Machine Learning

4771

Instructors:
Adrian Weller and Ilia Vovsha

Lecture 8: Statistical Learning Theory (Intro)

- 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)
- Bounds
- VC Dimension
- Structural Risk Minimization (SRM)

Parametric Paradigm (Philosophy)

- Heyday: 1930 – 1960's
- Standard assumptions: familiar problem & underlying physical process
- Problem: set of parameters that needs to be estimated
- Approach: adopt the Maximum-Likelihood / MAP / Bayesian method
- Strength:
    1. If assumptions are correct, we obtain more accurate estimates
    2. Math is simpler & faster to compute.
- Principle: if it works for the *asymptotic* case, should work for a small sample too.

Parametric Paradigm (Beliefs?)

A. It is possible to find a good approximation to any function with few parameters
   - Evidence (?): Weierstrass Approximation Theorem
   - Strength: computationally simple

B. The underlying law behind many real-life problems is the normal law
   - Evidence: Central Limit Theorem

C. MLE / MAP / Bayesian are good approaches for estimating the parameters
   - Evidence: conditional optimality (restricted set or asymptotic case)

Parametric Paradigm (Deficiencies?)

A. Singularities of high-dimensional problems (curse of dimensionality)
   - Increasing required accuracy → exponentially more resources
   - Resources: parameters, degree of polynomial, hidden units
   - A small set of functions is not sufficient

B. What if normal law is not applicable?
   - Wrong assumption → inaccurate estimates

C. MLE / MAP / Bayesian might not be optimal
   - General set of functions
   - Small sample case

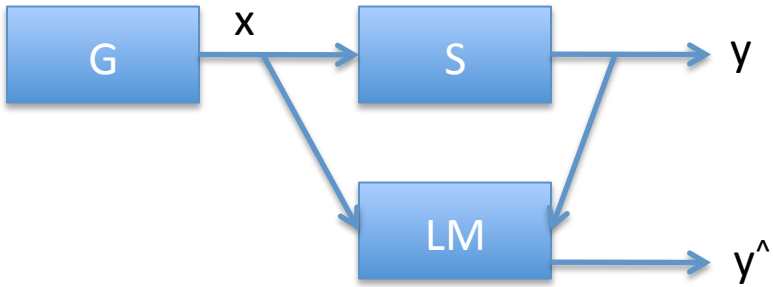
General Model of Learning

Model of learning from examples:

A. Data generator (G):
   - Generates iid vectors according to unknown, fixed pdf F(**x**)

B. Supervisor (target) operator (S):
   - Outputs labels for each vector
   - Unknown & fixed

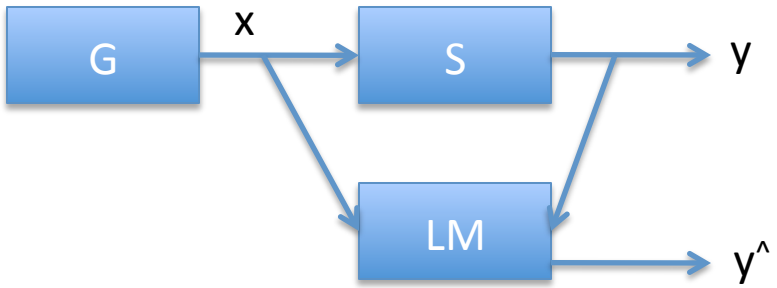
C. Learning Machine (LM):
   - Receives a training set and constructs an operator



$$X = \left\{ \left(\mathbf{x}_1, y_1 \right), \ldots, \left(\mathbf{x}_\ell, y_\ell \right) \right\}$$

General Learning Machine

- Goal of LM: construct best approximation to S
- Specific Goals:
  - Imitate S: construct best predictor of supervisor's output
  - Identify S: construct similar operator
- Practical Goal:
  - Imitation is easier, possible to develop non-asymptotic (small sample) theory
  - Choose best approximating function from a set



Minimizing Risk from Data

- Goal: among a set of functions, find the one that best satisfies a given quality criterion
- Problem: how do we choose the "best" function?
- Formal Problem Statement:

