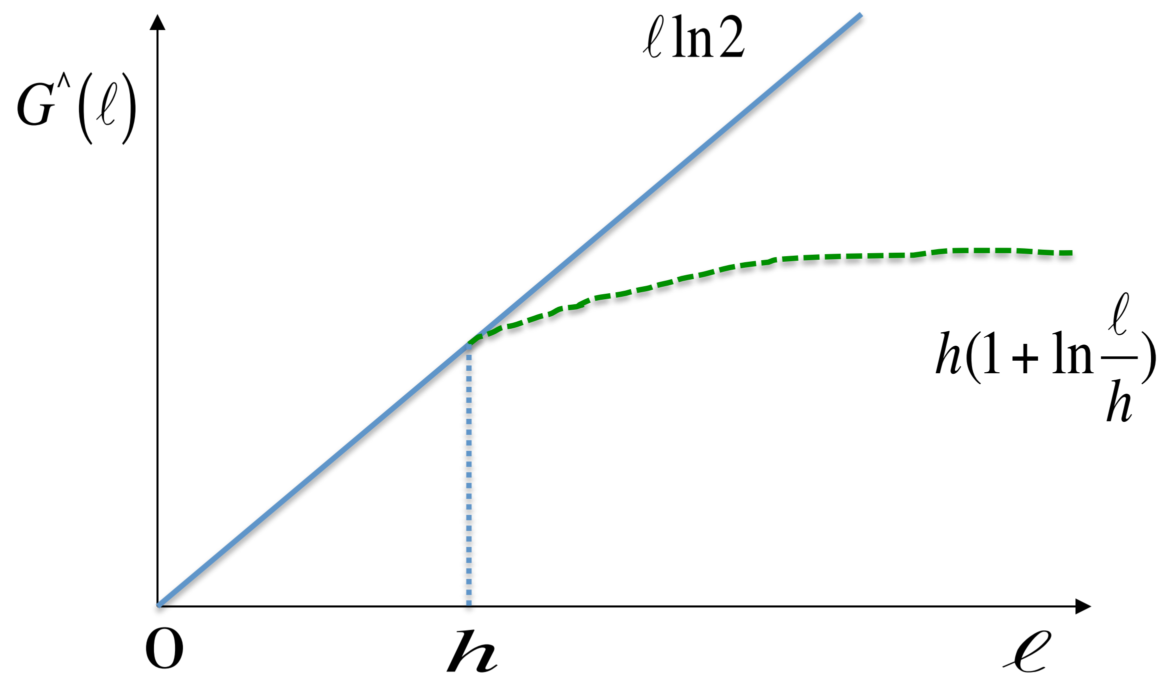
$$(a) \quad G^{\wedge}(\ell) = \ell \ln 2$$

$$(b) \quad G^{\wedge}(\ell) \equiv \begin{Bmatrix} \ell \ln 2 & if \;\; \ell \le h \\ \le h\left(1 + \ln\frac{\ell}{h}\right) & if \;\; \ell > h \end{Bmatrix}$$

where h is the largest integer for which $G^{\wedge}(h) = h \ln 2$

Growth Function Behavior

- The growth function is either linear or bounded by a logarithmic function with coefficient h. It cannot be of any intermediate form!
- This is crucial to prove sufficiency & necessity for the VC dimension capacity concept (with respect to ERM consistency)



General Idea: Subsets

• Can talk about subsets of a set instead of clusters (also known as Sauer's Lemma). Here we assume that Z is an (infinite) set of elements, and the sample is a particular subset

$$(a) \quad \sup_{z_1,\ldots,z_\ell} N^S(z_1,\ldots,z_\ell) = 2^\ell$$

$$(b) \quad \sup_{z_1,\ldots,z_\ell} N^S(z_1,\ldots,z_\ell) \equiv \begin{cases} 2^\ell & if \;\; \ell \le h \\ \le \left(\sum_{k=0}^{h} \binom{\ell}{k}\right) \le \left(\frac{e\ell}{h}\right)^h & if \;\; \ell > h \end{cases}$$

where h is the largest integer for which equality is valid.

Note: the above is not a precise argument, just an outline

Note: Sauer's Lemma is just the growth function theorem (result) stated for the general case of subsets of a set

VC Dimension (Definition)

- **Definition:** The coefficient h which characterizes the capacity of a set of functions with logarithmic-bounded growth function is called the VC dimension (of a set of indicator functions). When the growth function is linear, the VC dimension is defined to be infinite.

- We can modify the definition to stress the constructive method of estimating the VC dimension

VC-dim (Constructive Definition)

- **Definition:** The VC dimension of a set of indicator functions is equal to the largest number (h) of vectors $\left(x_1,\ldots,x_\ell\right)$ that can be separated into two different classes in all the 2^h possible ways using this set of functions.

