Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Lecture 10: Statistical Learning Theory (Bounds)

- General model of learning & ERM (Vapnik 0.1-1.11)
- Consistency (Vapnik 3.1-3.2.1)
- Uniform Convergence (Vapnik 3.3, 3.4, 3.7)
- Entropy, Capacity (Vapnik 3.7, 3.10, 3.13)
- Capacity (Vapnik 3.13)
- Bounds (Vapnik 4.1, 4.8)
- VC Dimension (Vapnik 4.9.1, 4.11)
- Structural Risk Minimization (SRM)

Recap

- We introduced a capacity concept for a set of indicator functions
  - One-function case: just a particular case of LLN
  - Finite case: just number of functions in the set
  - General (infinite) case: entropy of functions on a sample
- Using this concept we obtained conditions for 2-sided U.C. However, we need conditions for 1-sided U.C
- Obviously if 2-sided holds, we have 1-sided, but what about cases where only 1-sided holds? Perhaps we can relax the conditions we obtained for 2-sided U.C ?
- Not a trivial problem!

Models of Reasoning

- Two models of reasoning: *deductive* and *inductive*
  - Deductive: from general to particular (true consequences from true premises)
  - Inductive: general judgments from particular assertions
- But general judgments from true particular assertions are not always true!
- *Demarcation* problem (I.Kant): when is the inductive step justified? (What is the difference between cases where it is and is not?)
- The problem can be discussed in the context of scientific theories: is there a way to distinguish between scientific and non-scientific theories?

Non-Falsifiability

- Is there a formal way to distinguish between scientific and non-scientific theories?
- Necessary condition to justify a theory (K. Popper): *feasibility* of its *falsification*
    - ➢ Existence of particular assertions which fall into the theory's domain but cannot be explained by it
    - ➢ If a theory can be falsified, it satisfies the conditions of a scientific theory
    - ➢ If there is no example that can falsify the theory, it should be considered a non-scientific theory

Mathematical Non-Falsifiability

- Suppose the following equality holds (for indicator functions):

$$\forall \ell : \frac{H^{\wedge}(\ell)}{\ell} = \ln 2 \Rightarrow N^{\wedge}(z_1,\ldots,z_\ell) = 2^{\ell}$$

- In other words, almost any sample (of arbitrary size) can be separated in all possible ways by the set of functions of the machine
- Therefore the minimum of empirical risk is zero
- This is a *nonfalsifiable learning machine*, it can give a general explanation for almost any data
- "Almost any data" since the entropy is defined in terms of the integral:

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

From 2-sided to 1-sided (idea)

- Suppose we have a non-falsifiable machine "A" (2-sided U.C does NOT take place)
- It is possible that the machine can generalize using ERM (one-sided U.C)
- If we can find a second, falsifiable, machine "B" that is arbitrarily close to "A", we can deduce U.C(1) for "A"

Formally:

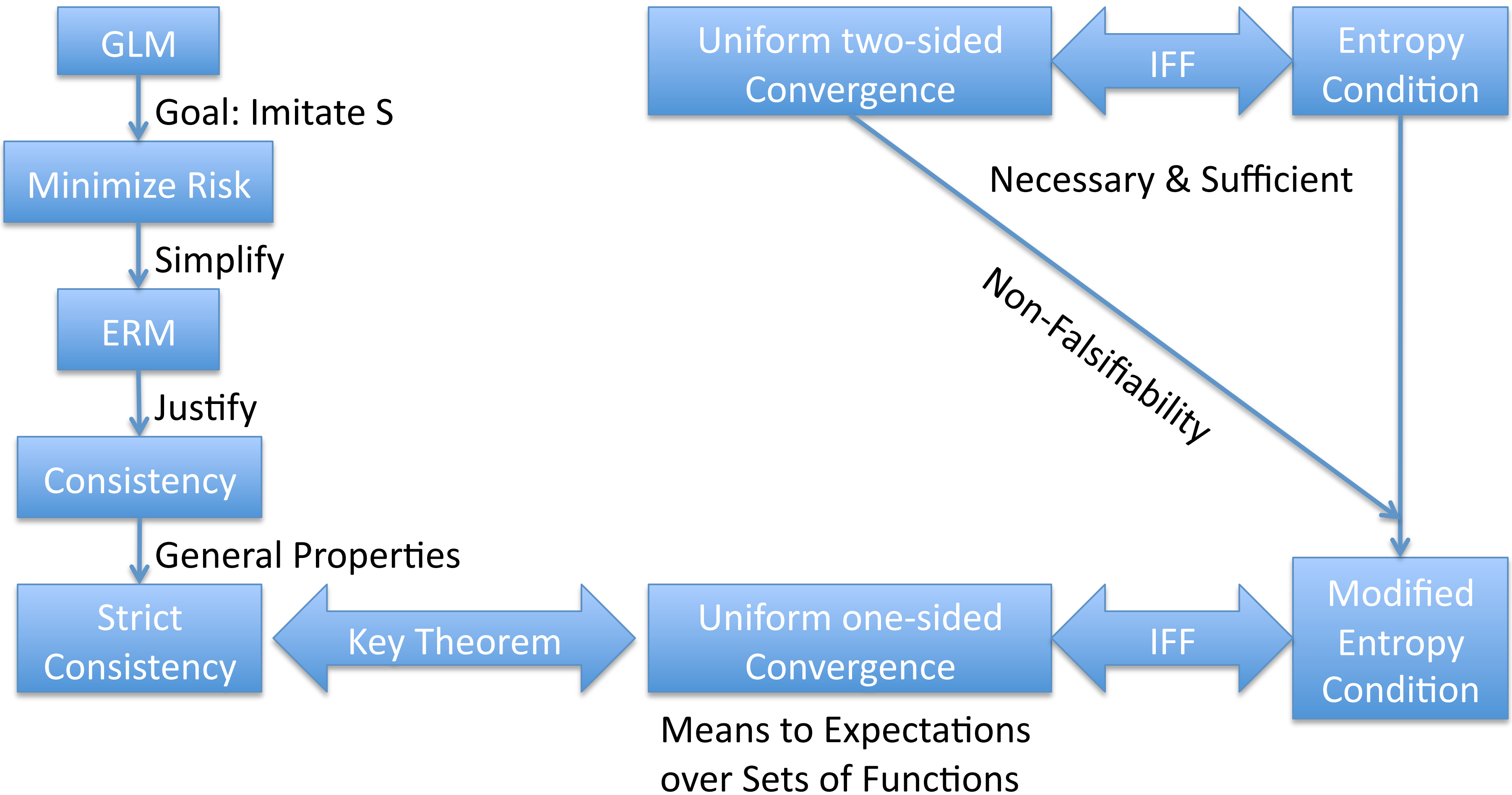
- Suppose we have a set of functions {L} for which 2-sided U.C does NOT take place
- Now introduce a new set of functions {L*} with the following property:

$$\forall \varepsilon,\ \forall L(\mathbf{z},\alpha),\ \exists L^*\left(\mathbf{z},\alpha^k\right):$$

$$\int \left(L(\mathbf{z},\alpha) - L^*\left(\mathbf{z},\alpha^k\right)\right) dF(\mathbf{z}) < \varepsilon$$

- If for the second set {L*}, U.C(2) is valid, then for the first set {L}, U.C(1) holds

Road Map (4)


GLM
Goal: Imitate S
Minimize Risk
Simplify
ERM
Justify
Consistency
General Properties
Strict Consistency
Key Theorem
Uniform two-sided Convergence
IFF
Entropy Condition
Necessary & Sufficient
Non-Falsifiability
Uniform one-sided Convergence
IFF
Modified Entropy Condition
Means to Expectations over Sets of Functions

Recap

• We introduced a capacity concept (entropy) which completely defines the qualitative behavior of the learning processes (we are specifically referring to ERM)

• Do capacity concepts completely define the quantitative theory (bounds) as well?

• Quantitative theory → Rate of Convergence → Bounds

• Note: there are some shortcomings to entropy, therefore we are motivated to introduce a whole structure of concepts (which motivates VC dimension)

• What are the conditions for the existence of a *fast* asymptotic rate of U.C for a given probability measure?

> ➢ Conditions for existence of two positive constants {b,c} such that for a sufficiently large sample:

$$P\left\{ \sup_{\alpha} \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell}\sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right| > \varepsilon \right\} < b\exp\{-c\varepsilon^2\ell\}$$

Types of Bounds

- Bounds determine the generalization ability of the learning machine (utilizing ERM)
- We focus on indicator loss functions
- We would like to estimate two quantities:
  - (1) The value of achieved risk (for the rule selected by ERM)
  - (2) The difference between achieved and minimal risk for a given function set

$$Suppose: \quad \inf_{\alpha} R(\alpha) @ \alpha_0, \quad \inf_{\alpha} R_{emp}(\alpha) @ \alpha_\ell$$

$$(1) \quad R(\alpha_\ell)$$

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

Comments

• Estimating difference (2) is easy to do once the value (1) is estimated. Hence we focus on the first quantity R(alpha_L}