A. Specify
   - Domain [Z], PDF over Z [F(z)] // where Z is a subset of R^n, F(z) joint over (x,y)
   - Admissible set of functions: $\{g(\mathbf{z},\alpha)\}$
   - Quality criterion through loss function: $L(\mathbf{z}, g(\mathbf{z},\alpha))$

B. Minimize Risk Functional (risk)

$$R(g(\mathbf{z},\alpha)) = \int L(\mathbf{z}, g(\mathbf{z},\alpha))\, dF(\mathbf{z})$$

   - Expected loss for chosen function "g"
   - "α" denotes a set of parameters

Empirical Risk Minimization

- Problem: how do we minimize the risk functional?
- Solution: too difficult to do this directly, hence consider empirical risk instead

$$R(\alpha) = \int L(\mathbf{z}, \alpha)\, dF(\mathbf{z})$$

$$R_{emp}(\alpha) = \frac{1}{\ell} \sum_{i=1}^{\ell} L(\mathbf{z}_i, \alpha)$$

- General induction principle: to achieve good generalization (test error on unseen examples), the ERM principle constructs a decision rule that minimizes training error (empirical risk)
- Task: develop a theory for this principle
- Approach: develop theory for indicator {0,1} functions (classification), then generalize to real-valued functions (regression)

General Induction Principle

- Consider applying ERM given a very "expressive" (with high capacity) set of functions (e.g. the set of polynomials of any degree)
- Might lead to over-fitting, poor generalization
- This observation suggest that we can find conditions on the set of functions which can guarantee whether ERM is "good" (consistent) or not.
- Note: we sometimes distinguish between sets of loss functions and the set of admissible functions (e.g. polynomials), though they are implicitly lumped together
- For example, we can consider indicator {0,1} loss functions with the set of admissible functions {g(x)} being polynomials

ERM Examples

1. Classification: Perceptron

$$R_{emp}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} step(-y_i w^T x_i) \qquad \alpha \equiv W$$

$$R_{emp}^{Per}(\alpha) = -\sum_{i \in misclassified} (y_i w^T x_i)$$

2. Regression: Least Squares

$$R_{emp}(\alpha) = \frac{1}{N}\sum_{i=1}^{N} (y_i - w^T x_i)^2 \Rightarrow \frac{1}{N}\sum_{i=1}^{N} (y_i - f(x_i,\alpha))^2$$

$$R(\alpha) = \int (y - f(x,\alpha))^2 \, dF(x,y)$$

3. Density Estimation: Maximum Log-Likelihood

ERM Examples

1. Classification: Perceptron

$$R_{emp}^{Per}(\alpha) = -\sum_{i \in misclassified} (y_i w^T x_i) \qquad \alpha \equiv W$$

2. Regression: Least Squares

$$R_{emp}(\alpha) = \frac{1}{N} \sum_{i=1}^{N} (y_i - f(x_i, \alpha))^2$$

3. Density Estimation: Maximum Likelihood

$$\max p(D \mid \alpha) = \max \prod_{i=1}^{N} p(\mathbf{x}_i \mid \alpha)$$

$$\Rightarrow \min R_{emp}(\alpha) = -\sum_{i=1}^{N} \log p(\mathbf{x}_i \mid \alpha)$$

$$R(\alpha) = -\int \log p(\mathbf{x}, \alpha)\, dF(\mathbf{x})$$

Method Consistency

- What is consistency?
  - ➢ Convergence to the best solution with increasing number of examples
- Is ERM consistent?
  - ➢ No guarantee!
- Goal: describe situations under which the method is consistent
- Approach:
  1. Find the necessary and sufficient conditions for consistency
  2. Estimate the quality of the solution (rate of convergence)
- Theory:
  1. Theory of consistency (Qualitative)
  2. Theory of bounds (Quantitative, characterizes generalization)

Convergence Modes

"Find the necessary and sufficient conditions for consistency"

- Find conditions for convergence to best rule as $\ell \to \infty$
- Recall: ERM principle defines a decision rule

