Bayesian Learning

Benjamin Guedj, Ph.D.

https://bguedj.github.io
Inria Lille - Nord Europe

2017–2018

benjamin.guedj@inria.fr

https://bguedj.github.io

10h de cours (5, 6, 12, 19 et 20 mars 2018)

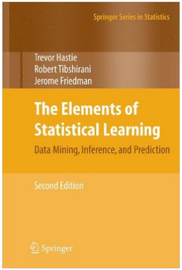
Projet à rendre. Deadline : **vendredi 30 mars à 18h**.

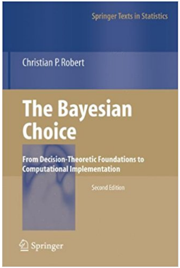
References

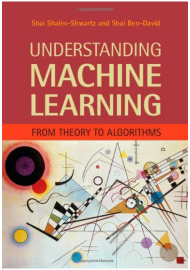
T. Hastie, R. Tibshirani and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, Springer, 2009. [Link]

C. P. Robert. *The Bayesian Choice*, Springer, 2007. [Link]

S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*, Cambridge University Press, 2014. [Link]







Outline

The rising of AI

Introduction

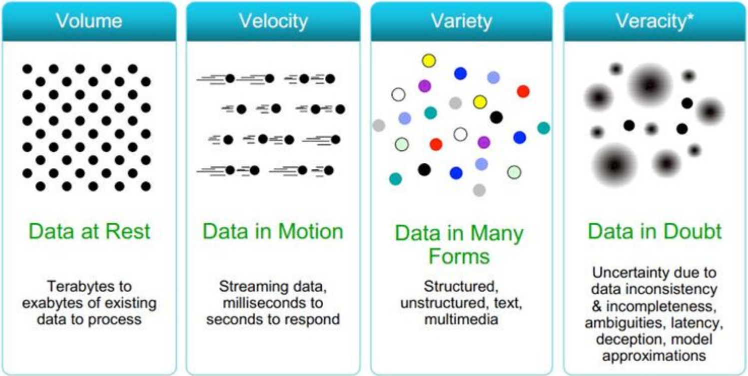
“We are drowning in information and starving for knowledge.”

-Rutherford D. Roger

It is vital to remember that information - in the sense of raw data - is not knowledge, that knowledge is not wisdom, and that wisdom is not foresight. But information is the first essential step to all of these.

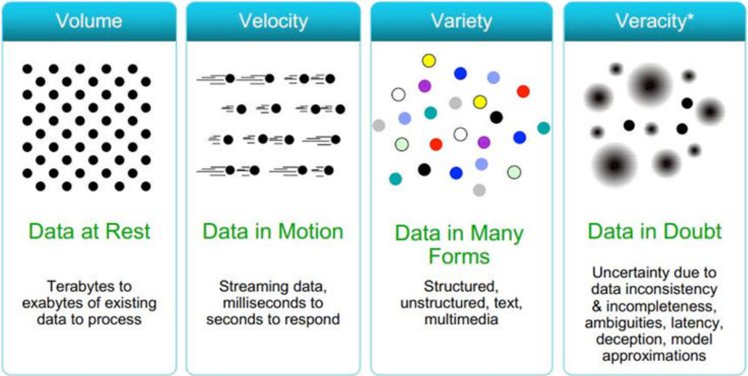
Arthur C Clarke

Big Data 4 V's



Volume
Velocity
Variety
Veracity*
Data at Rest
Terabytes to exabytes of existing data to process
Data in Motion
Streaming data, milliseconds to seconds to respond
Data in Many Forms
Structured, unstructured, text, multimedia
Data in Doubt
Uncertainty due to data inconsistency & incompleteness, ambiguities, latency, deception, model approximations

Big Data 4 V's



$\longrightarrow$ Value (\$)

Data Scientists: 100,000 jobs by 2020. Demand is expected to exceed supply by 50 to 60% (McKinsey, 2015)

