



Bayesian Learning

Benjamin Guedj, Ph.D.

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

2017–2018

[benjamin.guedj@inria.fr - Link]

[<https://bguedj.github.io> - Link]

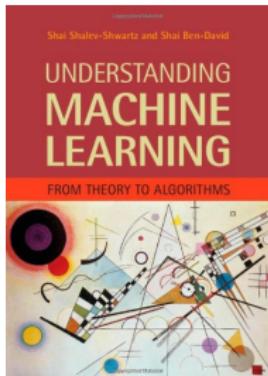
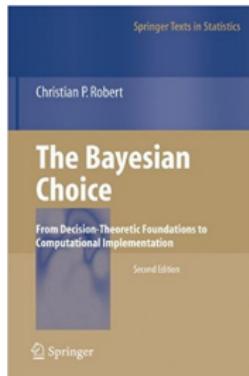
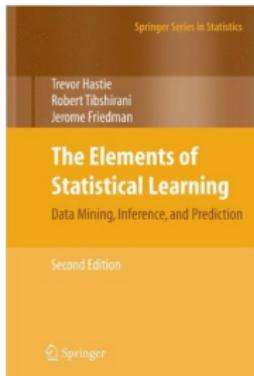
10h de cours (5, 6, 12, 19 et 21 mars 2018)

Projet à rendre. Deadline : **lundi 2 avril 2018 à 23h59.**

[<https://bguedj.github.io/teaching/projet.pdf> - Link]

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]

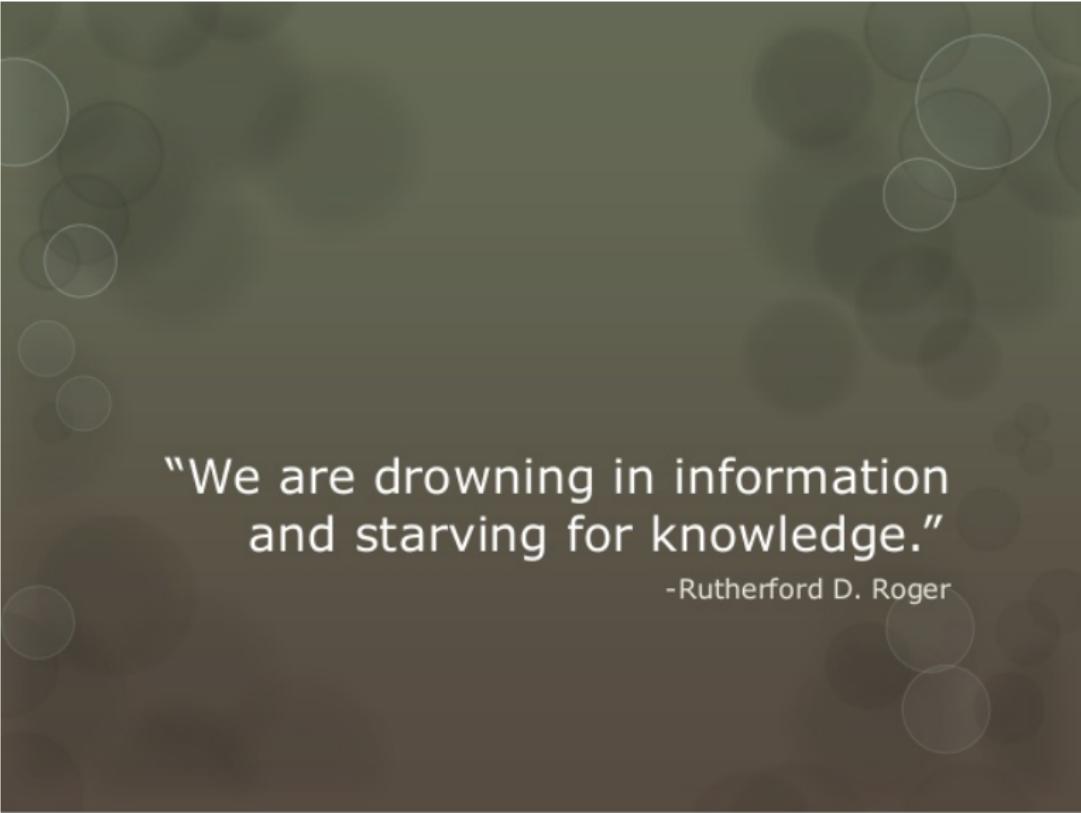


Outline

1. Introduction to statistical / machine learning
2. The Bayesian framework
3. Quasi-Bayesian learning
4. Bayesian learning in practice

The rising of AI

Introduction



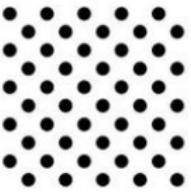
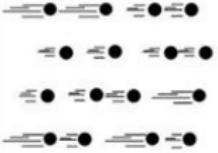
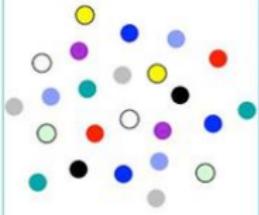
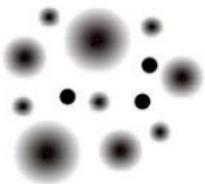
“We are drowning in information
and starving for knowledge.”

