pertinence
simple solutions for complex processes

www.pertinence.com

Introduction to Learning Theory

Olivier Bousquet
Pertinence
Paris, France
olivier.bousquet@pertinence.com

Machine Learning Summer School, Canberra, February 2006

pertinence

Outline

1. Learning Theory: What and Why?
   (a) Settings and assumptions
   (b) No Free Lunch
   (c) Consistency
2. Ingredients for Learning Bounds
   (a) Fundamental inequalities
   (b) Deviation/concentration inequalities
   (c) Union bound
3. Implications
   (a) How to use bounds
   (b) Connection to the design of algorithms

pertinence

Goal of this course

- Some thoughts about learning and its foundations (main focus)
- Keep things simple: basic techniques, finite sets of hypotheses
- Show some of the important phenomena, give insights into proof techniques
- Discuss about the meaning: how to use and how not to use a bound

pertinence

Not covered

- History of the domain
- General results, detailed proofs
- Advanced topics, e.g. influence of noise and loss functions

Learning Theory: What?

First of all, what is a theory?

> [THEORY] A set of statements or principles devised to **explain** a group of facts or phenomena, especially one that has been **repeatedly** tested or is widely accepted and can be used to **make predictions** about natural phenomena.

- Aim is to **model** a phenomenon (i.e. provide a schematic description of it, that accounts for its properties)
- so that we can **understand** it (i.e. use the model to further study the characteristics of the phenomenon)
- and **predict** it (i.e. derive consequences that can be tested)

Some more definitions

Difference between deduction and induction.

> [DEDUCTION] The process of reasoning in which a conclusion follows necessarily from the stated premises; inference by reasoning from the general to the specific.

This is how mathematicians prove theorems from axioms.

> [INDUCTION] The process of deriving general principles from particular facts or instances.

This is how physicists create theories from observing Nature.
Also: Transduction (from specific to specific in one step, without making the axioms/theory/model explicit)

What is a good theory?

- Intuitively, the goal is to provide a simple and concise account of observations.
- A good theory should be able to explain/predict many phenomena, with a simple model.
- The principles on which relies a theory are called assumptions. For example:

  > [Principle of Relativity] The Laws of Physics are the same in all inertial frames of reference
- Assumptions cannot be proved or disproved (there is no principle from which they are deduced) but their consequences can be tested.
- A good theory should make few assumptions.

What is Learning?

> [LEARNING] To gain knowledge, comprehension, or mastery of through experience or study.

- We will focus on experience, not study (too easy).
- Learning becomes interesting when what is gained is more than what was explicitly given, i.e. more than an accumulation of facts.

We will consider these as synonymous

- learning
- generalization
- induction
- theory building
- modelling

Recursion

- So far, we have investigated what is a theory and how it should be constructed.
- Why is this needed? (Do we study what is a theory when we study the theory of relativity?)
- Building a theory is the process of induction (i.e. learning).

→ Are we not going in circles?

Recursion

Rephrasing:

- Induction is the process of building theories
- Learning Theory's main focus is the phenomenon of induction
- How can we create a theory about theory building?

A lot of philosophical issues are involved, we need to formalize things in order to go further.

Learning Theory: Why?

[Kurt Lewin] There is nothing so practical as a good theory

As for all theories, learning theory should

- Try to understand systems that learn from data
- Provide a framework for studying their properties
- Allow to derive consequences in the form of "predictions" of which systems may work best

Hopefully, it should guide us toward designing better learning algorithms!

Inductive principles

Unfortunately, Induction is not like Deduction.

- Deduction can be justified (if the principles are correct, so do the consequences).
- Justifying Induction means justifying principles. This raises issues of philosophical nature.

Example 1: Probability of Sunrise Tomorrow

What is the probability p that the sun will rise tomorrow? (given we observed it rising on each of the previous d days)

Example 1: Probability of Sunrise Tomorrow

What is the probability p that the sun will rise tomorrow? (given we observed it rising on each of the previous d days)

- p cannot be defined (tomorrow is not identical to yesterday and there cannot be an experiment to test what happens tomorrow).
- $p = 1$, because the sun always rose in the past.
- $p = \frac{d+1}{d+2}$ obtained from Bayes rule (assuming uniform prior).
- Use physics to estimate the probability of the sun to explode tomorrow
  - ⋆ compute the proportion of stars that explode per day
  - ⋆ compare the sun to other stars with similar properties

Results are a high probability of rising again. Justification involves comparison with past "similar" situations (similar is highly subjective).

Example 2: extend a sequence of integers

You observe the sequence

1,2,4,7,...

What comes next?

Example 2: sequence of integers

528 hits on The On-Line Encyclopedia of Integer Sequences

- Maximum number of pieces formed when slicing a pancake with n cuts $u_{n+1} = u_n + n \Rightarrow$ 1,2,4,7,11,16... (A000124)
- $u_{n+2} = u_{n+1} + u_n + 1 \Rightarrow$ 1,2,4,7,12,20,... (A000071)
- Tribonacci numbers $u_{n+3} = u_{n+2} + u_{n+1} + u_n \Rightarrow$ 1,2,4,7,13,24,... (A000073)
- Binary expansion: 1,10,100,111,1000,1011,... odd number of 1's (Odious numbers) $\Rightarrow$ 1,2,4,7,8,11,13,14... (A000069)
- Or decimal expansions of π and e interleaved $\Rightarrow$ **1**,2,**4**,7,**1**,1,**5**,8,**9**
- and why not : $u_{n+1} = u_n$ for $n > 3 \Rightarrow$ 1,2,4,7,7,7...

