Understanding Machine Learning – A theory Perspective

Shai Ben-David

University of Waterloo

MLSS at MPI Tubingen, 2017

Disclaimer – Warning ….

This talk is NOT about how cool machine learning is.

I am sure you are already convinced of that.

I am NOT going to show any videos of amazing applications of ML.

You will hear a lot about the great applications of ML throughout this MLSS.

I wish to focus on understanding the principles underlying Machine Learning.

High level view of (Statistical) Machine Learning

"The purpose of science is

to find meaningful simplicity

in the midst of

disorderly complexity"

Herbert Simon

More concretely

- Statistical learning is concerned with algorithms that detect ***meaningful regularities*** in large complex data sets.
- We focus on data that is ***too complex*** for humans to figure out its meaningful regularities.
- We consider the task of finding such regularities from ***random samples*** of the data population.

A typical setting

- Imagine you want a computer program to help decide which email messages are spam and which are important.
- Might represent each message by features.
  (e.g., return address, keywords, spelling, etc.)
- Train on a sample S of emails, labeled according to whether they are/aren't spam.
- Goal of algorithm is to produce good prediction rule (program) to classify future emails.

The concept learning setting

The learner is given some Training Sample

E.g.,

sales	sex	Mr.	bad spelling	known-sender	spam?
Y	N	Y	Y	N	Y
N	N	N	Y	Y	N
N	Y	N	N	N	Y
Y	N	N	N	Y	N
N	N	Y	N	Y	N
Y	N	N	Y	N	Y
N	N	Y	N	N	N
N	Y	N	Y	N	Y

The learner's output – a classification rule

Given data, some reasonable rules might be:

- Predict SPAM if [unknown AND (sex OR sales)]
- Predict SPAM if [sales + sex – known > 0].
- …

These kind of tasks are called Classification Prediction

Some typical classification prediction tasks

- Medical Diagnosis (*Patient info → High/Low risk*).
- Sequence-based classifications of proteins.
- Detection of fraudulent use of credit cards.
- Stock market prediction (*today's news → tomorrow's market trend*).

The formal setup (for label prediction tasks)

- Domain set – X
- Label set - Y (often {0,1})
- Training data – $S=((x_1,y_1), \ldots(x_m, y_m))$
- Learner's output – h: X ➔ Y

Data generation and measures of success

- An unknown distribution D generates instances (x_1, x_2, …) independently.
- An unknown function f: X ➔ Y labels them.
- The error of a classifier h is the probability (over D) that it will fail, $Pr_{x\sim D}[h(x) \neq f(x)]$

Empirical Risk Minimization (ERM)

Given a labeled sample

$$S=((x_1,y_1), \ldots(x_m, y_m))$$

and some candidate classifier h,

Define the empirical error of h as

$$L_S(h) = |\{i : h(x_i) \neq f(x_i)\}|/m$$

(the proportion of sample points on which h errs)

ERM – find h the minimizes $L_S(h)$.

Not so simple – Risk of Overfitting

- Given any training sample

  $S=((x_1,y_1), \ldots (x_m, y_m))$

- Let,

  $h(x)=y_i$ if $x=x_i$ for some $i \leq m$

  and $h(x)=0$ for any other x.

- Clearly $L_S(h) =0$.
- It is also pretty obvious that in many cases this h has high error probability.

The missing component –

- Leaners need of some **prior knowledge**

How is learning handled in nature (1)?

Bait Shyness in rats

Successful animal learning

The *Bait Shyness* phenomena in rats:

When rats encounter poisoned food, they learn very fast the causal relationship between the taste and smell of the food and sickness that follows a few hours later.

How is learning handled in nature (2)?
Pigeon Superstition (Skinner 1948)

What is the source of difference?

- In what way are the rats “smarter” than the pigeons?

Bait shyness and inductive bias

Garcia et al (1989) *:*

Replace the stimulus associated with the poisonous baits by making a sound when they taste it (rather than having a slightly different taste or smell).

How well do the rats pick the relation of sickness to bait in this experiment?

Surprisingly (?)

The rats fail to detect the association!

They do not refrain from eating when the same warning sound occurs again.

What about "improved rats"?

- Why aren't there rats that will also pay attention to the noise when they are eating?
- And to light, and temperature, time-of-day, and so on?
- Wouldn't such "improved rats" survive better?