• Recall that we already have some bounds (Chernoff bounds) on the probability of two-sided convergence

• Therefore we would like to use these results (which we have for the maximum over all alphas in the set) to derive a bound on a particular risk value (particular since it is for the function that minimizes empirical risk)

• Our approach will once again be to start from the finite case and then derive the infinite case using the obtained forms

$$P\left\{ \sup_{\alpha} \left| R(\alpha) - R_{emp}(\alpha) \right| > \varepsilon \right\}$$

$$P\left\{ \max_{1 \le k \le n} \left| p_{L>0} - v_{\ell} \right| > \varepsilon \right\} \le 2N \exp\left\{ -2\varepsilon^2 \ell \right\}$$

Recall: Chernoff Bounds

- Recall: we considered Chernoff bounds for U.C

$$P\left\{ \sup_{\alpha} \left| R(\alpha) - R_{emp}(\alpha) \right| > \varepsilon \right\}$$

$$\equiv P\left\{ \sup_{\alpha} \left| \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) - \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{z}_i,\alpha) \right| > \varepsilon \right\}$$

$$\equiv P\left\{ \sup_{\alpha} \left| P\{L(\mathbf{z},\alpha) > 0\} - v_{\ell}\{L(\mathbf{z},\alpha) > 0\} \right| > \varepsilon \right\}$$

$$\equiv P\left\{ \sup_{\alpha} \left| p_{L>0} - v_{\ell} \right| > \varepsilon \right\}$$

- For finite set of functions case:

$$P\left\{ \max_{1 \le k \le n} \left| p_{L>0} - v_{\ell} \right| > \varepsilon \right\} \le 2N \exp\left\{-2\varepsilon^2 \ell\right\} = 2\exp\left\{\left(\frac{\ln N}{\ell} - 2\varepsilon^2\right)\ell\right\}$$

Relative Uniform Convergence

- Now we are interested in *relative* convergence:

$$P\left\{ \sup_{\alpha} \frac{|p_{L>0} - \nu_{\ell}|}{\sqrt{p_{L>0}}} > \varepsilon \right\} < ?$$

- Why?

- Suppose our set of functions (set of alphas) contains only "bad" functions that provide probability of error close to ½: then in this pessimistic case, the bounds (using additive Chernoff inequalities) we can obtain on U.C are *tight*. In other words we can't improve the bound.

- But what if the set contains at least one good function which provides probability of error equal (close) to zero: then in this optimistic case, the bounds for U.C actually "penalize" us for considering the entire set of functions equally.

- By considering convergence relative to the expectation we take all cases (including intermediate between the above) into account (and hence we get better bounds).

Multiplicative Chernoff Bounds

- Notation: $S = X_1 + \ldots + X_m,\ X_i \in \{0,1\},\ 0 \le \varepsilon \le 1$

$$\Pr[X_i = 1] = p, \quad \mu = E[S] = pm, \quad \hat{p} = \frac{S}{m}$$

- Additive Form:

$$\Pr[\hat{p} - p > \varepsilon] \le \exp\left\{-2\varepsilon^2 m\right\} \qquad \Pr[p - \hat{p} > \varepsilon] \le \exp\left\{-2\varepsilon^2 m\right\}$$

- Multiplicative Form (in terms of standard deviation):

$$\Pr[\hat{p} - p > \varepsilon p] \le \exp\left\{-\frac{\varepsilon^2 pm}{3}\right\} \qquad \Pr[p - \hat{p} > \varepsilon p] \le \exp\left\{-\frac{\varepsilon^2 pm}{2}\right\}$$

$$\Pr[\hat{p} - p > \varepsilon p] \equiv \Pr[\frac{p - \hat{p}}{\sqrt{p}} > \varepsilon\sqrt{p}]$$

$$\varepsilon^* = \varepsilon\sqrt{p} \Rightarrow \quad \Pr[\frac{p - \hat{p}}{\sqrt{p}} > \varepsilon^*] \le \exp\left\{-\frac{(\varepsilon^*)^2 m}{2}\right\}$$

