Algorithmic Foundations of Learning

Lecture 1
Introduction

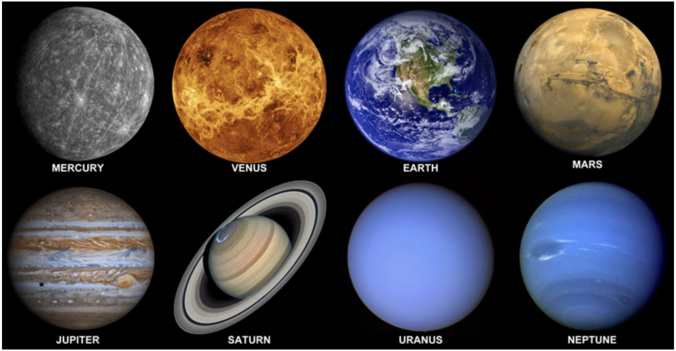
Patrick Rebeschini

Department of Statistics
University of Oxford

"New science is based on maximum likelihood rather than certainty"

Arthur C. Clarke and Gentry Lee, Rama Series Book 2, 1989

Old science...



http://eightplanets.org/planets.html

- Assume you are an astronomer in the 16th century
- You use observations on planets' movements to develop physical models
- **Q:** Assume you have $n = 100$ observations.
  How many observations you need to get 3 times better accuracy?

... also relies on statistics!

- Measurements in physics are modeled as random variables
- Central Limit Theorem: if $X_1, X_2, \ldots$ are i.i.d. $\mu = \mathbf{E} X_1$ $\sigma^2 = \mathbf{Var} X_1 < \infty$

$$\sqrt{n} \left\{ \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right\} \Longrightarrow \mathcal{N}(0, \sigma^2)$$

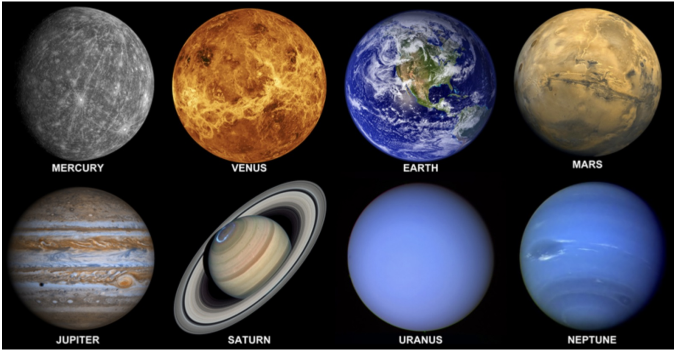
- A non-asymptotic statement (see **Problem 1.1** in the Problem Sheets):

$$\sqrt{\mathbf{E}\left[\left(\frac{1}{n} \sum_{i=1}^{n} X_i - \mu\right)^2\right]} = \frac{\sigma}{\sqrt{n}}$$

- General phenomenon that permeates statistics:

$$\text{Notion of error} \lesssim \frac{1}{\sqrt{n}}$$

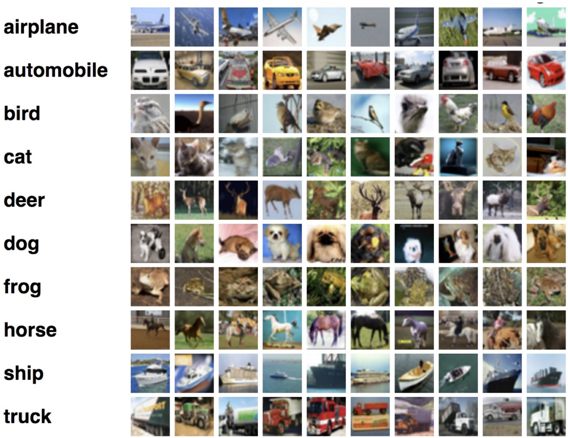
Old science...



http://eightplanets.org/planets.html

- **Q:** Assume you have $n = 100$ observations.
  How many observations you need to get 3 times better accuracy?
- **A:** You need $n' = 900$ observations
- The fact that accuracy growth quadratically and not linearly is **striking**

New Science?



Offline learning: prediction

Given a batch of observations (images & labels) interested in predicting the label of a new image

New Science?

User 1	★★★		★★★	
User 2	★★	★★★★		
User 3		★★★	★★	★★★★★

Offline learning: estimation
Given a batch of observations (users & ratings)
interested in estimating the missing ratings in a recommendation system

New Science?