$\longrightarrow$ which one is the simplest ?

Example 3: sequence of integers [Hutter]

Sequence:
$2, 3, 5, 7, 11, 13, 17, 19, 23, 29, 31, 37, 41, 43, 47, 53, 59, ?$
What comes next?

- 61, since this is the next prime
- 60, since this is the order of the next simple group

Conclusion: We prefer answer 61, since primes are a more familiar concept than simple groups.

Example 4: sequence of digits [Hutter]

Extend
$14159265358979323846264338327950288419716939937?$

- Looks random?!
- Frequency estimate: n = length of sequence. k_i = number of occured $i \Rightarrow$ Probability of next digit being i is $\frac{i}{n}$. Asymptotically $\frac{i}{n} \longrightarrow \frac{1}{10}$ (seems to be) true.
- But we have the strong feeling that (i.e. with high probability) the next digit will be 5 because the previous digits were the expansion of π.
- Conclusion: We prefer answer 5, since we see more structure in the sequence than just random digits.

Inductive Principle

- Is there a principle for providing a "good" answer?
- Occam's razor: Use the **simplest** explanation **consistent** with past data. In other terms: build the simplest theory that accounts for the observations and use it for predicting future observations.
- Issues:
  - ⋆ this cannot be justified or proved, it is a guidance that intuitively makes sense;
  - ⋆ simplicity is not an objective notion.

$\longrightarrow$ can we obtain general statements although there is no general principle?

Probability: a nice tool for reasoning

- Probability theory allows to formalize reasoning under uncertainty.
- Axioms
  - ⋆ $P(X) \geq 0$
  - ⋆ $P(\Omega) = 1$
  - ⋆ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- How are such quantities measured?

Probability interpretations

Interpretation = connection to the real world

- Frequentism: probabilities are relative frequencies. (e.g. the relative frequency of tossing head.)
- Objectivism: probabilities are real aspects of the world. (e.g. the probability that some atom decays in the next hour)
- Subjectivism: probabilities describe someone's degree of belief. (e.g. it is (im)plausible that extraterrestrians exist)

Examples from [Hutter]

Probabilities as Frequencies

- The frequentist interprets probabilities as relative frequencies.
- Repeat an experiment n times. If an event occurs $k(n)$ times, define the relative frequency of the event as $k(n)/n$.
- The limit $lim_{n\to\infty}k(n)/n$ is defined as the probability of the event.
- Easy but limited: not possible to perform such experiments in many cases!

Probabilities as Intrinsic Properties

- For the objectivist probabilities are real aspects of the world.
- The outcome of an experiment may be physically random.
- Probabilities give the expected frequency of each outcome if one were to repeat the experiment.
- Probabilities could be measured in a frequentist way, but they pre-exist.

Probabilities as Degrees of Belief

- For the subjectivist probabilities characterize someone's degree of belief in an event occurring.
- It is natural to assume that plausibilities/beliefs $Bel(\cdot|\cdot)$ can be represented by real numbers, that the rules qualitatively correspond to common sense, and that the rules are mathematically consistent.
- Cox's theorem: Beliefs have the same properties as probabilities

Bayes' Rule

- Allows to update probabilities/beliefs based on observations.
- Let D be some observation, $P(h)$ the probability of hypotheses prior to the observation, and $P(D|h)$ the probability to observe D under the hypothesis h.
- The posterior probability $P(h|D)$ of h given the observation is

$$P(h|D) = \frac{P(D|h)P(h)}{\sum_{h'} P(D|h')P(h')}$$

Note: this requires to have prior probabilities (i.e. someone's beliefs, or notion of simplicity).
In particular, $P(h)$ cannot be "measured" in a frequentist way.

Probabilities and Proofs

What do we gain using probability calculus?

- Nothing: if the starting point is wrong, the endpoint also.
- Probability-based results are not more valid than others (e.g. using Bayes rule does not guarantee anything)
- Probabilities simply provide a nice framework for reasoning (with uncertainty).

To relate probabilities to a real-world phenomenon, one needs an "interpretative hypothesis" (e.g. events with probability zero never occur).

The need for assumptions

- We mentioned we want to "predict" or "generalize"
- This can work only if
  - ⋆ the future looks like the past (prediction/forecasting)
  - ⋆ or the unseen looks like the seen (generalization)
- Hence we need to assume some kind of common underlying phenomenon

The need for assumptions (2)

- Our goal is to keep assumptions to a minimum
- Can we completely avoid assumptions?
- Yes if we aim at competitive results: we do not aim at predicting well, but predicting almost as well as someone else.

Hence there are two points of view:

- Given assumptions, what can we do at best?
- Given no assumptions, can we match others' performance?

Settings

Examples of commonly considered settings

1. Off-line supervised learning: training pairs are given, the goal is to produce a model that predicts well
2. Semi-supervised: training pairs and unlabeled data are given, the goal is to produce a model that predicts well
3. Transductive: training pairs and unlabeled data are given, the goal is to predict well on the unlabeled examples
4. On-line: repeated transductive learning (one new example at a time, prediction to be performed at each step)
5. Variants of on-line: actions instead of predictions (e.g. reinforcement learning)

Settings vs Assumptions

It is important to distinguish between the assumptions, the goals and the algorithms!