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day

Volume / Velocity

i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

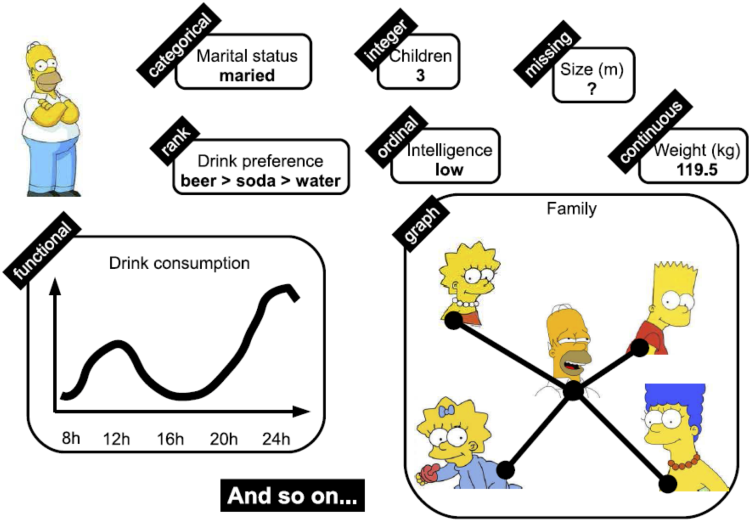
- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day
- Human brain may store about 2.5 petabytes of binary data

Volume / Velocity

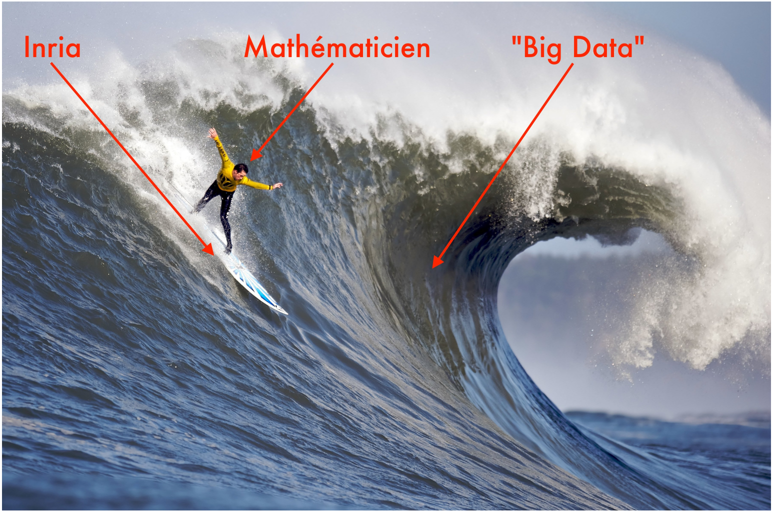
i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

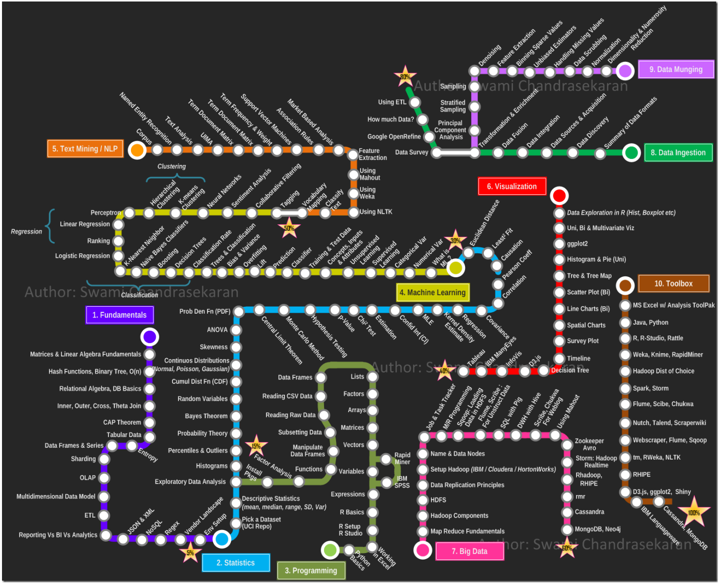
- Google: 24 petabytes/day
- Facebook: 10 terabytes/day
- Twitter: 7 terabytes/day
- Large Hadron Collider: 40 terabytes/second
- Google Street View: 20 petabytes
- AT&T network: 30 petabytes/day
- Human brain may store about 2.5 petabytes of binary data
- ...