• Consider a sequence of random variables $r_1 \ldots r_L$:

• We say that the sequence converges to a random variable r_0

- "In Probability": $r_\ell \xrightarrow[\ell \to \infty]{P} r_0$
- "Almost surely": $r_\ell \xrightarrow[\ell \to \infty]{A.S} r_0$

Convergence Modes

- Consider a sequence of random variables converging to a random variable r_0:

A. Convergence in Probability

$$\forall \delta > 0 : \; P\left\{ \left| r_\ell - r_0 \right| > \delta \right\} \xrightarrow[\ell \to \infty]{} 0 \qquad r_\ell \xrightarrow[\ell \to \infty]{P} r_0$$

B. Almost Sure Convergence

$$\forall \delta > 0 : \; P\left\{ \sup_{\ell > n} \left| r_\ell - r_0 \right| > \delta \right\} \xrightarrow[n \to \infty]{} 0 \qquad r_\ell \xrightarrow[\ell \to \infty]{A.S} r_0$$

- Which convergence mode is stronger?

Convergence Modes

• Consider a sequence of random variables, measuring distance (using the uniform metric) between random functions and some fixed function:

$$r_\ell = \rho\big(F(x), F_\ell(x)\big) = \sup_x \big|F(x) - F_\ell(x)\big|$$

• We say that the sequence converges in probability to a random variable $r_0 = 0$

$$\forall \delta > 0 : \; P\Big\{ \big|r_\ell - r_0\big| > \delta \Big\} \xrightarrow[\ell \to \infty]{} 0$$

$$P\Big\{ \big|r_\ell - r_0\big| > \delta \Big\} \equiv P\Big\{ \big|r_\ell - 0\big| > \delta \Big\} \equiv P\left\{ \left| \sup_x \big|F(x) - F_\ell(x)\big| \right| > \delta \right\}$$

$$\Rightarrow \forall \delta > 0 : \; P\left\{ \sup_x \big|F(x) - F_\ell(x)\big| > \delta \right\} \xrightarrow[\ell \to \infty]{} 0$$

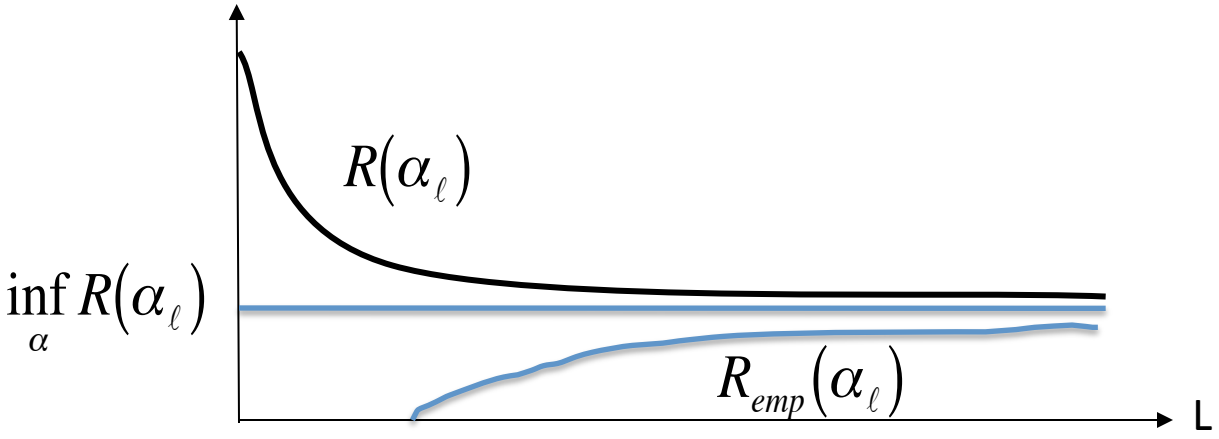
Consistency (Definition)