- Data generation mechanism: only relevant for proving theorems.
- Protocol: important for designing algorithms
- Success measure: sometimes impossible to practically measure. May serve as a guide for designing algorithms.
- Type of analysis: what we want to prove.
- Further restrictive assumptions: used for designing algorithms or proving restricted results.

Settings and Algorithms

$\longrightarrow$ Algorithms may do something different than optimizing directly the success (often impossible, or intractable)

$\longrightarrow$ Algorithms may perform well under various settings

Data Generation Mechanisms

- Bayes: the function is sampled and the data also
- IID: the data is sampled iid from unknown P
- Transduction: the data is fixed, the split is random
- On-line stochastic: random source but not necessarily independent (Markov...)
- On-line deterministic: no assumption, the data and its order is fixed
- On-line adversarial: the data is generated by an opponent

Protocols

- Off-line: examples given altogether
- On-line: examples given one at a time (order may or may not matter)

Information available to the learner may differ

- Error function
- Error value at past predictions
- Error of other predictors

Success Measures

Usually one consider loss measures

- Off-line: Expected error (average error on future instances generated in the same way)
- On-line: Cumulated error (sum of the errors made at each step)
- On-line: Expected future cumulated error (average cumulated error on future instances generated in the same way)

Type of analysis

- Worst-case: performance under the worst possible data generation mechanism
- Average-case: average over possible data generation mechanisms (requires some weighting)
- This case: performance on the given problem

What do we want to prove for a given learning algorithm?

- Expectation vs Probability
- Relative error bounds
- Oracle inequalities

Example: Bayesian Inference

- Data generation mechanism: function sampled from prior, data sampled iid from prior and labeled by function.
- Protocol: off-line, all examples given at once.
- Loss measure: expected error
- Type of analysis: average under the prior
- Further restrictive assumptions: noise of a specific form/intensity

These can be considered independently: one can use the same setting but perform a different analysis, or consider different generation mechanisms (e.g. IID)
Also, Bayes rule can be used in other settings. It is only optimal under all the above assumptions and for the above type of analysis.

Our Goal

We will

- consider a worst-case analysis
- try to avoid assumptions as much as possible
- look for best algorithms

The quantity of interest looks like

$$\min_{\text{predictor}} \max_{\text{problem}} \text{Loss(predictor, problem)}$$

Is this reasonable?

- We will show that this has no reason to be small. This will be the no-free-lunch theorems. They come in various flavours depending on how deep you dig.
- There are several ways in which you can make this quantity high: for fixed n, for varying n and fixed problem...

Minimax estimation

The classical minimax approach is to consider the following quantity:

$$\min_{\text{predictor}} \max_{\text{problem in a class}} \text{Loss(predictor, problem)}$$

Sometimes (if minimum loss is not zero) one considers

$$\min_{\text{predictor}} \max_{\text{problem}} \text{Loss(predictor, problem)} - \text{Best(problem)}$$

This is fine if we can guarantee that the data generation mechanism is indeed of the posited form. If this is not the case, this type of quantity is not very useful.

Learnability

- Another similar point of view is to tell whether a class can be learned.
- Several settings (identification in the limit, inductive inference, PAC learning)
- The question is whether, for a given class of functions, one can design an algorithm such that any function in the class would be recovered by the algorithm with enough training data.

Model identification

Sometimes one even measures the success by how similar is the predictive model to the "true" model.

- This is not interesting in practice (one can never verify the identity of two models since most often one can only access data)
- We prefer error minimization
- Similarly, you can hope but not ensure that the best function is in the class you consider

Updated goal

We prefer a competitive approach.
Indeed, we do not want to impose restrictions on the way the data is generated, but instead compare our performance to some references.

$$\min_{\text{predictor}} \max_{\text{problem,reference}} \text{Loss(predictor, problem)} - \text{Loss(reference, problem)}$$

This is an important point!

Upper and lower bounds

- To prove an upper bound: exhibit an algorithm for which you can prove a bound on the error which is independent of the data generation mechanism
- To prove a lower bound: for each possible algorithm, exhibit a data generation mechanism which misleads the algorithm

Notation

Let us introduce (finally) some notation

- $\mathcal{X}$ input space (often $\mathcal{X} \subset \mathbb{R}^d$)
- $\mathcal{Y}$ output space (often $\mathcal{Y} = \{0, 1\}$)
- $n \in \mathbb{N}$ sample size, i.e. number of observed pairs so far $(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y}$
- $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$ loss function (often $1_{[y \neq y']}$)

Notation (2)

Possible losses

- Cumulated loss for $g : \mathcal{X} \to \mathcal{Y}$:

$$L(g) = \sum_{i=1}^{n} \ell(g(x_i), y_i)$$

  (in the on-line setting, $g(x_i)$ may depend on the past observations)
- Expected loss

$$L(g) = \mathbb{E}\ell(g(X), Y)$$

- Predictor constructed from n observations: g_n
- Optimal function: $g^* = \arg\min L(g)$

Summary with notation

- Model identification: $\min d(g, g^*)$

$\to$ Risk minimization: $\min L(g)$

- Average case approach: $\min_g \sum_{P \in \mathcal{P}} L(g) - L(g^*)$
- Minimax approach: $\min \max_{P \in \mathcal{P}} L(g) - L(g^*)$

$\to$ Regret approach: $\min_g \max_P \max_{g' \in \mathcal{G}} L(g) - L(g')$

Lower bounds: weak and strong

Consider the quantity

$$\min_{g_n} \max_{P} \max_{g \in \mathcal{G}} L(g_n) - L(g)$$

- To prove a lower bound, we need to find a "bad" P.
- This P may depend on g_n and on n.
- Whether it depends on n or not makes a difference.
- Indeed, one is often interested as the decrease as n increases.