Variety / Veracity

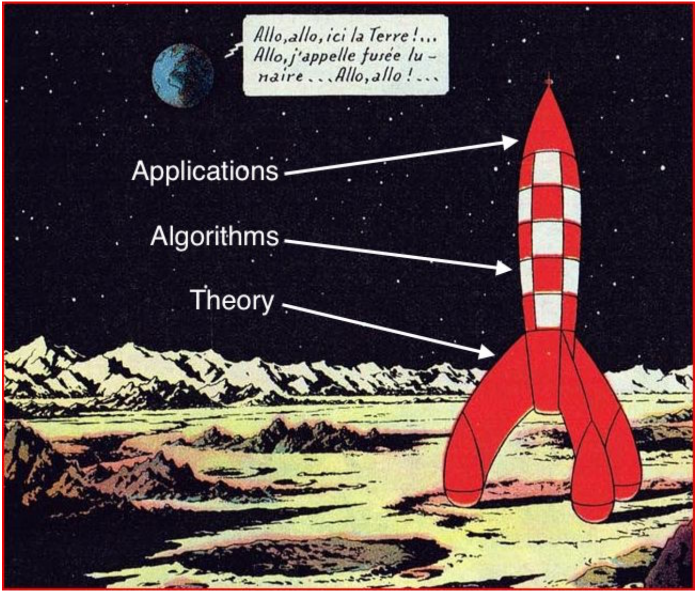


My job (allegory)





1. Fundamentals
Matrices & Linear Algebra Fundamentals
Hash Functions, Binary Tree, O(n)
Relational Algebra, DB Basics
Inner, Outer, Cross, Theta Join
CAP Theorem
Tabular Data
Data Frames & Series
Entropy
Sharding
OLAP
Multidimensional Data Model
ETL
Reporting Vs BI Vs Analytics
JSON & XML
NoSQL
Regex
Vendor Landscape
Env Setup
5%
2. Statistics
Prob Den Fn (PDF)
ANOVA
Skewness
Continuos Distributions (Normal, Poisson, Gaussian)
Cumul Dist Fn (CDF)
Random Variables
Bayes Theorem
Probability Theory
Percentiles & Outliers
Histograms
Exploratory Data Analysis
Descriptive Statistics (mean, median, range, SD, Var)
Pick a Dataset (UCI Repo)
Central Limit Theorem
Monte Carlo Method
Hypothesis Testing
p-Value
Chi² Test
Estimation
Confid Int (CI)
MLE
Kernel Density Estimate
Regression
Covariance
Correlation
Pearson Coeff
Causation
Least² Fit
Euclidean Distance
3. Programming
Python Basics
Working in Excel
R Setup R Studio
R Basics
Expressions
Variables
Vectors
Matrices
Arrays
Factors
Lists
Data Frames
Reading CSV Data
Reading Raw Data
Subsetting Data
Manipulate Data Frames
Functions
Factor Analysis
Install Pkgs
15%
Rapid Miner
IBM SPSS
4. Machine Learning
What is ML?
30%
Numerical Var
Categorical Var
Supervised Learning
Unsupervised Learning
Concepts, Inputs & Attributes
Training & Test Data
Classifier
Prediction
Lift
Overfitting
Bias & Variance
Trees & Classification
Classification Rate
Decision Trees
Boosting
Naive Bayes Classifiers
K-Nearest Neighbor
Logistic Regression
Ranking
Linear Regression
Perceptron
Regression
Classification
Hierarchical Clustering
K-means Clustering
Clustering
Neural Networks
Sentiment Analysis
Collaborative Filtering
Tagging
50%
Vocabulary Mapping
Classify Text
5. Text Mining / NLP
Corpus
Named Entity Recognition
Text Analysis
UIMA
Term Document Matrix
Term Frequency & Weight
Support Vector Machines
Association Rules
Market Based Analysis
Feature Extraction
Using Mahout
Using Weka
Using NLTK
6. Visualization
Data Exploration in R (Hist, Boxplot etc)
Uni, Bi & Multivariate Viz
ggplot2
Histogram & Pie (Uni)
Tree & Tree Map
Scatter Plot (Bi)
Line Charts (Bi)
Spatial Charts
Survey Plot
Timeline
Decision Tree
D3.js
InfoVis
IBM ManyEyes
Tableau
40%
7. Big Data
Map Reduce Fundamentals
Hadoop Components
HDFS
Data Replication Principles
Setup Hadoop (IBM / Cloudera / HortonWorks)
Name & Data Nodes
Job & Task Tracker
MR Programming
Sqoop: Loading Data in HDFS
Flume, Scribe: For Unstruct Data
SQL with Pig
DWH with Hive
Scribe, Chukwa For Weblog
Using Mahout
Zookeeper Avro
Storm: Hadoop Realtime
Rhadoop, RHIPE
rmr
Cassandra
MongoDB, Neo4j
60%
8. Data Ingestion
Summary of Data Formats
Data Discovery
Data Sources & Acquisition
Data Integration
Data Fusion
Transformation & Enrichment
Data Survey
Google OpenRefine
How much Data?
Using ETL
80%
9. Data Munging
Dimensionality & Numerosity Reduction
Normalization
Data Scrubbing
Handling Missing Values
Unbiased Estimators
Binning Sparse Values
Feature Extraction
Denoising
Sampling
Stratified Sampling
Principal Component Analysis
10. Toolbox
MS Excel w/ Analysis ToolPak
Java, Python
R, R-Studio, Rattle
Weka, Knime, RapidMiner
Hadoop Dist of Choice
Spark, Storm
Flume, Scibe, Chukwa
Nutch, Talend, Scraperwiki
Webscraper, Flume, Sqoop
tm, RWeka, NLTK
RHIPE
D3.js, ggplot2, Shiny
IBM Languageware
Cassandra, MongoDB
100%
Author: Swami Chandrasekaran