Online learning
Given a sequence of dynamic observations (game stages) interested in learning the best action

New Science

Machine learning: a "machine" (algorithm) that "learns" patterns from data

NB: "Artificial Intelligence" misleading? (talk by Michael Jordan: Link)

In this course we will cover three main learning paradigms:

- Offline statistical learning: prediction
- Offline statistical learning: estimation
- Online statistical learning

$$\text{Notion of error} \lesssim \frac{1}{\sqrt{n}}$$

In machine learning we can beat the slow rate $\frac{1}{\sqrt{n}}$ and get up to the fast rate $\frac{1}{n}$

Q. What about the **dimensionality** of the data? (old science $\neq$ new science)

Goals of this course

1. **Statistics:** Derive error bounds of the following type:

$$\text{Notion of error} \lesssim \frac{f(\text{dimension})}{n^{\alpha}}$$

- dimension (old science): 6 (degrees of freedom in Newtonian physics)
- dimension (new science): can be $\gg 10^6$ (e.g., number of pixels in an image)
- Ideally: $f(\text{dimension}) \ll \text{dimension}$, e.g., $f(\text{dimension}) \sim \log(\text{dimension})$

2. **Computation:** Understand number of basic computations required to solve problem up to the level of the statistical precision

$$\text{Run time} \sim g(n, \text{dimension})$$

Goals of this course

3. **Theory:**
   Develop non-asymptotic methods for studying random structures in:
   - high-dimensional probability
   - high-dimensional statistics
   - high-dimensional optimization

   Many settings in machine learning are encoded in the general formulation:

$$\text{Minimize}_{a\in\mathcal{A}} \;\; \mathbf{E}\,\ell(a, Z)$$

   - supervised learning (regression, classification, etc.)
   - unsupervised learning (k-means, etc.)
   - density estimation
   - ...

Statistical/computational learning theory

Problem formulation (out-of-sample prediction):

- Given n data $(X_1, Y_1), \ldots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $\mathbf{P}$ (**unknown**)
- Consider the *population risk* $r(a) = \mathbf{E}\, \phi(a(X), Y)$

Goal: Compute $A \in \sigma\{(X_i, Y_i)_{i=1}^n\}$ such that $\underbrace{r(A) - \inf_a r(a)}_{\text{excess risk}}$ is **small**

What does it mean to solve the problem **optimally**?

- **Statistics:** A is minimax-optimal w.r.t. the class of distrib. $\mathcal{P}$ if

$$\mathbf{E}\, r(A) - \inf_a r(a) \quad \sim \inf_{A \in \sigma\{Z_1, \ldots, Z_n\}} \sup_{\mathbf{P} \in \mathcal{P}} \left\{ \mathbf{E}\, r(A) - \inf_a r(a) \right\}$$

- **Runtime:** Computing A takes same time to read the data, i.e. $O(nd)$ cost
- **Memory:** Storing $O(1)$ data point at a time, i.e. $O(d)$ storage cost
- **Distributed computations:** Runtime $O(1/m)$ if we have m machines
- (communication, privacy, robustness...)

Offline statistical learning: prediction

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from unknown distribution
2. Choose action $A \in \mathcal{A} \subseteq \mathcal{B}$
3. Suffer an expected/population loss/risk $r(A)$, where

$$\boxed{a \in \mathcal{B} \longrightarrow r(a) := \mathbf{E}\, \ell(a, Z)}$$

with ℓ is an prediction loss function and Z is a new test data point

Goal: Minimize the estimation error defined by the following decomposition

$$\boxed{\underbrace{r(A) - \inf_{a \in \mathcal{B}} r(a)}_{\text{excess risk}} = \underbrace{r(A) - \inf_{a \in \mathcal{A}} r(a)}_{\text{estimation error}} + \underbrace{\inf_{a \in \mathcal{A}} r(a) - \inf_{a \in \mathcal{B}} r(a)}_{\text{approximation error}}}$$

as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Note: Estimation/Approximation trade-off, a.k.a. complexity/bias

Goal - Applications

- The data distribution is unknown so also the risk r can not be computed
- Nevertheless, r is used as a way to assess the performance of the algorithm
- **Goal:** Derive upper bounds for the estimation error
- **Bounds in expectation:**

$$\mathbf{E}\, r(A) - r(a^\star) \leq \boxed{\texttt{Expectation}}$$

- **Bounds in probability:** For any $\varepsilon \geq 0$,

$$\mathbf{P}\Big(r(A) - r(a^\star) \geq \varepsilon\Big) \leq \boxed{\texttt{UpperTail}(\varepsilon)}$$

or, equivalently, for any $\delta \in [0, 1]$,

$$\mathbf{P}\Big(r(A) - r(a^\star) < \boxed{\texttt{UpperTail}^{-1}(\delta)}\Big) \geq 1 - \delta$$

ERM and Uniform Learning

- A natural framework is given by the empirical risk minimization (ERM)

$$\boxed{a \in \mathcal{B} \longrightarrow R(a) := \frac{1}{n}\sum_{i=1}^{n} \ell(a, Z_i)}$$

- A natural algorithm is given by the minimizer of the ERM

$$A^\star \in \operatorname*{argmin}_{a \in \mathcal{A}} R(a)$$

- **Uniform Learning:** The estimation error is bounded by

$$\boxed{\underbrace{r(A^\star) - r(a^\star)}_{\text{estimation error for ERM}} \leq \underbrace{\sup_{a \in \mathcal{A}}\{r(a) - R(a)\} + \sup_{a \in \mathcal{A}}\{R(a) - r(a)\}}_{\texttt{Statistics}}}$$