-Rutherford D. Roger

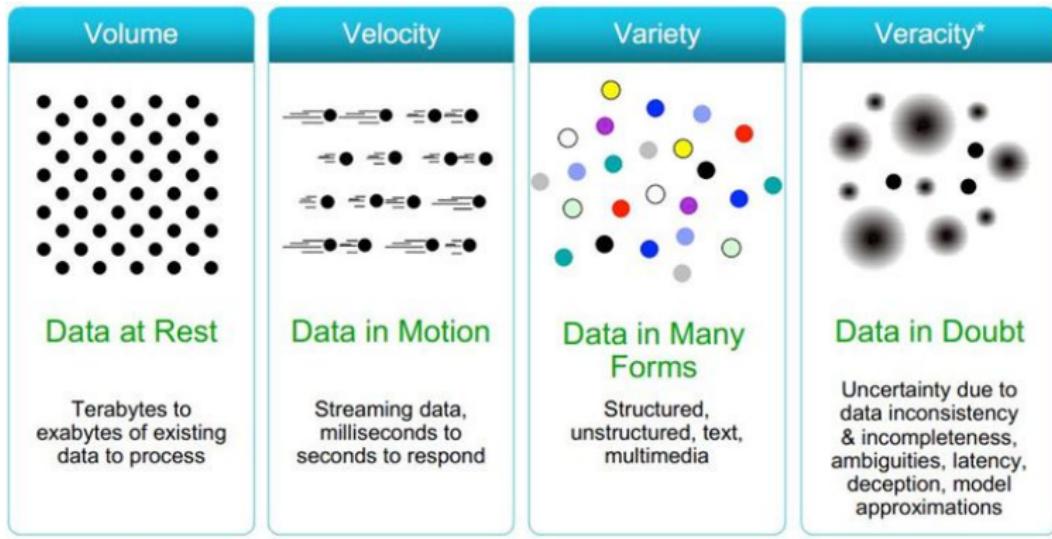
It is vital to remember
that information - in the
sense of raw data - is not
knowledge, that
knowledge is not wisdom,
and that wisdom is not
foresight. But information
is the first essential step
to all of these.

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



→ 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

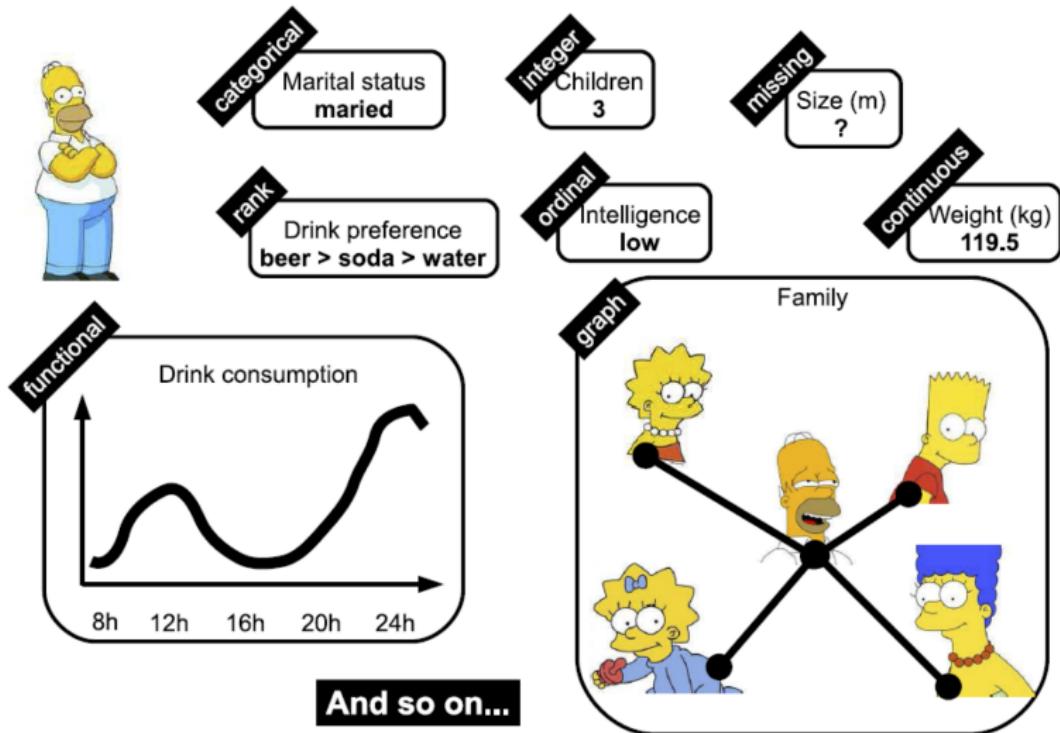
- ▶ Google: 24 petabytes/day
- ▶ Facebook: 10 terabytes/day
- ▶ Twitter: 7 terabytes/day
- ▶ Large Hadron Collider: 40 terabytes/second
- ▶ Google Street View: 20 petabytes
- ▶ AT&T network: 30 petabytes/day
- ▶ Human brain may store about 2.5 petabytes of binary data

Volume / Velocity

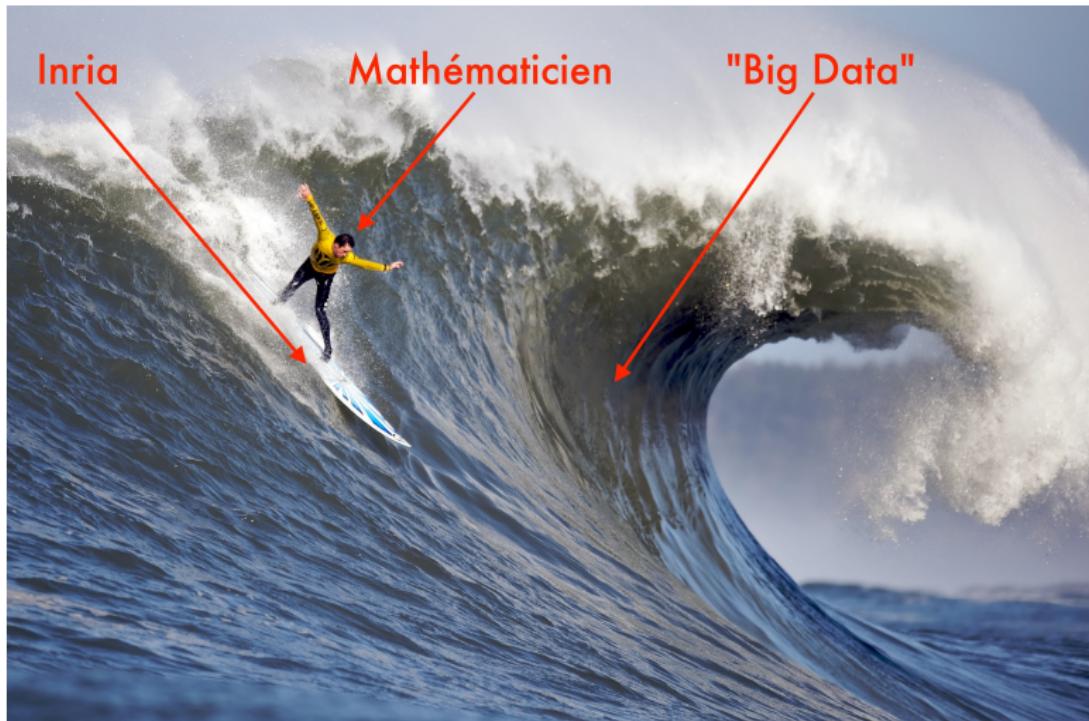
i	3	6	9	12	15	18	21	24	
10^i	kilo	mega	giga	tera	peta	exa	zeta	yotta	bytes