Allo, allo, ici la Terre !... Allo, j'appelle fusée lu-naire ... Allo, allo ! ...
Applications
Algorithms
Theory

A foretaste of Learning Theory

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

In the (~~rather~~ not so?) long term: mimic the inductive functioning of the humain brain to develop an artificial intelligence.

{Statistical,Machine} Learning: building automatic procedures to infer general rules from examples.

In the (~~rather~~ not so?) long term: mimic the inductive functioning of the humain brain to develop an artificial intelligence.

In the Big Data Era, very dynamic field at the crossroads of Computer Science, Optimization and Statistics.

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

We want to infer the link between the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{Y}$, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

Probabilistic framework: n-sample $\mathcal{D}_n = (\mathbf{X}_i, \mathbf{Y}_i)_{i=1}^n$ of i.i.d. replications of some random variable

$$(\mathbf{X}, \mathbf{Y}) \in \mathcal{X} \times \mathcal{Y}, \quad \dim(\mathcal{X}) = d.$$

We want to infer the link between the explanatory variable $\mathbf{X}$ and the response variable $\mathbf{Y}$, *i.e.*, use $\mathcal{D}_n$ to build up $\hat{\phi}$ such that $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

- Classification: $\mathcal{Y}$ is discrete.
- Regression: $\mathcal{Y}$ is a continuum.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathcal{X} = \mathbb{R}^d$ where d and n may be (extremely) large.

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathcal{X} = \mathbb{R}^d$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathcal{X} = \mathbb{R}^d$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$

- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathcal{X} = \mathbb{R}^d$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$
- Fat data: $d \gg n$

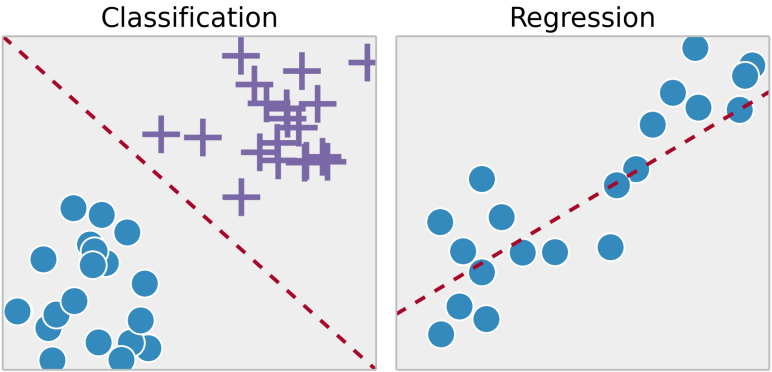
- Online learning: observations are revealed over time.
- Batch learning: all observations are revealed at once.

Big Data Era: easy/cheap to collect massive amounts of data, hence typically $\mathcal{X} = \mathbb{R}^d$ where d and n may be (extremely) large.