Second thoughts about our improved rats

- But then every tasting of food will be an "outlier" in some respect….
- How will they know which tasting should be blamed for the sickness?

The Basic No Free Lunch principle

No learning is possible without **applying prior knowledge.**

(we will phrase and prove a precise statement later)

First type of prior knowledge – Hypothesis classes

- A *hypothesis class H* is a set of hypotheses.
- We re-define the ERM rule by searching only inside such a prescribed H.

- *$ERM_H(S)$ picks a classifier* h in H that minimizes the empirical error over members of H

Our first theorem

Theorem: (Guaranteed success for ERM_H)

Let H be a finite class, and assume further that the unknown labeling rule, f, is a member of H.

Then for every $\varepsilon, \delta > 0$,

if $m > (\log(|H|) + \log(1/\delta))/\varepsilon$,

With probability $> 1 - \delta$ (over the choice of S)

any $ERM_H(S)$ hypothesis has error below ε.

Proof

All we need to apply are two basic probability rules:

1) *The probability of the AND of independent events is the product of their probabilities.*

2) The "unions bound" – *the probability of the OR of any events is at most sum of their probabilities.*

Not only finite classes

- The same holds, for example, for the class H of all intervals over the real line.

(we will see a proof of that in the afternoon)

A formal definition of learnability

H is ***PAC Learnable*** if

there is a function $m_H : (0,1)^2 \rightarrow N$

and a learning algorithm A,

such that for every distribution D over X,

every $\varepsilon, \delta > 0$, and every f in H,

for samples S of size $m > m_H(\varepsilon, \delta)$

generated by D and labeled by f,

$$Pr[L_D(A(S)) > \varepsilon] < \delta$$

More realistic setups

Relaxing the realizability assumption.

We wish to model scenarios in which the learner does not have a priori knowledge of a class to which the true classifier belongs.

Furthermore, often the labels are not fully determined by the instance attributes.

General loss functions

Our learning formalism applies well beyond counting classification errors.

Let Z be any domain set.

and ℓ: H x Z ➔ *R* quantify the loss of a "model" h on an instance z.

Given a probability distribution P over Z

Let $L_P(h) = Ex_{z\sim P}(\ell(h, z))$

Examples of such losses

- The 0-1 classification loss:

  $\ell(h, (x,y)) = 0$ *if* $h(x)=y$ *and 1 otherwise.*

- *Regression square loss:*

  $\ell(h, (x,y)) = (y-h(x))^2$

- *K-means clustering loss:*

  $\ell((c_1, \ldots c_k), z) = \min_i (c_i - z)^2$

Agnostic PAC learnability

H is **Agnostic *PAC Learnable*** if

there is a function $m_H : (0,1)^2 \rightarrow N$

and a learning algorithm A,

such that for every distribution P over XxY

and every $\varepsilon, \delta > 0$,

for samples S of size $m > m_H(\varepsilon, \delta)$

generated by P,

$$Pr[L_P(A(S)) > \mathbf{Inf_{[h \text{ in } H]}\, L_P(h)} + \varepsilon] < \delta$$

Note the different philosophy

Rather than making an “absolute” statement that is guaranteed to hold only when certain assumptions are met (like realizability),

provide a weaker, **relative** guarantee, that is guaranteed to **always** hold.

General Empirical loss

- For any loss ℓ : H x Z ➔ *R as above and a finite domain subset S, define the empirical loss w.r.t.* $S=(z_1, \ldots z_m)$ *as* $L_S(h) = \Sigma_i \ \ell(h, z_i)/m$

Representative samples

We say that a sample S is

ε- representative of a class H w.r.t. a distribution P

If for every h in H

$$| L_S(h) - L_P(h)| < \varepsilon$$

Representative samples and ERM

Why care about representative samples?

If S is an *ε- representative of a class H w.r.t. a distribution P*

then for every $ERM_H(S)$ classifier h,

$$\mathbf{L_P(h) < Inf_{[h' \text{ in } H]}\, L_P(h') + 2\varepsilon}$$

Uniform Convergence Property

- We say that a class H has the

Uniform Convergence Property

if there is a function $m_H : (0,1)^2 \rightarrow N$

such that for every distribution P over Z

and every $\varepsilon, \delta > 0$,

with probability> $(1-\delta)$, samples S of size

$m > m_H(\varepsilon, \delta)$ generated by P, are

ε- representative of H w.r.t. P

Learning via Uniform Convergence

Corollary:

If a class H has the uniform convergence property then it is agnostic PAC learnable, and any ERM algorithm is a successful PAC learner for it.