- **Definition:** we say that the ERM principle is *consistent* for $\{L(\mathbf{z},\alpha)\}, F(\mathbf{z})$ if the following two conditions hold:

$$(1) \quad R(\alpha_\ell) \xrightarrow[\ell\to\infty]{P} \inf_\alpha R(\alpha)$$

$$(2) \quad R_{emp}(\alpha_\ell) \xrightarrow[\ell\to\infty]{P} \inf_\alpha R(\alpha)$$

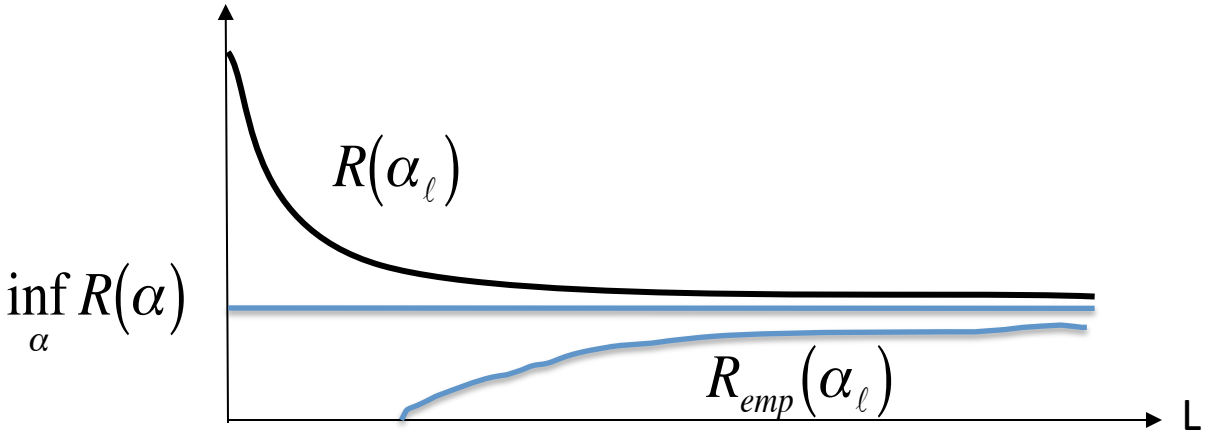
- Expected & Empirical risk must converge "in P" to minimal possible value of risk
- Why both achieved & estimated risk need to converge?



Two Conditions

• Suppose we generated an infinite sample from a given pdf. We fixed the order and we marked each example in the sample with a number {1…L}. For each iteration from 1 to L (infinity) we use the corresponding sample to do the following:

- Minimize risk on the sample (ERM) and obtain the decision surface (optimal set of parameters alpha_L) which yields the minimum value.
- Plug the optimal set of parameters into the integral with respect to the entire distribution and obtain the expected (achieved) risk value.
- Both the achieved & estimated risk need to converge to the smallest possible risk for a given set of functions (hence infimum or minimum over all alphas)



Trivial Consistency

- **Problem**: trivial cases of consistency
- Suppose ERM is ***not*** consistent for some set $\left\{L(\mathbf{z},\alpha)\right\}$
  - Add one "minorizing" function to the set such that: $\inf_{\alpha} L(\mathbf{z},\alpha) > \phi(\mathbf{z})$
  - For the extended set, ERM is consistent!
  - For every case, minimum of risk is attained at $\phi(\mathbf{z})$
- Problem since we are forced to take specific functions into account (consistency depends on whether such function exists)
- But we would like conditions that depend on general properties of a set