- Good ol' statistics: n and d small
- Tall data: $n \gg d$
- Fat data: $d \gg n$
- Big/massive data: n and d huge

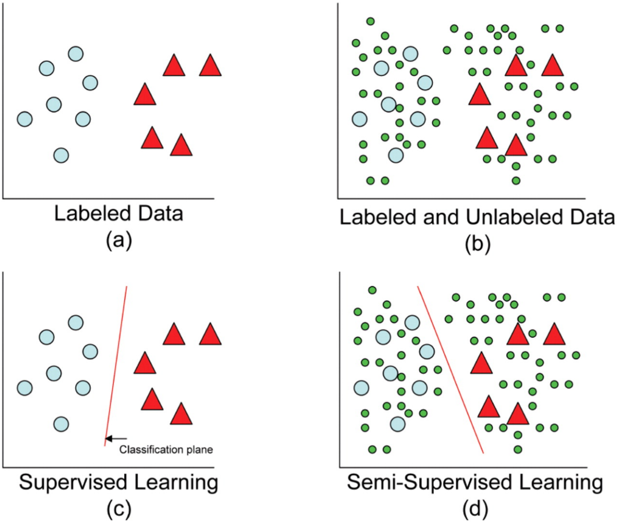
- Supervised learning: all of the $\mathbf{Y}_i$s are observed.

- Supervised learning: all of the $\mathbf{Y}_i$s are observed.



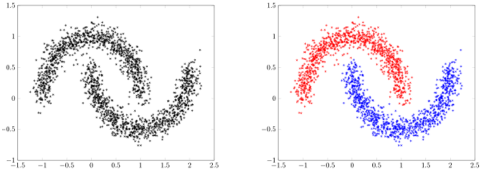
- Semi-supervised learning: some of the $\mathbf{Y}_i$s are observed (labeling is expensive or difficult).

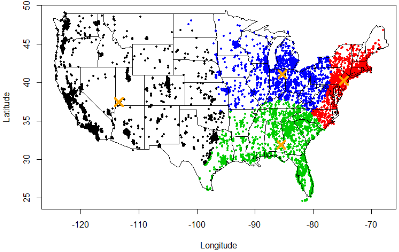
- Semi-supervised learning: some of the $\mathbf{Y}_i$s are observed (labeling is expensive or difficult).



- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).





- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Unsupervised learning: none of the $\mathbf{Y}_i$s are observed (detect patterns).

- Reinforcement learning: feedback (possibly adversarial) from the environment (robotics, adversarial environments, training...).

- Reinforcement learning: feedback (possibly adversarial) from the environment (robotics, adversarial environments, training...).

Loss function:

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

Loss function:

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

Loss function:

$$\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi}\colon (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Loss function:

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi} \colon (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Risk

$$R(\hat{\phi}) = \mathbb{E}\left[\ell\left(\hat{\phi}(\mathbf{X}), Y\right)\right],$$

Loss function:

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+.$$

$\ell(\hat{\phi}(\mathbf{X}), Y)$ (random) quantifies how a predictor $\hat{\phi}(\mathbf{X})$ is a "good" approximation of $\mathbf{Y}$.

A predictor is any mapping

$$\hat{\phi} \colon (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{Y}^{\mathcal{X}}.$$

Risk

$$R(\hat{\phi}) = \mathbb{E}\left[\ell\left(\hat{\phi}(\mathbf{X}), Y\right)\right],$$

Empirical risk

$$R_n(\hat{\phi}) = \frac{1}{n}\sum_{i=1}^{n}\left[\ell\left(\hat{\phi}(\mathbf{X}_i), Y_i\right)\right].$$

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.

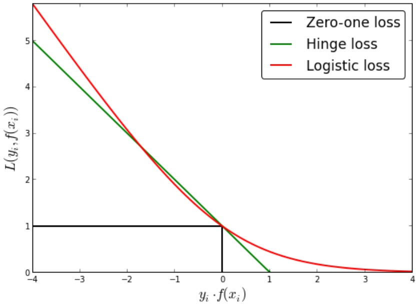
- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.