Finite classes enjoy Unif. Conv.

- **Theorem:** If H is finite, then it has the uniform convergence property.
- **Proof:** Hoeffding implies Unif Conv for single h's and then the Union Bound handles the full class.
- **Corollary:** Finite classes are PAC learnable and ERM suffices for that.

Do we need the restriction to an H?

- Note that agnostic PAC learning requires only relative-to-H accuracy

  $L_D(A(S)) < \mathbf{Inf_{[h \text{ in } H]}}\ \mathbf{L_D(h)} + \varepsilon$

Why restrict to some H?

Can we have a *Universal Learner*, capable of competing with *every* function?

For the proof we need

The Hoeffding inequality:

Let $\theta_1 \ldots\ldots \theta_m$ be random variables over [0,1] with a common expectation μ, then

$$\Pr\left[\left| 1/m \sum_{i=1\ldots m} \theta_i - \mu \right| > \varepsilon \right] < 2 \exp(-2m\varepsilon^2)$$

We will apply it for $\theta_i = \ell(h, z_i)$

The No-Free-Lunch theorem

Let A be any learning algorithm over some domain set X.

Let m be < |X|/2, then

there is a distribution P over X x {0,1}

and f:X ➔{0,1} such that

1) $L_P(f)=0$ and
2) for P- samples S of size m

with probability > 1/7, $L_P(A(S))> 1/8$

The Bias-Complexity tradeoff

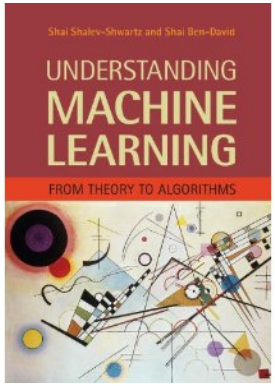
Corollary: Any class of infinite VC dimension is not PAC learnable – we cannot learn w.r.t. *the universal* H.

For every learner $L_P(A(S))$ can be viewed as the sum of

the Approximation error $\mathbf{Inf_{[h \text{ in } H]}\, L_D(h)}$

and **the Generalization error – the ε**

Want to know more?



See all 1 image(s)

Publisher: learn how customers can search inside this book.

Tell the Publisher!
I'd like to read this book on Kindle

Don't have a Kindle? Get your Kindle here, or download a **FREE** Kindle Reading App.

Understanding Machine Learning: From Theory to Algorithms [Hardcover]

Shai Shalev-Shwartz (Author), Shai Ben-David (Author)

List Price: ~~CDN$ 62.95~~
Price: CDN$ 50.36 Prime
You Save: CDN$ 12.59 (20%)
Pre-order Price Guarantee. Learn more.

This title has not yet been released.
You may pre-order it now and we will deliver it to you when it arrives.
Ships from and sold by **Amazon.ca**. Gift-wrap available.

Save Up to 90% on Textbooks
Hit the books in Amazon.ca's Textbook Store and save up to 90% on used textbooks and 35% on new textbooks. Learn more.

Vous voulez voir cette page en français ? Cliquez ici.

Quantity: 1

or
Sign in to turn on 1-Click ordering.

Share

Book Description

Publication Date: **May 31 2014** | ISBN-10: **1107057132** | ISBN-13: **978-1107057135**

Machine learning is one of the fastest growing areas of computer science, with far-reaching applications. The aim of this textbook is to introduce machine learning, and the algorithmic paradigms it offers, in a principled way. The book provides an extensive theoretical account of the fundamental ideas underlying machine learning and the mathematical derivations that transform these principles into practical algorithms. Following a presentation of the basics of the field, the book covers a wide array of central topics that have not been addressed by previous textbooks. These include a discussion of the computational complexity of learning and the concepts of convexity and stability; important algorithmic paradigms including stochastic gradient descent, neural networks, and structured output learning; and emerging theoretical concepts such as the PAC-Bayes approach and compression-based bounds. Designed for an advanced undergraduate or beginning graduate course, the text makes the fundamentals and algorithms of machine learning accessible to students and non-expert readers in statistics, computer science, mathematics, and engineering.