- The VC dimension is the maximum number of vectors that can be *shattered by the set of functions*

- If for any n, there exists a set of n vectors that can be shattered by the given set of functions, then the VC dimension is equal to infinity

Shattering

- Shattering:
  - ➢ We pick h points & place them at $\left(x_1,\ldots,x_h\right)$
  - ➢ They challenge us with every possible (2^h in total) assignment (labeling)

    $\left(y_1,\ldots,y_h\right) \in (\pm 1,\ldots,\pm 1)$
  - ➢ If our set of admissible functions (i.e. concept class, classifiers) can satisfy every possible assignment (correctly classify for every labeling), then the VC dimension is at least h
  - ➢ Recall: growth function is "supremum over every set". Therefore, it is enough to demonstrate just ***one*** placement of points to show VC dim is at least h
  - ➢ To show VC dim is less than h+1, we need to show that for ***every*** possible placement of h+1 points (every set) there exists some labeling that can't be achieved

Constructive Bound

- With probability (1-eta), for the function that minimizes empirical risk, the inequality below holds true

$$R(\alpha_\ell) < R_{emp}(\alpha_\ell) + \frac{\mathrm{E}(\ell)}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_\ell)}{\mathrm{E}(\ell)}}\right)$$

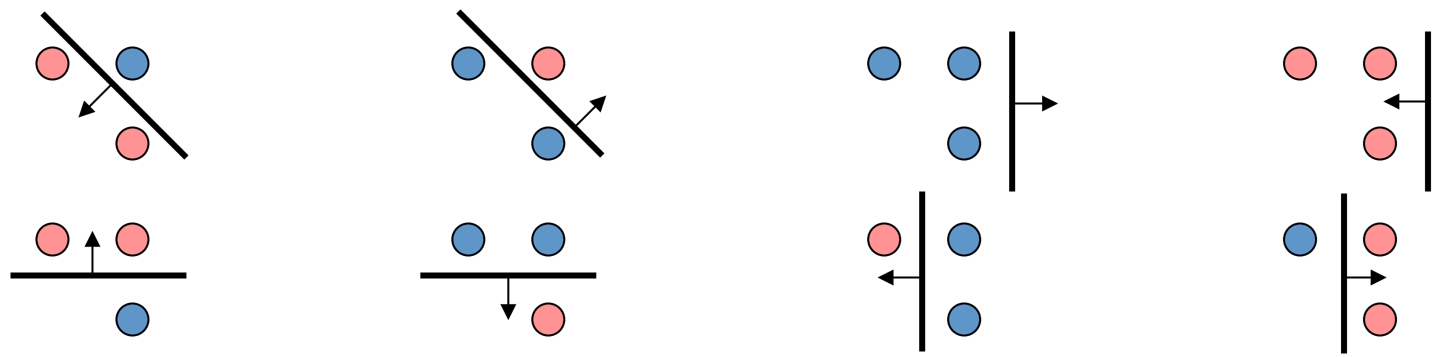
where

$$\mathrm{E}(\ell) = 4\,\frac{h\big(1 + \ln(2\ell/h)\big) - \ln(\eta/4)}{\ell}$$

Example: 2D Linear Classifiers

- Linear classifiers ➔ h = 3
- Can't ever shatter 4 points!
- Can't shatter 3 points on a straight line (but that doesn't matter)
- Note: # of parameters = VC dimension

$$f(x;w) = w_0 + w_1 x_1 + w_2 x_2$$



Example: N-D Linear Classifiers

- Consider a more general case: linear classifier in N dimensions
- A hyperplane in R^N shatters any set of *affinely independent* points
- Affine combination is a weighted average of the points (where sum of weights = 1)
- Can choose N+1 affinely independent points ➔ h = N+1

- Not quite satisfactory in practice!
- What if I have lots of redundant features (dimensions)? h should be less than N+1
- But VC estimate does not distinguish between such cases and cases where features are valuable!
- Solution: gap tolerant classifiers, bound on VC dimension in terms of margin

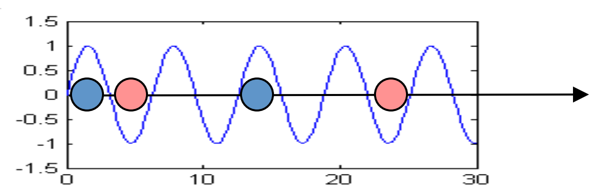
Example: 1D Sinusoidal Classifiers

- Consider the set of functions $f(x;\theta) = sign\left(\sin\left(\theta x\right)\right)$
- Number of parameters = 1, but h = infinity
- Can choose points wisely and shatter perfectly for every n
- Note: h not proportional to # of parameters

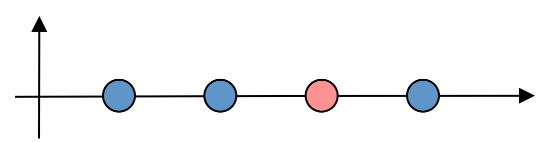
$$choose: \ x_i = 10^{-i}, \ i = 1, \ldots, h$$

$$given: \ y_1, \ldots, y_h$$

$$set: \ \theta = \pi\left(1 + \sum_{i=1}^{h} \frac{1}{2}\left(1 - y_i\right)10^{-i}\right)$$



But, as a side note, if I choose 4 equally spaced x's then cannot shatter



Example: Nearest Neighbor Classifier

- K-Nearest Neighbor (K-NN) Algorithm: classify each data point by a majority vote of its K neighbors
- K=1 ➔ classify by nearest neighbor (1-NN)
- 1-NN shatters any set of points ➔ h = infinity
- Empirical risk is always zero, but classifier can still perform well in practice!
- Infinite capacity does not guarantee poor performance (Note, there is no contradiction here: infinite VC implies that U.C doesn't take place, and hence ERM is not consistent, but that doesn't mean that the algorithm doesn't do well in a particular situation)