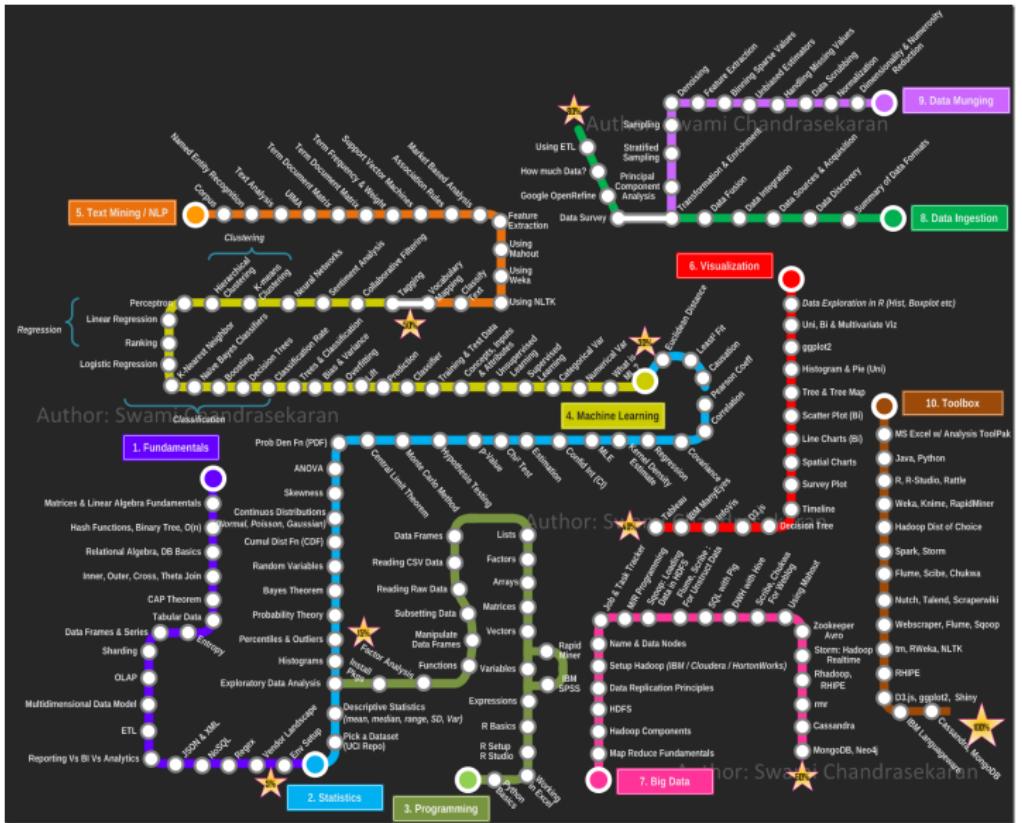
- ▶ Google: 24 petabytes/day
- ▶ Facebook: 10 terabytes/day
- ▶ Twitter: 7 terabytes/day
- ▶ Large Hadron Collider: 40 terabytes/second
- ▶ Google Street View: 20 petabytes
- ▶ AT&T network: 30 petabytes/day
- ▶ Human brain may store about 2.5 petabytes of binary data
- ▶ ...

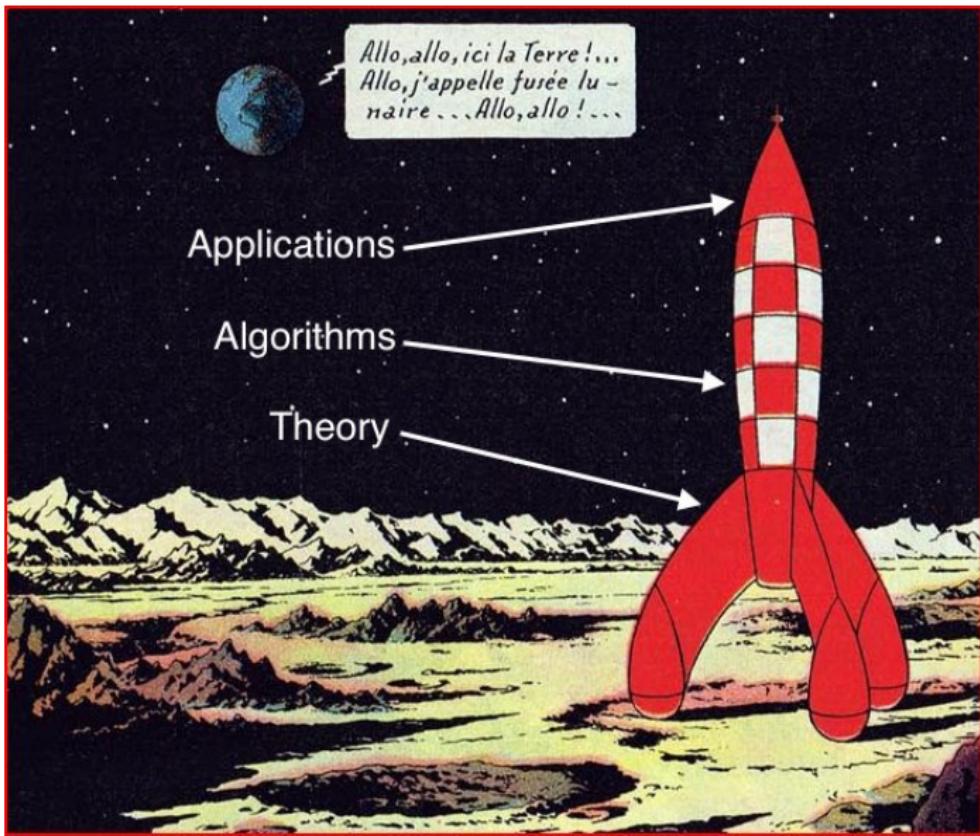
Variety / Veracity



My job (allegory)