Strong: P does not depend on n

No free lunch

Simplest form due to [Wolpert96]

> On average over all possible data generation mechanisms (assuming there is a finite number of them), all algorithms have the same error when measured on future instances

$$\sum_{P \in \mathcal{P}} Loss(g_n, P) = \sum_{P \in \mathcal{P}} Loss(g'_n, P)$$

Interpretation

- If g_n is better than g'_n on a problem, it will be worse on another one
- One has to restrict the considered problems
- "my algorithm is better" means "the prior implemented by my algorithm is better suited to these databases"

So what do we do?

- Given values of f: $f(x_1), \ldots, f(x_n)$, what could be the value at some $x \neq x_1, \ldots, x_n$?
- If we allow all possible functions, it is clear that no generalization will be possible: all values are equally possible
- We need to make choices and express preferences or assumptions about the possible forms of functions we expect: we need to choose a prior
- Data only does not lead to generalization (if all possibilities look the same, nothing will be inferred).

The IID assumption

Examples are pairs (Xi,Yi) drawn

- independently
- from an unknown
- but fixed distribution P

This implies that there is some relationship between current and future instances.

Criticism of NFL1 under IID

- Finite setting not always realistic (although you always discretize to make things tractable on computers)
- In the finite setting, with enough examples one achieves minimal error (fixed problem, n increasing)
- Wolpert considers off-sample error. But there is a slight difference between expected and off-sample errors ($\sqrt{(1 + r \log n)/n}$ where r is the number of repetitions in the training set).
- In the continuous case ($\mathcal{X}$ continuous) they are the same.

Good and Bad news

Within the IID setting, considering the expected loss:

- Good news: there exists universally consistent algorithms
  - ⋆ When the sample size goes to infinity, the error of the algorithm converges to the best possible error (under P).
- Bad news: this does not help (Devroye et al. 96)
  - ⋆ NFL2: for a fixed sample size, the error of the algorithm can be arbitrarily close to the worst possible error (for some P)
  - ⋆ NFL3: the rate of convergence of a universally consistent algorithm can be arbitrarily slow (for some P)

Consistency under IID

Consistency means

$$\mathbb{E}L(g_n) - L(g^*) \to 0$$

when n increases.

- In a countable space, with enough data, all $x \in \mathcal{X}$ will eventually be observed (several times) so that consistency can be achieved without generalization!
- In a continuous space, measurability comes to the rescue: a measurable function can be approached by smoothed versions, and consistency can thus be obtained by smoothing (less and less as the sample size increases).

No-Free-Lunch 2

Binary classification setting.

Theorem 1. [DGL96, Thm 7.1] *For any $\epsilon > 0$, any n and any algorithm g_n, there exists P (with $L^* = 0$) such that*

$$\mathbb{E}L(g_n) \geq 1/2 - \epsilon.$$

Even though there are consistent algorithms, on finite samples their performance may be arbitrarily bad.
But this is a 'weak' statement: P depends on n (and ϵ).

No-Free-Lunch 3

Slow rates

Theorem 2. [DGL96, Thm 7.2] *For any non-increasing sequence a_n converging to zero, and any algorithm g_n, there exists P (with $L^* = 0$) such that*

$$\forall n,\ \mathbb{E}L(g_n) \geq a_n.$$

This is a 'strong' statement.

Proof ideas

- Construct a problem for which generalization is not possible: you cannot tell the labels of the points you have not seen yet
- Eventually you will have seen (almost) all the points so that you can still be consistent
- However, it may take a while to see enough of the points
  - ⋆ NFL2: can be done with a finite set (large enough).
  - ⋆ NFL3: same thing on a countable space (the probability mass of the unseen example has to be larger than a_n)

Comments

- Consistency is easy in the countable case
- Consistency relies on measurability in the continuous case
- No-Free-Lunch results only rely on not having seen enough data (and building problems where generalization is not possible: you cannot predict what you have not seen yet).

$\longrightarrow$ If we do not put assumptions, we cannot prove that we generalize, so what can we prove?

Estimation vs approximation

- NFL theorems make $L(g_n) - L(g^*)$ large
- Decomposition

$$L(g_n) - L(g^*) = (L(g_n) - L(g_{\mathcal{G}})) + (L(g_{\mathcal{G}}) - L(g^*))$$

where $g_{\mathcal{G}} = \arg\min_{\mathcal{G}} L(g)$

Estimation vs approximation (2)

- Approximation error $L(g_{\mathcal{G}}) - L(g^*)$
  - ⋆ the best we can do within $\mathcal{G}$.
  - ⋆ Deterministic quantity (i.e. does not depend on the data).
  - ⋆ Can only be estimated if we put assumptions
- Estimation error $L(g_n) - L(g_{\mathcal{G}}) = \max_{g \in \mathcal{G}} L(g_n) - L(g)$
  - ⋆ maximum regret = how much the data misleads us.
  - ⋆ Random quantity.
  - ⋆ Can be quantified even without assumptions on the data generating mechanism!