- Statistical Learning deals with bounding the `Statistics` term (Vapnik 1995)

- **Generalization Error:** $r(a) - R(a) \approx \frac{1}{n^{(\text{test})}} \sum_{i=1}^{n^{(\text{test})}} \ell(a, Z_i^{(\text{test})}) - \frac{1}{n}\sum_{i=1}^{n} \ell(a, Z_i)$

Goal - Theory

To analyse the ERM algorithm, we need to develop tools to:

- Control the suprema of random processes:

$$\mathbf{E} f(Z_1, \ldots, Z_n) \leq \boxed{\;?\;}$$

with $f(Z_1, \ldots, Z_n) = \sup_{a \in \mathcal{A}} \{R(a) - r(a)\}$

- Control the concentration of random processes:

$$\mathbf{P}\Big(f(Z_1, \ldots, Z_n) - \mathbf{E}\, f(Z_1, \ldots, Z_n) \geq \varepsilon\Big) \leq \boxed{\texttt{UpperTail}_f(\varepsilon)}$$

$$\mathbf{P}\Big(f(Z_1, \ldots, Z_n) - \mathbf{E}\, f(Z_1, \ldots, Z_n) < \boxed{\texttt{UpperTail}_f^{-1}(\delta)}\Big) \geq 1 - \delta$$

Q. Can the ERM rule/algorithm $A^\star$ be computed?
(we depart from classical learning theory and also consider computational issues)

Computational aspects

- The ERM is in general intractable
- We need to approximately compute it
- We will consider stochastic optimisation methods to minimize R.
- New error decomposition that highlight the statistical/computational parts

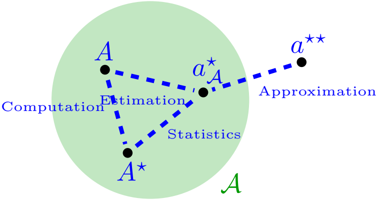
$$\boxed{r(A) - r(a^\star) \le \underbrace{R(A) - R(A^\star)}_{\texttt{Optimization}} + \underbrace{\sup_{a \in \mathcal{A}}\{r(a) - R(a)\} + \sup_{a \in \mathcal{A}}\{R(a) - r(a)\}}_{\texttt{Statistics}}}$$

- Key insight (Bousquet and Bottou 2008)

$$\boxed{\texttt{Bound on Optimisation}} \sim \boxed{\texttt{Bound on Statistics}}$$

It is only necessary to run an optimization algorithm until we are guaranteed to find a estimator with an accuracy of the same order as the statistical fluctuations of the problem

Explicit regularization: uniform convergence



- Estimation/approximation: $r(A) - r(a^{\star\star}) = \underbrace{r(A) - r(a^{\star})}_{\text{Estimation}} + \underbrace{r(a^{\star}) - r(a^{\star\star})}_{\text{Approximation}}$

- Classical error decomposition for estimation error:

$$\underbrace{r(A) - r(a^{\star})}_{\text{Estimation}} = r(A) - R(A) + R(A) - R(A^{\star}) + \underbrace{R(A^{\star}) - R(a^{\star})}_{\leq 0} + R(a^{\star}) - r(a^{\star})$$

$$\boxed{r(A) - r(a^{\star\star}) \leq \underbrace{2 \sup_{a \in \mathcal{A}} |r(a) - R(a)|}_{\text{Statistics}} + \underbrace{R(A) - R(A^{\star})}_{\text{Computation}} + \underbrace{r(a^{\star}) - r(a^{\star\star})}_{\text{Approximation}}}$$

Offline statistical learning: estimation

1. Observe training data $Z_1, \ldots, Z_n$ i.i.d. from distr. parametrized by $a^\star \in \mathcal{A}$
2. Choose a parameter $A \in \mathcal{A}$
3. Suffer a loss $\ell(A, a^\star)$ where ℓ is an estimation loss function

Goal: Minimize the estimation loss $\ell(A, a^\star)$ as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

Online statistical learning

At every time step $t = 1, 2, \ldots, n$:

1. Choose an action $A_t \in \mathcal{A}$
2. A dynamic data point Z_t is sampled from an unknown distribution
3. Suffer an expected/population loss/risk $r(A_t)$, where

$$\boxed{a \in \mathcal{B} \longrightarrow r(a) := \mathbf{E}\,\ell(a, Z)}$$

with ℓ a prediction loss function and Z is a new data point

Goal: Minimize the (normalized) (pseudo-)regret defined as

$$\boxed{\frac{1}{n} \sum_{t=1}^{n} r(A_t) - \inf_{a \in \mathcal{A}} r(a)}$$

as a function of n and notions of "complexity" of the set $\mathcal{A}$ of the function ℓ

On the course

- Slides provide the high-level narrative and highlight the main results
- Lecture notes are self-contained and contain <u>more</u> information than slides
- Main ideas of proofs and graphical illustrations are given during lectures
- <u>Any material covered in the lecture notes is fair game for the exam</u>
- Students are expected to study lecture notes in details,
<u>even if the lecturer does not fully cover them during lecture</u>

This is a theory course!