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 human 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 human 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
- ▶ Fat data: $d \gg n$
- ▶ Tall data: $n \gg d$

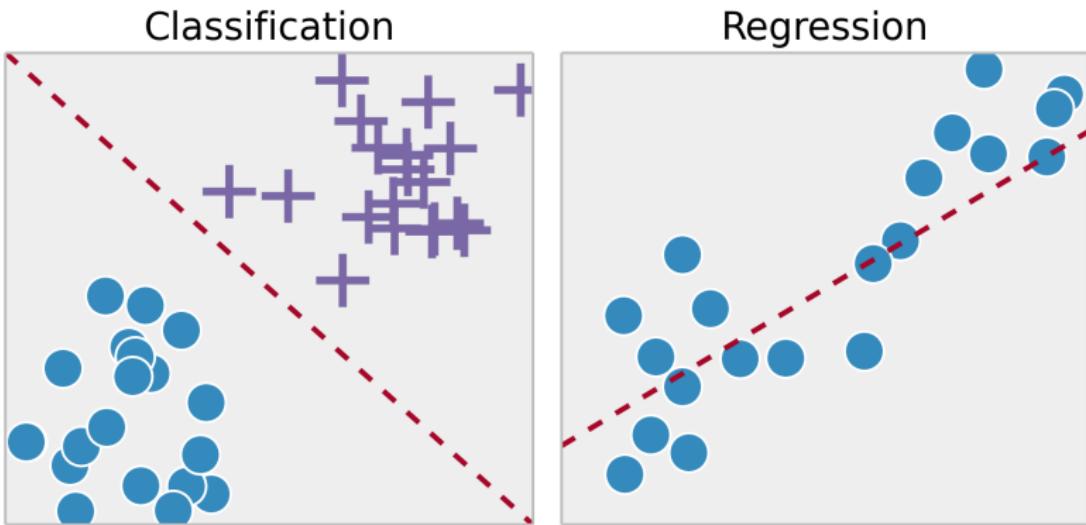
- ▶ Online learning: observations are revealed over time.
- ▶ Batch learning: all observations are revealed at once.

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: $d \gg n$
- ▶ Big/massive data: n and d huge

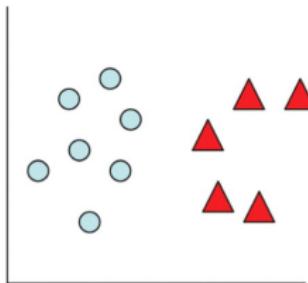
- ▶ Supervised learning: all of the \mathbf{Y}_i s are observed.

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

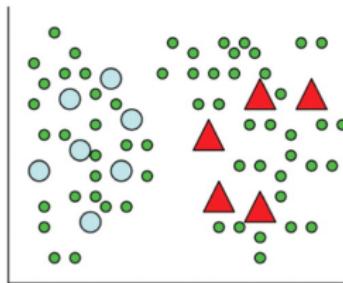


- ▶ Semi-supervised learning: some of the \mathbf{Y}_i s are observed (labeling is expensive or difficult).

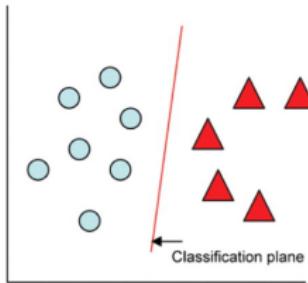
- ▶ Semi-supervised learning: some of the \mathbf{Y}_i s are observed (labeling is expensive or difficult).



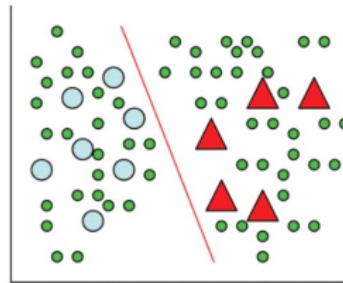
Labeled Data
(a)



Labeled and Unlabeled Data
(b)



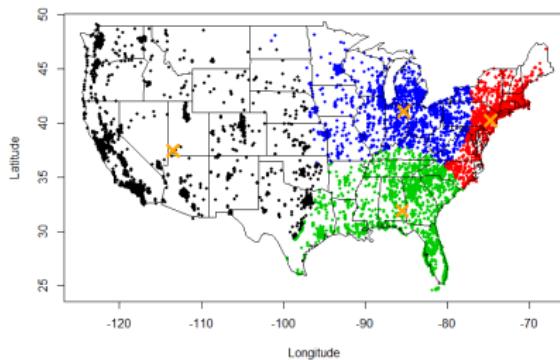
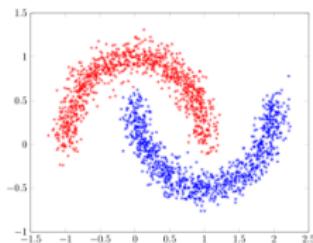
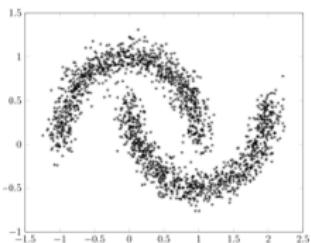
Supervised Learning
(c)



Semi-Supervised Learning
(d)

- ▶ 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:

$$\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

Loss function:

$$\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

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

Loss function:

$$\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

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

A predictor is any mapping

$$\hat{\phi}: (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{Y}^{\mathcal{X}}.$$

Loss function:

$$\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

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

A predictor is any mapping

$$\hat{\phi}: (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{Y}^{\mathcal{X}}.$$

Risk

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

Loss function:

$$\ell: \mathcal{Y} \times \mathcal{Y} \rightarrow \mathbb{R}_+.$$

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

A predictor is any mapping

$$\hat{\phi}: (\mathcal{X} \times \mathcal{Y})^n \rightarrow \mathcal{Y}^{\mathcal{X}}.$$

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

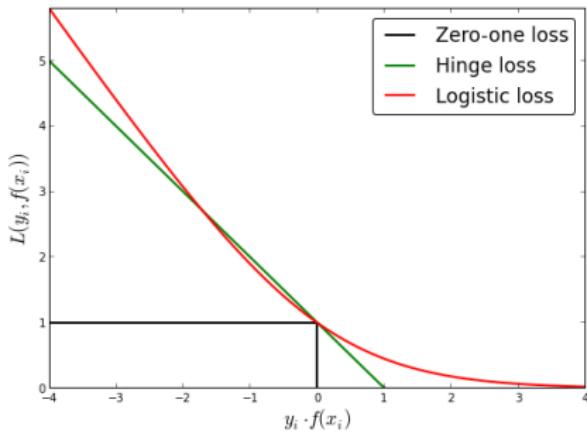
- ▶ Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- ▶ Absolute loss (regression): $\ell(a, b) = |a - b|$.