Strict Consistency

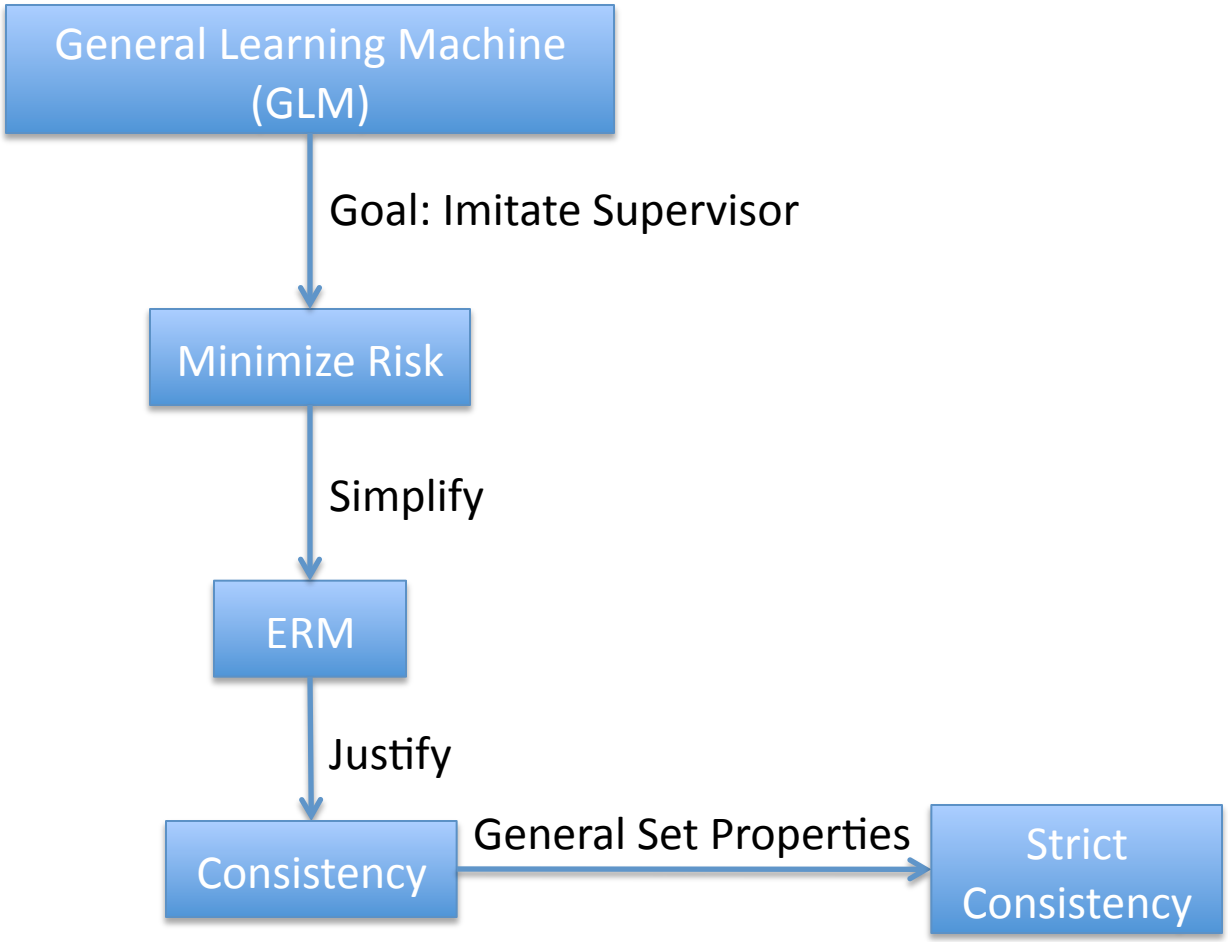
- **Definition:** we say that the ERM principle is *strictly consistent* for $\{L(\mathbf{z},\alpha)\}, F(\mathbf{z})$ if for any nonempty subset S(c) of this set, the convergence below is valid:

$$S(c) = \left\{\alpha : \quad \int L(\mathbf{z},\alpha)\, dF(\mathbf{z}) \geq c\right\}$$

$$\inf_{\alpha_\ell \in S(c)} R_{emp}(\alpha_\ell) \xrightarrow[\ell \to \infty]{P} \inf_{\alpha \in S(c)} R(\alpha)$$

- Trivial cases are excluded
- NOTE: previously two conditions (expected & empirical) but now just one
- Empirical convergence is sufficient since it implies expected (but not vice versa!)

Road Map (1)


General Learning Machine (GLM)
Goal: Imitate Supervisor
Minimize Risk
Simplify
ERM
Justify
Consistency
General Set Properties
Strict Consistency