Summary

- Make as few assumptions as possible (here IID but this can be removed and leads to similar results)
- Focus on estimation error (or maximum regret)
- Theory will not justify your assumptions but will tell you how to best use them (in the sense of how you can avoid being misled by the data)

Ingredients for Learning Bounds

- Deviation inequalities
- Fundamental inequalities
- Union bound

Expectations versus Tails

Considering the random quantity $L(g_n) - \inf L(g)$, we can be interested in

- Expectation
$$\mathbb{E}L(g_n) - \inf L(g)$$
- Tail probability
$$\mathbb{P}\left[L(g_n) - \inf L(g) \geq \epsilon\right]$$
- The latter gives more information (one can recover the whole distribution)

We could even look at $\max_{(x_1,y_1),\ldots,(x_n,y_n)} L(g_n) - \inf L(g)$ (e.g. in the online setting).

Empirical loss

- Recall that $L(g) = \mathbb{E}\ell(g(X), Y)$
- This can be estimated from the data by the empirical loss as
$$L_n(g) = \frac{1}{n}\sum_{i=1}^{n} \ell(g(X_i), Y_i)$$
- One has $\mathbb{E}L_n(g) = L(g)$ (unbiased estimator for any fixed function)

Probability Tools (1)

Basic facts

- Union: $\mathbb{P}\left[A \text{ or } B\right] \leq \mathbb{P}\left[A\right] + \mathbb{P}\left[B\right]$
- Inclusion: If $A \Rightarrow B$, then $\mathbb{P}\left[A\right] \leq \mathbb{P}\left[B\right]$.
- Inversion: If $\mathbb{P}\left[X \geq t\right] \leq F(t)$ then with probability at least $1-\delta$, $X \leq F^{-1}(\delta)$.
- Expectation: If $X \geq 0$, $\mathbb{E}\left[X\right] = \int_0^\infty \mathbb{P}\left[X \geq t\right] dt$.

Probability Tools (2)

Basic inequalities

- Jensen: for f convex, $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$
- Markov: If $X \geq 0$ then for all $t > 0$, $\mathbb{P}[X \geq t] \leq \frac{\mathbb{E}[X]}{t}$
- Chebyshev: for $t > 0$, $\mathbb{P}[|X - \mathbb{E}[X]| \geq t] \leq \frac{\mathrm{Var}[X]}{t^2}$
- Chernoff: for all $t \in \mathbb{R}$, $\mathbb{P}[X \geq t] \leq \inf_{\lambda \geq 0} \mathbb{E}\left[e^{\lambda(X-t)}\right]$

Loss class

For convenience, let $Z_i = (X_i, Y_i)$ and $Z = (X, Y)$. Given $\mathcal{G}$ define the loss class

$$\mathcal{F} = \{f : (x, y) \mapsto 1_{[g(x) \neq y]} : g \in \mathcal{G}\}$$

Denote $Pf = \mathbb{E}[f(X, Y)]$ and $P_n f = \frac{1}{n}\sum_{i=1}^n f(X_i, Y_i)$

Quantity of interest:

$$Pf - P_n f$$

We will go back and forth between $\mathcal{F}$ and $\mathcal{G}$ (bijection)

Empirical process

Empirical process:

$$\{Pf - P_n f\}_{f \in \mathcal{F}}$$

- Process = collection of random variables (here indexed by functions in $\mathcal{F}$)
- Empirical = distribution of each random variable

$\Rightarrow$ Many techniques exist to control the supremum

$$\sup_{f \in \mathcal{F}} Pf - P_n f$$

The Law of Large Numbers

$$L(g) - L_n(g) = \mathbb{E}[f(Z)] - \frac{1}{n}\sum_{i=1}^n f(Z_i)$$

$\longrightarrow$ difference between the expectation and the empirical average of the r.v. $f(Z)$

Law of large numbers

$$\mathbb{P}\left[\lim_{n \to \infty} \frac{1}{n}\sum_{i=1}^n f(Z_i) - \mathbb{E}[f(Z)] = 0\right] = 1\,.$$

$\Rightarrow$ can we quantify it ?

Hoeffding's Inequality

Quantitative version of law of large numbers.

Assumes bounded random variables

Theorem 3. *Let $Z_1, \dots, Z_n$ be n i.i.d. random variables. If $f(Z) \in [a, b]$. Then for all $\epsilon > 0$, we have*

$$\mathbb{P}\left[\left|\frac{1}{n}\sum_{i=1}^{n} f(Z_i) - \mathbb{E}\left[f(Z)\right]\right| > \epsilon\right] \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right).$$

$\Rightarrow$ Let's rewrite it to better understand

Hoeffding's Inequality

Write

$$\delta = 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right)$$

Then

$$\mathbb{P}\left[|P_n f - Pf| > (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}\right] \leq \delta$$

or [Inversion] with probability at least $1-\delta$,

$$|P_n f - Pf| \leq (b-a)\sqrt{\frac{\log\frac{2}{\delta}}{2n}}$$

Hoeffding's inequality

Let's apply to $f(Z) = 1_{[g(X)\neq Y]}$.

For any g, and any $\delta > 0$, with probability at least $1-\delta$

$$L(g) \leq L_n(g) + \sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \qquad (1)$$

Notice that one has to consider a fixed function f and the probability is with respect to the sampling of the data.

If the function depends on the data this does not apply !

Limitations

- For each fixed function $f \in \mathcal{F}$, there is a set S of samples for which $Pf - P_n f \leq \sqrt{\frac{\log\frac{2}{\delta}}{2n}}$ $(\mathbb{P}[S] \geq 1-\delta)$
- They may be different for different functions
- The function chosen by the algorithm depends on the sample

$\Rightarrow$ For the observed sample, only some of the functions in $\mathcal{F}$ will satisfy this inequality !