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

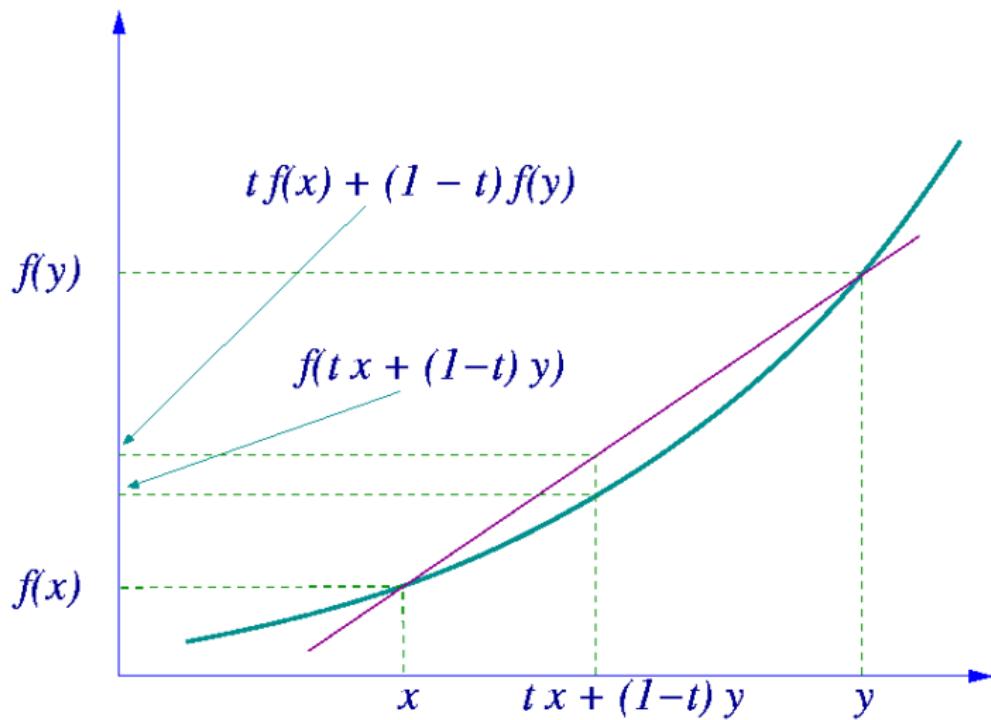
- ▶ Quadratic loss (regression): $\ell(a, b) = (a - b)^2$.
- ▶ Absolute loss (regression): $\ell(a, b) = |a - b|$.
- ▶ 0-1 loss (classification): $\ell(a, b) = \mathbb{1}_{\{a \neq b\}}$.
- ▶ Hinge loss (classification): $\ell(a, b) = \max(0, 1 - ab)$.

- ▶ 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

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

Statistical Learning vs. Machine Learning

Same task, different approaches:

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

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

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

$$\hat{\phi}(\cdot) = \arg \max_m \left\{ \sum_{i=1}^n \log dP(Y_i; m(\mathbf{X}_i)) \right\}.$$

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

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

SL vs. ML in the simple parametric case

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

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

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

$$\hat{\phi}(\cdot) = \arg \max_{\theta} \left\{ \sum_{i=1}^n \log dP(Y_i | \mathbf{x}_i, \theta) \right\}.$$

*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)



A primer on probability distributions

All words are hyperlinks.

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]
- ▶ [Beta - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]
- ▶ [Beta - Link]
- ▶ [Poisson - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]
- ▶ [Beta - Link]
- ▶ [Poisson - Link]
- ▶ [Binomial - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]
- ▶ [Beta - Link]
- ▶ [Poisson - Link]
- ▶ [Binomial - Link]
- ▶ [Bernoulli - Link]

A primer on probability distributions

All words are hyperlinks.

- ▶ [Normal - Link]
- ▶ [Inverse Gaussian (a.k.a Wald) - Link]
- ▶ [Beta - Link]
- ▶ [Poisson - Link]
- ▶ [Binomial - Link]
- ▶ [Bernoulli - Link]
- ▶ [Gamma - Link]

A primer on probability distributions

All words are hyperlinks.

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

A primer on probability distributions

All words are hyperlinks.

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

The Bayesian paradigm

Introductory example

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

Introductory example

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

The associated likelihood is the inverted density:

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

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*

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

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

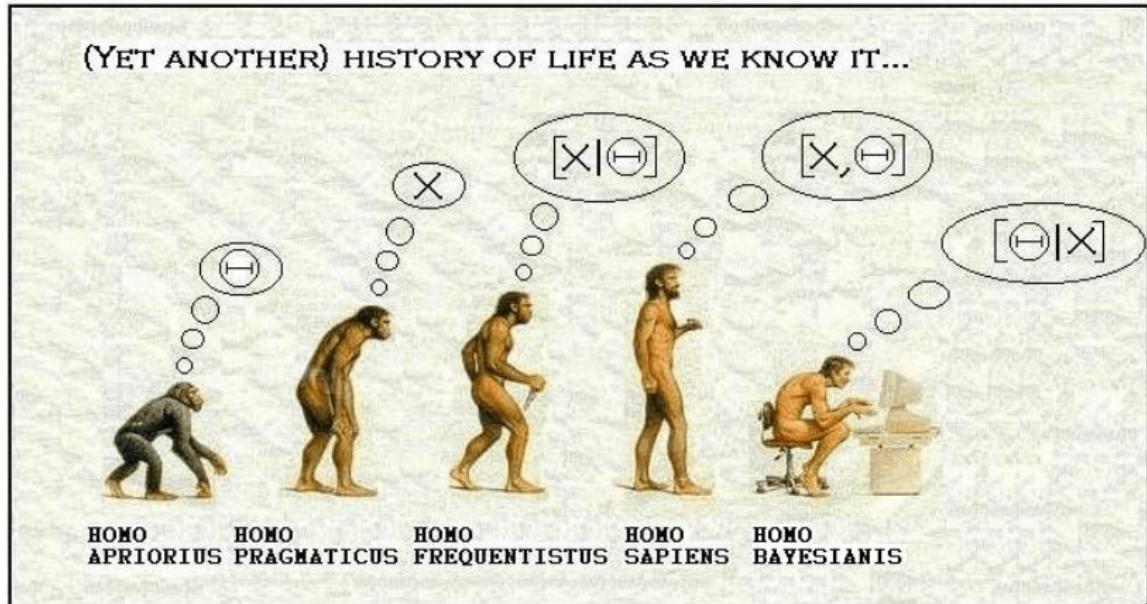
Consequences

- ▶ Semantic drift from unknown to random
- ▶ Actualization of θ by extracting the information contained in the observation x

Consequences

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

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

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

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

Distributions (2/2)

Distributions (2/2)

- ▶ The *posterior distribution* of θ

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

Distributions (2/2)

- ▶ The *posterior distribution* of θ

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

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

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

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

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

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

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

Resolution 1/2

Since

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

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

and

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

Resolution 2/2

then

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

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

Pour se remettre dans le bain

Pour se remettre dans le bain

1. Quelle est la différence entre *statistical learning* et *machine learning* ?
2. Donner la définition d'un algorithme d'apprentissage.
3. Quels sont les quatre grands types d'apprentissage ?
4. Que définissent *fat data* et *tall data* ?
5. Comment compare-t-on les performances d'algorithmes d'apprentissage ?
6. Quelle notion est souvent cruciale au moment de choisir une bonne fonction de perte ? En donner la définition.
7. Donner quelques exemples de fonctions de perte.
8. Enoncer le théorème de Bayes.
9. Quel est le rôle de la distribution *a priori* ?
10. Donner la définition d'un modèle bayésien.
11. Quelles sont les quatre distributions importantes en bayésien ?

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- ▶ Operates conditional upon the observations.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- ▶ Operates conditional upon the observations.
- ▶ Integrates simultaneously prior knowledge and information brought by data.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- ▶ Operates conditional upon the observations.
- ▶ Integrates simultaneously prior knowledge and information brought by data.
- ▶ Coherent and complete inferential scope and unique motor of inference.

The posterior distribution

$$\pi(\theta|\mathbf{x}) \propto f(\mathbf{x}|\theta)\pi(\theta)$$

- ▶ Operates conditional upon the observations.
- ▶ Integrates simultaneously prior knowledge and information brought by data.
- ▶ Coherent and complete inferential scope and unique motor of inference.
- ▶ Usually known up to a constant! $m(\mathbf{x})$ may be intractable.

Prior distributions

There is no such thing as *the* prior distribution!

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ .

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ .

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ .

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta) d\theta = +\infty$.

Prior distributions

There is no such thing as *the* prior distribution!

Usually encapsulates *prior* knowledge on θ .

Vague priors (such as $\theta \sim \mathcal{N}(0, 100)$).

Improper priors: $\int \pi(\theta) d\theta = +\infty$.

A prior on θ may depend on additional parameters: those are called hyperparameters.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family \mathcal{F} of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to \mathcal{F} .

Conjugacy and explicit calculus of posteriors

Conjugate priors are a specific family of distributions with nice analytical properties.

A family \mathcal{F} of probability distributions is *conjugate* for a likelihood $f(x|\theta)$ if for every $\pi \in \mathcal{F}$, the posterior distribution $\pi(\theta|x)$ also belongs to \mathcal{F} .

Only of interest when \mathcal{F} is parameterized: switching from the prior to the posterior is reduced to an update of parameters.

Advantages

- ▶ Limited/finite information conveyed by data x

Advantages

- ▶ Limited/finite information conveyed by data x
- ▶ Preservation of the structure of the prior $\pi(\theta)$

Advantages

- ▶ Limited/finite information conveyed by data x
- ▶ Preservation of the structure of the prior $\pi(\theta)$
- ▶ Exchangeability

Advantages

- ▶ Limited/finite information conveyed by data x
- ▶ Preservation of the structure of the prior $\pi(\theta)$
- ▶ Exchangeability
- ▶ Allows for generation of "virtual observations"

Advantages

- ▶ Limited/finite information conveyed by data x
- ▶ Preservation of the structure of the prior $\pi(\theta)$
- ▶ Exchangeability
- ▶ Allows for generation of "virtual observations"
- ▶ Most importantly: **tractability and simplicity**

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Exponential families

The family of distributions

$$f(x|\theta) = C(\theta)h(x)\exp(R(\theta)T(x))$$

is called an *exponential family*. When

$$f(x|\theta) = h(x)\exp(-\theta x - \psi(\theta))$$

the family is said to be *natural*.

Main interest: allow for conjugate priors

$$\pi(\theta|\mu, \lambda) = K(\mu, \lambda)\exp(\theta\mu - \lambda\psi(\theta)), \quad \lambda > 0.$$

Classical exponential families and conjugate priors

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$$f(x|\theta)$$

$$\pi(\theta)$$

$$\pi(\theta|x)$$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + x, \beta + n - x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + x, \beta + n - x)$
Negative binomial $\text{Neg}(m, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + m, \beta + x)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + x, \beta + n - x)$
Negative binomial $\text{Neg}(m, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \dots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$

Classical exponential families and conjugate priors

[Examples of exponential families - Link]