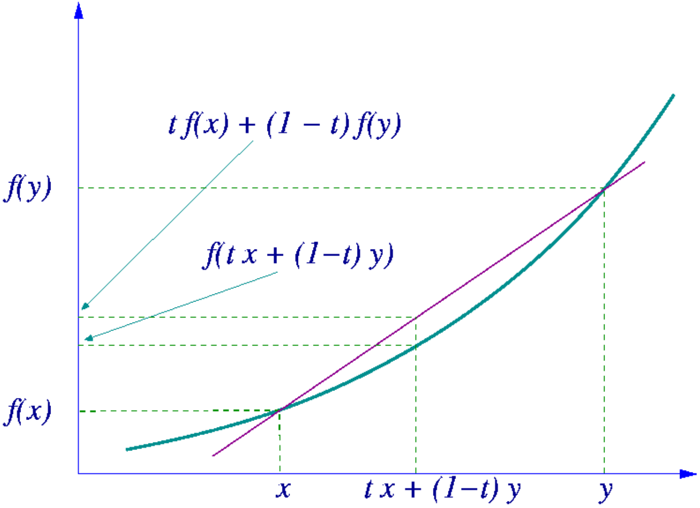
- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.
- Logistic loss (classification): $\ell(a, b) = \log[1 + \exp(-ab)]$.

- Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- Absolute loss (regression): $\ell(a, b) = |a - b|$.
- 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.
- Logistic loss (classification): $\ell(a, b) = \log[1 + \exp(-ab)]$.



Convexity is (often) crucial



Statistical Learning vs. Machine Learning

Same task, different approaches:

Statistical Learning vs. Machine Learning

Same task, different approaches:

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{m} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i)) \right\}.$$

Statistical Learning vs. Machine Learning

Same task, different approaches:

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{m} \left\{ \sum_{i=1}^{n} \ell(y_i, m(\mathbf{x}_i)) \right\}.$$

- In statistical learning, assume that the Y_is are realisations of some random variable Y (given $\mathbf{X}$) with distribution P. Solve

$$\widehat{\phi}(\cdot) = \arg\max_{m} \left\{ \sum_{i=1}^{n} \log \mathrm{d}P(Y_i; m(\mathbf{X}_i)) \right\}.$$

SL vs. ML in the simple parametric case

SL vs. ML in the simple parametric case

- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

$$\widehat{\phi}(\cdot) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \ell(y_i, \langle \theta, \mathbf{x}_i \rangle) \right\}.$$

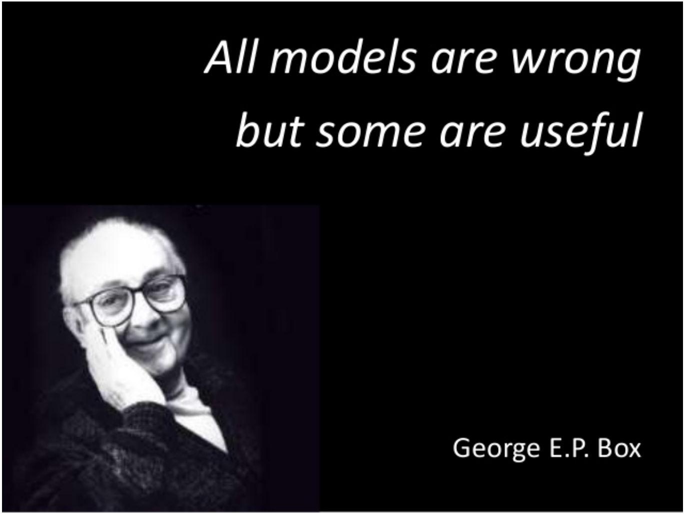
SL vs. ML in the simple parametric case

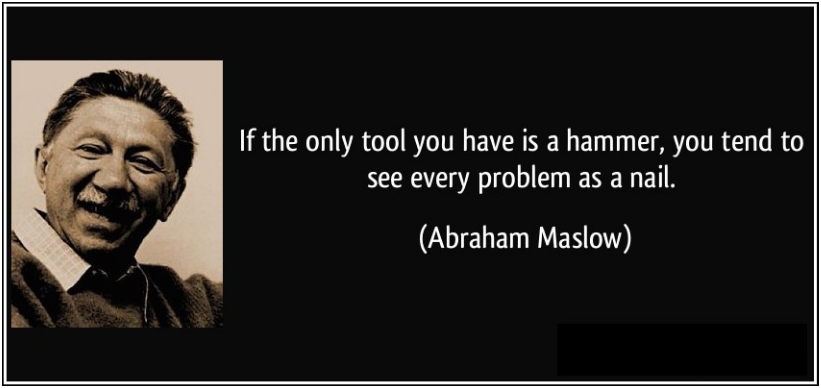
- In machine learning, given some deterministic sequence $(\mathbf{x}_i, y_i)$, solve