Bounds: Finite Case

- Suppose our set contains N functions (where N is finite)

$$\alpha_{1,\ldots,N} \in \Lambda, |\Lambda| = N \Rightarrow \sup_{\alpha} \equiv \max_{\alpha}$$

$$\Pr\left[\frac{p - \hat{p}}{\sqrt{p}} > \varepsilon\right] \le \exp\left\{-\frac{\varepsilon^2 \ell}{2}\right\}$$

- Using Multiplicative Chernoff bounds:

$$P\left\{\max_{1\le k\le n} \frac{p_{L>0} - v_\ell}{\sqrt{p_{L>0}}} > \varepsilon\right\} \le \sum_{k=1}^{N} P\left\{\frac{p_{L>0}(k) - v_\ell(k)}{\sqrt{p_{L>0}(k)}} > \varepsilon\right\} \le N \exp\left\{-\frac{\varepsilon^2 \ell}{2}\right\}$$

$$\Rightarrow P\left\{\max_{1\le k\le n} \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \varepsilon\right\} \le N \exp\left\{-\frac{\varepsilon^2 \ell}{2}\right\}$$

- Why did we rewrite the quantity? We want to bound the value of achieved risk (for the rule selected by ERM)

Bounds: Finite Case

• We want to bound R(alpha). It would be simpler to make a statement of the form: with probability very close to 1, simultaneously for all functions in the set, the quantity R(alpha) is bounded by something

$$\textit{Let}\ \ 0 < \eta \leq 1, \quad N \exp\left\{-\varepsilon^2 \ell / 2\right\} = \eta$$

$$\Rightarrow \ln \exp\left\{-\varepsilon^2 \ell/2\right\} = \ln \frac{\eta}{N} \quad \Rightarrow \frac{\varepsilon^2 \ell}{2} = -(\ln \eta - \ln N) \quad \Rightarrow \varepsilon = \sqrt{2 \frac{\ln N - \ln \eta}{\ell}}$$

$$P\left\{ \max_{1 \leq k \leq n} \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} > \varepsilon \right\} \leq \eta \ \equiv \ \forall k : P\left\{ \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} \leq \varepsilon \right\} \geq 1 - \eta$$

$$\textit{From}\ \ \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} \leq \varepsilon \quad \textit{to}\ \ R(\alpha_k) < ?\{R_{emp}(\alpha_k), \varepsilon\}$$

$$\text{From } \frac{R(\alpha_k) - R_{emp}(\alpha_k)}{\sqrt{R(\alpha_k)}} \leq \varepsilon \quad \text{to} \quad R(\alpha_k) < ?\{R_{emp}(\alpha_k), \varepsilon\}$$

$$\frac{X - C}{\sqrt{X}} \leq \varepsilon \Rightarrow X - C \leq \varepsilon\sqrt{X} \Rightarrow (X - C)^2 \leq \varepsilon^2 X$$

$$\Rightarrow X^2 - 2CX - \varepsilon^2 X + C^2 \leq 0 \Rightarrow X^2 - (2C + \varepsilon^2)X + C^2 \leq 0$$

$$\Rightarrow X \leq \frac{2C + \varepsilon^2 \pm \sqrt{(2C + \varepsilon^2)^2 - 4C^2}}{2} = \frac{2C + \varepsilon^2 \pm \sqrt{4C^2 + 4C\varepsilon^2 + \varepsilon^4 - 4C^2}}{2}$$

$$\Rightarrow X \leq C + \frac{\varepsilon^2 \pm \varepsilon^2 \sqrt{4C/\varepsilon^2 + 1}}{2} = C + \frac{\varepsilon^2}{2}\left(1 \pm \sqrt{1 + \frac{4C}{\varepsilon^2}}\right)$$

$$\text{In our case:} \quad X = R(\alpha_k), C = R_{emp}(\alpha_k), \varepsilon = \sqrt{2\frac{\ln N - \ln \eta}{\ell}}$$

Bound Form

- We want to bound R(alpha). It would be simpler to make a statement of the form: with probability very close to 1, simultaneously for all functions in the set, the quantity R(alpha) is bounded by something

$$X \leq C + \frac{\varepsilon^2}{2}\left(1 \pm \sqrt{1 + \frac{4C}{\varepsilon^2}}\right) \quad \textit{In our case}: X = R(\alpha_k), C = R_{emp}(\alpha_k), \varepsilon = \sqrt{2\frac{\ln N - \ln \eta}{\ell}}$$

$$\Rightarrow R(\alpha_k) < R_{emp}(\alpha_k) + \frac{\varepsilon^2}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha_k)}{\varepsilon^2}}\right)$$

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

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 quite 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