$f(x \theta)$	$\pi(\theta)$	$\pi(\theta x)$
Normal $\mathcal{N}(\theta, \sigma^2)$	Normal $\mathcal{N}(\mu, \tau^2)$	Normal $\mathcal{N}(\rho(\sigma^2\mu + \tau^2x), \rho\sigma^2\tau^2)$ $\rho^{-1} = \sigma^2 + \tau^2$
Poisson $\mathcal{P}(\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + x, \beta + 1)$
Gamma $\mathcal{G}(\nu, \theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + \nu, \beta + x)$
Binomial $\mathcal{B}(n, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + x, \beta + n - x)$
Negative binomial $\text{Neg}(m, \theta)$	Beta $\mathcal{Be}(\alpha, \beta)$	Beta $\mathcal{Be}(\alpha + m, \beta + x)$
Multinomial $\mathcal{M}_k(\theta_1, \dots, \theta_k)$	Dirichlet $\mathcal{D}(\alpha_1, \dots, \alpha_k)$	Dirichlet $\mathcal{D}(\alpha_1 + x_1, \dots, \alpha_k + x_k)$
Normal $\mathcal{N}(\mu, 1/\theta)$	Gamma $\mathcal{G}(\alpha, \beta)$	Gamma $\mathcal{G}(\alpha + 1/2, \beta + (\mu - x)^2/2)$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^\star(\theta) \propto \det(I(\theta))^{1/2}.$$

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^*(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^*(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- ▶ Relates to information theory

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^*(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- ▶ Relates to information theory
- ▶ Parameterization invariant

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^*(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- ▶ Relates to information theory
- ▶ Parameterization invariant
- ▶ Suffers from dimensionality curse

Non-informative priors: Jeffreys priors

Based on Fisher information:

$$I(\theta) = -\mathbb{E} \left[\frac{\partial^2 \log \mathcal{L}(\theta|x)}{\partial \theta^2} \right].$$

Jeffreys prior is defined as

$$\pi^*(\theta) \propto \det(I(\theta))^{1/2}.$$

Pros & Cons:

- ▶ Relates to information theory
- ▶ Parameterization invariant
- ▶ Suffers from dimensionality curse
- ▶ Depends on data: incoherence with the likelihood principle

Example

If $x \sim \mathcal{B}(n, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \mathcal{Be}(1/2, 1/2).$$

If $n \sim \text{Neg}(x, \theta)$, Jeffreys' prior is

$$\pi(\theta) \propto \theta^{-1}(1 - \theta)^{-1/2}$$

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \dots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Non-informative priors: Laplace priors

With a finite set $\{\theta_1, \dots, \theta_p\}$, uniform prior $\pi(\theta_i) = 1/p$.

Continuous extension: $\pi(\theta) \propto 1$. This is no longer a probability distribution yet if $\int f(x|\theta)d\theta < +\infty$, the posterior is well-defined as a probability distribution. Modeling is crucial. Weakness: lack of reparameterization invariance.

Pour se remettre dans le bain

- ▶ Expliquer l'intérêt de la conjugaison, et en donner la définition.
- ▶ Quel est l'intérêt d'utiliser une vraisemblance issue d'une famille exponentielle naturelle ?
- ▶ Donner des exemples de lois conjuguées.
- ▶ Donner deux exemples de méthodes de construction de priors non-informatifs. Quelles sont les limites de ces méthodes ?

Bayesian estimators

Bayesian paradigm is based on the posterior distribution.

Bayesian estimators

Bayesian paradigm is based on the posterior distribution.

Many estimators may be derived from the posterior.

MAP

The maximum a posteriori estimator is defined as

$$\arg \max_{\theta} f(x|\theta)\pi(\theta)$$

(penalized likelihood estimator).

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$$\pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1 - \theta)^{-1/2} \quad |$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$$\pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1 - \theta)^{-1/2} \quad \Big| \quad \hat{\theta} = \max \left(\frac{x-1/2}{n-1}, 0 \right)$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$$\frac{\pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1-\theta)^{-1/2}}{\pi(\theta) = 1} \quad \left| \quad \hat{\theta} = \max \left(\frac{x-1/2}{n-1}, 0 \right) \right.$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$$\begin{array}{c|c} \pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1-\theta)^{-1/2} & \hat{\theta} = \max \left(\frac{x-1/2}{n-1}, 0 \right) \\ \hline \pi(\theta) = 1 & \hat{\theta} = x/n \end{array}$$

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1 - \theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	$\hat{\theta} = x/n$
$\pi(\theta) = \theta^{-1} (1 - \theta)^{-1}$	

Binomial example

Consider $x|\theta \sim \mathcal{B}(n, \theta)$.

Possible priors and corresponding MAP estimators:

$\pi(\theta) = \frac{1}{\mathcal{B}(1/2, 1/2)} \theta^{-1/2} (1 - \theta)^{-1/2}$	$\hat{\theta} = \max\left(\frac{x-1/2}{n-1}, 0\right)$
$\pi(\theta) = 1$	$\hat{\theta} = x/n$
$\pi(\theta) = \theta^{-1} (1 - \theta)^{-1}$	$\hat{\theta} = \max\left(\frac{x-1}{n-2}, 0\right)$

Not always appropriate!

Consider $f(x|\theta) = \frac{1}{\pi} (1 + (x - \theta)^2)^{-1}$ and $\pi(\theta) = \frac{1}{2} \exp(-|\theta|)$.

Not always appropriate!

Consider $f(x|\theta) = \frac{1}{\pi} (1 + (x - \theta)^2)^{-1}$ and $\pi(\theta) = \frac{1}{2} \exp(-|\theta|)$.

The MAP is $\hat{\theta} = 0$!

Other possible estimators

- ▶ Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) d\theta.$

Other possible estimators

- ▶ Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) d\theta.$
- ▶ Median: $\hat{\theta} = \text{med}(\pi(\cdot|x)).$

Other possible estimators

- ▶ Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) d\theta.$
- ▶ Median: $\hat{\theta} = \text{med}(\pi(\cdot|x)).$
- ▶ Realization: $\hat{\theta} \sim \pi(\cdot|x).$

Other possible estimators

- ▶ Mean: $\hat{\theta} = \mathbb{E}_{\theta \sim \pi(\cdot|x)} \theta = \int \theta \pi(\theta|x) d\theta.$
- ▶ Median: $\hat{\theta} = \text{med}(\pi(\cdot|x)).$
- ▶ Realization: $\hat{\theta} \sim \pi(\cdot|x).$
- ▶ ...