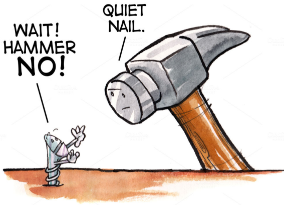
$$\widehat{\phi}(\cdot) = \arg\min_{\theta} \left\{ \sum_{i=1}^{n} \ell(y_i, \langle \theta, \mathbf{x}_i \rangle) \right\}.$$

- In statistical learning, assume that the Y_is are realisations of some random variable Y (given $\mathbf{X}$) with distribution P. Solve

$$\widehat{\phi}(\cdot) = \arg\max_{\theta} \left\{ \sum_{i=1}^{n} \log \mathrm{d}P(Y_i | \mathbf{X}_i, \theta) \right\}.$$


All models are wrong
but some are useful
George E.P. Box


If the only tool you have is a hammer, you tend to see every problem as a nail.
(Abraham Maslow)


WAIT! HAMMER NO!
QUIET NAIL.

A primer on probability distributions

All words point to hyperlinks.

A primer on probability distributions

All words point to hyperlinks.

- Normal

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson
- Binomial

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson
- Binomial
- Bernoulli

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson
- Binomial
- Bernoulli
- Gamma

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson
- Binomial
- Bernoulli
- Gamma
- Student

A primer on probability distributions

All words point to hyperlinks.

- Normal
- Inverse Gaussian (a.k.a Wald)
- Beta
- Poisson
- Binomial
- Bernoulli
- Gamma
- Student
- ...

The Bayesian paradigm

Introductory example

Consider observations $\mathbf{x} = (x_1, \ldots, x_n)$ generated from a probability distribution with density $f(\cdot|\theta)$.

Introductory example

Consider observations $\mathbf{x} = (x_1, \ldots, x_n)$ generated from a probability distribution with density $f(\cdot|\theta)$.

The associated likelihood is the inverted density:

$$\mathcal{L}(\theta|x) = f(x|\theta).$$

Example $f(\cdot|\theta) = \mathcal{N}(\theta, 1)$.

Bayes' Theorem

Bayes' Theorem

Inversion of probabilities a.k.a actualisation principle.

Bayes' Theorem

Inversion of probabilities a.k.a actualisation principle.

If A and B are events such that $\mathbb{P}(B) \neq 0$,

$$\begin{aligned}\mathbb{P}(A|B) &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)} \\ &= \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.\end{aligned}$$

(due to Thomas Bayes, published in 1764)

Who was Thomas Bayes?

Who was Thomas Bayes?

Reverend Thomas Bayes (ca. 1702–1761) Presbyterian minister in Kent from 1731. Election to the Royal Society based on a tract of 1736 where he defended the views and philosophy of Newton. Sole probability paper, "Essay Towards Solving a Problem in the Doctrine of Chances", published posthumously in 1763 and containing the seeds of Bayes' Theorem.

A new paradigm

Bayes introduces a whole new perspective.

A new paradigm

Bayes introduces a whole new perspective.

- Uncertainty on the parameter θ, modeled through a probability distribution π, called *prior distribution*.

A new paradigm

Bayes introduces a whole new perspective.

- Uncertainty on the parameter θ, modeled through a probability distribution π, called *prior distribution*.
- Inference based on the distribution of θ conditional on $\mathbf{X}$ $\pi(\theta|\mathbf{x})$, called *posterior distribution*

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta}.$$

A Bayesian model

…is made of a parametric (in this course) statistical model defined through its likelihood $f(\mathbf{x}|\theta)$ and a prior distribution on the parameter $\pi(\theta)$.

Consequences

- Semantic drift from unknown to random

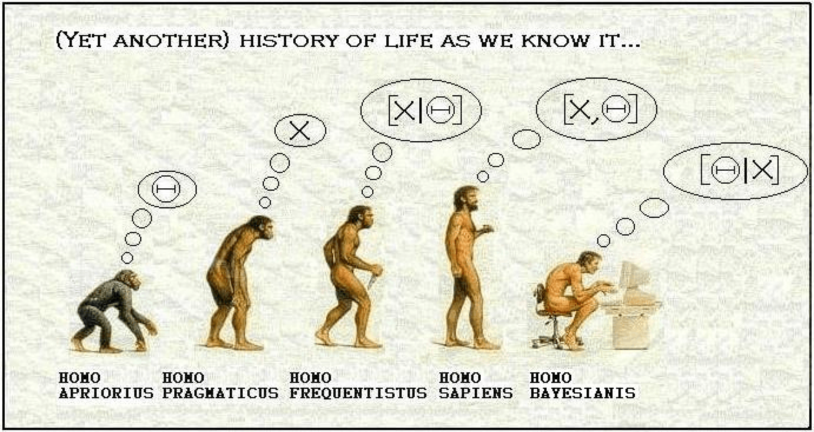
Consequences

- Semantic drift from unknown to random
- Actualization of θ by extracting the information contained in the observation $\mathbf{x}$