Limitations

What we need to bound is

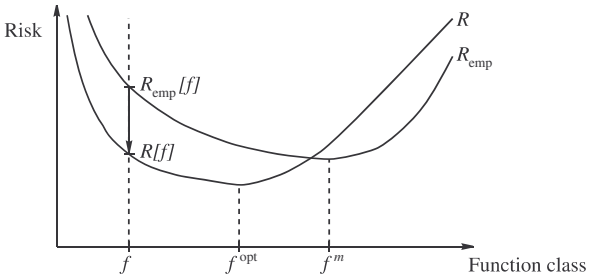
$$Pf_n - P_n f_n$$

where f_n is the function chosen by the algorithm based on the data.
For any fixed sample, there exists a function f such that

$$Pf - P_n f = 1$$

Take the function which is $f(Z_i) = 0$ on the data and $f(Z) = 1$ everywhere else.
This does not contradict Hoeffding but shows it is not enough

Limitations



Hoeffding's inequality quantifies differences for a fixed function

Fundamental Inequalities

- The goal is to upper bound $L(g_n) - L(g_{\mathcal{G}})$ by quantities that one can control
- The idea is to introduce the empirical estimate of the loss
- This is done for a specific algorithm

ERM in one class

Assume that $g_n = \arg\min_{g \in \mathcal{G}} L_n(g)$ (empirical risk minimizer)

- $$L(g_n) - \inf_{g \in \mathcal{G}} L(g) \leq 2 \sup_{g \in \mathcal{G}} |L(g) - L_n(g)| \,,$$
- $$\mathbb{E}\left[L(g_n) - \inf_{g \in \mathcal{G}} L(g)\right] \leq 2\mathbb{E}\left[\sup_{g \in \mathcal{G}} |L(g) - L_n(g)|\right] \,.$$

Arbitrary algorithm

- If $g_n \in \mathcal{G}$, one has

$$L(g_n) - L(g_{\mathcal{G}}) \leq L_n(g_n) - L_n(g_{\mathcal{G}}) + 2 \sup_{g \in \mathcal{G}} |L(g) - L_n(g)|$$

- Drop the $L_n(g_{\mathcal{G}})$ term and you obtain a bound which depends on the observed error of g_n.

$\rightarrow$ Not really interesting though (lhs may go to zero while rhs does not)

- only interesting for providing bounds of the form

$$L(g_n) \leq L_n(g_n) + \sup_{g \in \mathcal{G}} L(g) - L_n(g)$$

Refinement

$$L(g_n) - \inf_{g \in \mathcal{G}} L(g) \leq \sup_{g \in \mathcal{G}} (L(g) - L_n(g)) + L_n(g_{\mathcal{G}}) - L(g_{\mathcal{G}}) \,,$$

and

$$\mathbb{E}\left[L(g_n) - \inf_{g \in \mathcal{G}} L(g) \right] \leq \mathbb{E}\left[\sup_{g \in \mathcal{G}} (L(g) - L_n(g)) \right] .$$

Allows to use a one-sided supremum (sometimes better constants can be obtained)

Union Bound

Consider two functions f_1, f_2 and define

$$C_i = \{(x_1, y_1), \ldots, (x_n, y_n) : P f_i - P_n f_i > \epsilon\}$$

From Hoeffding's inequality, for each i

$$\mathbb{P}\left[C_i\right] \leq \delta$$

We want to bound the probability of being 'bad' for $i = 1$ or $i = 2$

$$\mathbb{P}\left[C_1 \cup C_2\right] \leq \mathbb{P}\left[C_1\right] + \mathbb{P}\left[C_2\right]$$

Finite Case

More generally

$$\mathbb{P}\left[C_1 \cup \ldots \cup C_N\right] \leq \sum_{i=1}^{N} \mathbb{P}\left[C_i\right]$$

We have

$$\begin{aligned} &\mathbb{P}\left[\exists f \in \{f_1, \ldots, f_N\} : P f - P_n f > \epsilon\right] \\ &\leq \sum_{i=1}^{N} \mathbb{P}\left[P f_i - P_n f_i > \epsilon\right] \\ &\leq N \exp\left(-2n\epsilon^2\right) \end{aligned}$$

Finite Case

We obtain, for $\mathcal{G} = \{g_1, \ldots, g_N\}$, for all $\delta > 0$

with probability at least $1 - \delta$,

$$\forall g \in \mathcal{G},\ L(g) - L_n(g) \leq \sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

Coding interpretation
$\log N$ is the number of bits to specify a function in $\mathcal{F}$

Consequence on the regret

- For the empirical risk minimizer in $\mathcal{G}$, one obtains

$$L(g_n) - L(g_{\mathcal{G}}) \leq 2\sqrt{\frac{\log N + \log \frac{1}{\delta}}{2n}}$$

- This is independent of the distribution (just assumes iid)
- Hence the worst case regret goes to zero: ERM matches the performance of the best function in the class

Summary

- Bounds are valid w.r.t. repeated sampling
- For a fixed function g, for most of the samples

$$L(g) - L_n(g) \approx 1/\sqrt{n}$$

- For most of the samples if $|\mathcal{G}| = N$

$$\sup_{g \in \mathcal{G}} (L(g) - L_n(g)) \approx \sqrt{\log N / n}$$

$\Rightarrow$ Extra variability because the chosen g_n changes with the data

Improvements

We obtained

$$\sup_{g \in \mathcal{G}} (L(g) - L_n(g)) \leq \sqrt{\frac{\log N + \log \frac{2}{\delta}}{2n}}$$

To be improved

- Hoeffding only uses boundedness, not the variance
- Union bound can be very loose
- Supremum is not what the algorithm chooses.