Consequences

- Semantic drift from unknown to random
- Actualization of θ by extracting the information contained in the observation $\mathbf{x}$
- Allows incorporation of imperfect information in the decision process

The advantages of being a Bayesian



Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

- The *joint distribution* of $(\theta, \mathbf{x})$

$$\varphi(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta).$$

Distributions (1/2)

Given the likelihood $f(\mathbf{x}|\theta)$ and the prior $\pi(\theta)$, several distributions of interest:

- The *joint distribution* of $(\theta, \mathbf{x})$

$$\varphi(\theta, \mathbf{x}) = f(\mathbf{x}|\theta)\pi(\theta).$$

- The *marginal distribution* of $\mathbf{x}$

$$m(\mathbf{x}) = \int \varphi(\theta, \mathbf{x})\mathrm{d}\theta = \int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta.$$

Distributions (2/2)

Distributions (2/2)

- The *posterior distribution* of θ

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

Distributions (2/2)

- The *posterior distribution* of θ

$$\pi(\theta|\mathbf{x}) = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{\int f(\mathbf{x}|\theta)\pi(\theta)\mathrm{d}\theta} = \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})}.$$

- The *predictive distribution* of y when $y \sim g(\cdot|\theta, \mathbf{x})$

$$g(y|\mathbf{x}) = \int g(y|\theta, \mathbf{x})\pi(\theta|\mathbf{x})\mathrm{d}\theta.$$

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

$$
\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta) \\
&\propto \exp\left(-\frac{(\mathbf{x}-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11}{20}\theta^2 + \theta(\mathbf{x} + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left(\theta - \frac{10\mathbf{x}+a}{11}\right)^2\right)
\end{aligned}
$$

A comprehensive example normal-normal

Assume that we model $\mathbf{x} \sim \mathcal{N}(\theta, 1)$ and use the prior $\theta \sim \mathcal{N}(a, 10)$.

$$
\begin{aligned}
\pi(\theta|\mathbf{x}) &\propto f(\mathbf{x}|\theta)\pi(\theta) \\
&\propto \exp\left(-\frac{(\mathbf{x}-\theta)^2}{2} - \frac{(\theta-a)^2}{20}\right) \\
&\propto \exp\left(-\frac{11}{20}\theta^2 + \theta(\mathbf{x} + a/10)\right) \\
&\propto \exp\left(-\frac{11}{20}\left(\theta - \frac{10\mathbf{x}+a}{11}\right)^2\right)
\end{aligned}
$$

which means $\theta|\mathbf{x} \sim \mathcal{N}\left(\frac{10x+a}{11}, \frac{10}{11}\right)$.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A second ball O then rolls n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

A comprehensive example uniform-binomial

Bayes' very own example: a billiard ball W rolls on a line of length one, with a uniform probability of stopping anywhere: W stops at p.

A second ball O then rolls n times under the same assumptions. X denotes the number of times the ball O stopped on the left of W.

Bayes' question: given X, what inference can we make on p?

Mathematical translation

Derive the posterior distribution of p given X, when $p \sim \mathcal{U}(0,1)$ and $X \sim \mathcal{B}(n,p)$.

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\mathbb{P}(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

Resolution 1/2

Since

$$\mathbb{P}(X = x|p) = \binom{n}{x} p^x (1-p)^{n-x},$$

$$\mathbb{P}(a < p < b \text{ and } X = x) = \int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

and

$$\mathbb{P}(X = x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p,$$

Resolution 2/2

then

$$\mathbb{P}(a < p < b | X = x) = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p}{\int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p} = \frac{\int_a^b \binom{n}{x} p^x (1-p)^{n-x} \mathrm{d}p}{\mathcal{B}(x+1, n-x+1)},$$

i.e., $p|x \sim \mathcal{B}(x+1, n-x+1)$.
(Beta distribution)