Main objective: refine the union bound

Binomial tails

- $P_n f \sim B(p,n)$ binomial distribution $p = Pf$
- $\mathbb{P}\left[Pf - P_n f \geq t\right] = \sum_{k=0}^{\lfloor n(p-t) \rfloor} \binom{n}{k} p^k (1-p)^{n-k}$
- Can be upper bounded
  - ⋆ Exponential $\left(\frac{1-p}{1-p-t}\right)^{n(1-p-t)} \left(\frac{p}{p+t}\right)^{n(p+t)}$
  - ⋆ Bennett $e^{-\frac{np}{1-p}((1-t/p)\log(1-t/p)+t/p)}$
  - ⋆ Bernstein $e^{-\frac{nt^2}{2p(1-p)+2t/3}}$
  - ⋆ Hoeffding e^{-2nt^2}

Tail behavior

- For small deviations, Gaussian behavior $\approx \exp(-nt^2/2p(1-p))$
  $\Rightarrow$ Gaussian with variance $p(1-p)$
- For large deviations, Poisson behavior $\approx \exp(-3nt/2)$
  $\Rightarrow$ Tails heavier than Gaussian
- Can upper bound with a Gaussian with large (maximum) variance $\exp(-2nt^2)$

Illustration (1)

Maximum variance ($p = 0.5$)

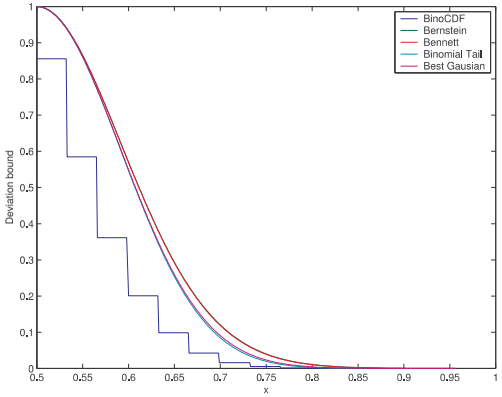
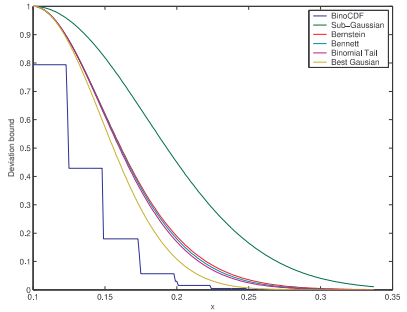
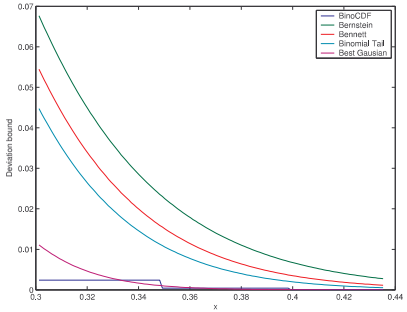


Illustration (2)

Small variance ($p = 0.1$)





Taking the variance into account

- Each function $f \in \mathcal{F}$ has a different variance $Pf(1-Pf) \leq Pf$.
- For each $f \in \mathcal{F}$, by Bernstein's inequality

$$Pf \leq P_n f + \sqrt{\frac{2Pf \log \frac{1}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}$$

- The Gaussian part dominates (for Pf not too small, or n large enough), it depends on Pf

Combining with union bound

With probability at least $1-\delta$

$$\forall g \in \mathcal{G},\ L(g) \leq L_n(g) + \sqrt{\frac{2L(g) \log \frac{N}{\delta}}{n}} + \frac{2 \log \frac{N}{\delta}}{3n}$$

Same trick, leads to the extra $\log N$ factor.

Consequences

From the fact

$$A \leq B + C\sqrt{A} \Rightarrow A \leq B + C^2 + \sqrt{B}C$$

we get

$$\forall g \in \mathcal{G},\ L(g) \leq L_n(g) + \sqrt{2L_n(g)\frac{\log \frac{N}{\delta}}{n}} + c\frac{\log \frac{N}{\delta}}{n}$$

Zero noise

Ideal situation

- g_n empirical risk minimizer
- Target function deterministic and belongs to $\mathcal{G}$
- $L(g^*) = 0$

In that case

- $L_n(g_n) = 0$

$\Rightarrow\ L(g_n) - L(g_{\mathcal{G}}) = O(\frac{\log N}{n})$.

Back to the union bound

- Recall that our goal was to improve the union bound
- First approach: use weights
- Second approach: use the variance

Weighted union bound (1)

For each $f \in \mathcal{F}$,

$$\mathbb{P}\left[Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}}\right] \leq \delta(f)$$

$$\mathbb{P}\left[\exists f \in \mathcal{F} : Pf - P_n f > \sqrt{\frac{\log \frac{1}{\delta(f)}}{2n}}\right] \leq \sum_{f \in \mathcal{F}} \delta(f)$$

Choose $\delta(f) = \delta p(f)$ with $\sum_{f \in \mathcal{F}} p(f) = 1$

Weighted union bound (2)

With probability at least $1 - \delta$,

$$\forall f \in \mathcal{F},\ Pf \leq P_n f + \sqrt{\frac{\log \frac{1}{p(f)} + \log \frac{1}{\delta}}{2n}}$$

- Applies to countably infinite $\mathcal{F}$
- Can put knowledge about the algorithm into $p(f)$
- But p chosen before seeing the data

Weighted union bound (3)

- Good p means good bound. The bound can be improved if you know ahead of time the chosen function (knowledge improves the bound)
- In the infinite case, how to choose the p (since it implies an ordering)

Union bound and variance (1)

Another way to use variance: even if the functions have large variance, bounds can be improved

- Consider a set of functions which make "similar" predictions
- This can be quantified by $\mathbb{P}(f_i(Z) \neq f_j(Z)) \leq \alpha$
- In this case the union bound can be refined

Union bound and variance (2)

- Write $Pf - P_n f = Pf_0 - P_n f_0 + (P - P_n)(f - f_0)$
- Hence deviations are those of a fixed function plus those of a function with variance bounded by α
- Result will have the form

$$Pf - P_n f \leq \sqrt{\frac{\log 1/\delta}{n}} + \sqrt{\frac{\alpha \log N/\delta}{n}} + o(1/\sqrt{n})$$

$\longrightarrow$ the complexity was measured by the cardinality of the class but it should involve the structure

Covers

- The goal is to exploit as much as possible the stucture of the class
- Repeat previous construction
  - ⋆ Find a subset of the functions $\{f_1, \ldots, f_d\}$ with $d << N$ such that

$$\forall f \in \mathcal{F}, \exists i \in \{1, \ldots, d\} \; s.t. \; \mathbb{P}(f_i(Z) \neq f(Z)) \leq \alpha$$

  - ⋆ Such a set is called an α-cover
- Leads to "covering numbers" bounds

Optimal union bound

- One cover: $\log N$ replaced by $\alpha \log N + \log N(\alpha)$
- Chaining: make covers at all scales, $\log N$ replaced by

$$\int \sqrt{\log N(\alpha)} d\alpha$$

- Generic chaining: progressively refine the covers in the regions where it is required
- Sometimes geometry can be taken into account directly (using Rademacher averages for example)

Summary

- $\log N$ is a very crude estimate of the complexity
- Using the geometry or structure of the class, one can get much better bounds
- Using refined deviations does not make such a huge difference (Bennett, Bernstein, Binomial : all have the same asymptotic behavior)

Occam's razor

- Bound with $\log N$ is usually called an Occam's razor bound: it indeed says the smaller the better.
- Sometimes it is misinterpreted: simplicity is taken directly.
- However, it depends on two things: how you weight hypotheses, and the geometry of the space!
- Weighting is an arbitrary choice, geometry is an intrinsic property: one should be careful not to mix them

Interpretation of bounds

- The iid assumption
- Possible usage of bounds

The IID assumption

- Whether it makes sense in real life is arguable
- However, similar results can be obtained without this assumption
  - ⋆ In the "transductive" setting: data is assumed to be fixed, but randomly splitted (this is exactly what experimentalists do)
  - ⋆ In the worst-case on-line setting: data is given, nothing is random. One can only get regret type bounds but they have the same form as above

What is a good bound ?

- Classification error between 0 and $1/2$
- Most theoretical bounds are useless (value $>> 1$)
- How to make them non-trivial ?

$\Rightarrow$ Here trivial does not mean easy but larger than 1

What is a good bound ?

- Depends on what you want to do with it
- Three levels of usage
  1. Quantitative
  2. Model selection
  3. Qualitative

First level

Obstacles

- Behavior of the error is complex
- Used techniques sharp in the asymptotic regime
- More precise techniques may exist but are much more messy
- Small bounds are unreadable

$\Rightarrow$ Hopeless ! use CV

Second level

Model selection

- Typical bounds behavior (picture)
- What matters is the location of the minimum

$\Rightarrow$ Little hope ! use CV if possible

Third level

Qualitative

- Use the quantities appearing in the bound to get inspiration for new algorithms
- Does not give the best choice of the parameters
- But indicates how complexity should be traded with errors
- Avoid a posteriori justifications !

$\Rightarrow$ Very reasonable !

Why a posteriori justifications are wrong ?

- Given a class of functions $\mathcal{F}$
- Define a (non-negative) functional $\Omega(f)$
- Obviously if $x \leq y$

$$\{\Omega(f) \leq x\} \subset \{\Omega(f) \leq y\}$$

- Hence $C\{\Omega(f) \leq x\}$ is a non-decreasing function of x !

$\Rightarrow$ Algorithm should minimize $\Omega(f)$!

$\Rightarrow$ Arbitrary !

Weighted union bound

- Choose p that gives more weight to functions satisfying certain properties
- Look at the bound you obtain: it tells that small p is better
- Build an algorithm that minimizes error $+ \log 1/p(f)$
- Claim that your algorithm is theoretically justified!

$\Rightarrow$ Arbitrary !

Summary

- Forget about the value of the bound
- Try to capture meaningful behavior
- Do not put quantities in by hand (or distinguish between what you put and what is new)
- Find what is responsible for deviations and how it influences them
- Do not hope to justify your prior but to get guidance on how to exploit it

Take home messages

- Induction cannot be justified, i.e. priors are needed but cannot be justified
- Only quantity that can be controlled without making assumptions: maximum regret
- Shape of bounds: $1/\sqrt{n}$, $\sqrt{\log N/n}$, $\sqrt{L_n(f)\log N/n}$, geometric structure
- Using bounds: distinguish between what are the assumptions, what was artificially put in the bound (e.g. weight) and what comes out from the learning phenomenon