Credible regions

Natural confidence region: highest posterior density (HPD).

$$\mathcal{C}_\alpha^\pi = \{\theta; \pi(\theta|x) > \alpha\}.$$

Prediction

Reminder: if $x \sim f(\cdot|\theta)$ and $z \sim g(\cdot|x, \theta)$, the *predictive distribution* is

$$g^\pi(z|x) = \int g(z|x, \theta)\pi(\theta|x)d\theta.$$

Example: normal prediction

Assume that $(x_1, \dots, x_n) \sim \mathcal{N}(\mu, \sigma^2)^{\otimes n}$ and

$$\pi(\mu, \sigma^2) \propto (\sigma^2)^{-\lambda_\sigma - 3/2} \exp\left(\frac{-\lambda_\mu(\mu - \xi)^2 + \alpha}{2\sigma^2}\right).$$

The posterior is

$$\mathcal{N}\left(\frac{\lambda_\mu \xi + n\bar{x}_n}{\lambda_\mu + n}, \frac{\sigma^2}{\lambda_\mu + n}\right) \times \mathcal{IG}\left(\lambda_\sigma + n/2, \frac{\alpha + s_x^2 + \frac{n\lambda_\mu(\bar{x}_n - \xi)^2}{\lambda_\mu + n}}{2}\right).$$

(where $s_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}_n)^2$)

$$\begin{aligned}
g^\pi(z|x_1, \dots, x_n) &\propto \int (\sigma^2)^{-\lambda_\sigma - 2 - n/2} \exp(-(z - \mu)^2/2\sigma^2) \\
&\quad \times \exp\left(-(\lambda_\mu + n) \left(\mu - \frac{\lambda_\mu \xi + n \bar{x}_n}{\lambda_\mu + n}\right)^2 + \alpha + s_x^2 + \frac{n \lambda_\mu (\bar{x}_n - \xi)^2}{\lambda_\mu + n}\right) / 2\sigma^2 \\
&\quad \times d(\mu, \sigma^2) \\
&\propto \left[\alpha + s_x^2 + \frac{n \lambda_\mu (\bar{x}_n - \xi)^2}{\lambda_\mu + n} + \frac{\lambda_\mu + n + 1}{\lambda_\mu + n} \left(z - \frac{\lambda_\mu \xi + n \bar{x}_n}{\lambda_\mu + n}\right)^2 \right]^{-(2\lambda_\sigma + n + 1)/2}
\end{aligned}$$

$$\begin{aligned}
g^\pi(z|x_1, \dots, x_n) &\propto \int (\sigma^2)^{-\lambda_\sigma - 2 - n/2} \exp(-(z - \mu)^2/2\sigma^2) \\
&\quad \times \exp\left(-(\lambda_\mu + n) \left(\mu - \frac{\lambda_\mu \xi + n \bar{x}_n}{\lambda_\mu + n}\right)^2 + \alpha + s_x^2 + \frac{n \lambda_\mu (\bar{x}_n - \xi)^2}{\lambda_\mu + n}\right) / 2\sigma^2 \\
&\quad \times d(\mu, \sigma^2) \\
&\propto \left[\alpha + s_x^2 + \frac{n \lambda_\mu (\bar{x}_n - \xi)^2}{\lambda_\mu + n} + \frac{\lambda_\mu + n + 1}{\lambda_\mu + n} \left(z - \frac{\lambda_\mu \xi + n \bar{x}_n}{\lambda_\mu + n}\right)^2 \right]^{-(2\lambda_\sigma + n + 1)/2}
\end{aligned}$$

Student t distribution with mean $\frac{\lambda_\mu \xi + n \bar{x}_n}{\lambda_\mu + n}$ and $2\lambda_\sigma + n$ degrees of freedom.

Quasi-Bayesian learning, and a foretaste of PAC-Bayesian theory

The quasi-Bayesian approach

The quasi-Bayesian approach

Set of candidates \mathcal{F} equipped with a probability measure π (prior).

The quasi-Bayesian approach

Set of candidates \mathcal{F} equipped with a probability measure π (prior).

Quasi-posterior

$$\hat{\rho}_\lambda(\cdot) \propto \exp(-\lambda R_n(\cdot)) \pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

The quasi-Bayesian approach

Set of candidates \mathcal{F} equipped with a probability measure π (prior).

Quasi-posterior

$$\hat{\rho}_\lambda(\cdot) \propto \exp(-\lambda R_n(\cdot)) \pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

Key fact!

The quasi-Bayesian approach

Set of candidates \mathcal{F} equipped with a probability measure π (prior).

Quasi-posterior

$$\hat{\rho}_\lambda(\cdot) \propto \exp(-\lambda R_n(\cdot)) \pi(\cdot),$$

for some inverse temperature $\lambda > 0$.

Key fact! In general, $\exp(-\lambda R_n(\cdot))$ is not a likelihood, hence the term quasi-Bayesian.

A generalization of Bayesian learning

The pseudo-likelihood term $\exp(-\lambda R_n(\cdot))$ is to be seen as a data fit term. However no model is attached to this representation!
Quasi-Bayesian learning natively is a model-free learning paradigm.

A generalization of Bayesian learning

The pseudo-likelihood term $\exp(-\lambda R_n(\cdot))$ is to be seen as a data fit term. However no model is attached to this representation!
Quasi-Bayesian learning natively is a model-free learning paradigm.

Tradeoff between interpretability (Bayesian modeling) and performance (quasi-Bayesian prediction). Echoes the celebrated similar tradeoff between ML and SL!

The missing link between machine learning and statistical learning?

Reminder:

- ▶ In ML, deterministic sequence (\mathbf{x}_i, y_i) ,

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

- ▶ In SL, random variables,

$$\hat{\phi}(\cdot) = \arg \max_m \left\{ \sum_{i=1}^n \log dP(Y_i; m(\mathbf{X}_i)) \right\}.$$

The missing link between machine learning and statistical learning?

Reminder:

- ▶ In ML, deterministic sequence (\mathbf{x}_i, y_i) ,

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

- ▶ In SL, random variables,

$$\hat{\phi}(\cdot) = \arg \max_m \left\{ \sum_{i=1}^n \log dP(Y_i; m(\mathbf{X}_i)) \right\}.$$

Quasi-Bayesian learning is a model-free approach yet relies on a stochastic assumption! Joining the best of two worlds.

A variational perspective

A variational perspective

With the classical quadratic loss $\ell: (a, b) \mapsto (a - b)^2$,

$$\hat{\rho}_\lambda \in \arg \inf_{\rho \ll \pi} \left\{ \int_{\mathcal{F}} R_n(\phi) \rho(d\phi) + \frac{\mathcal{K}(\rho, \pi)}{\lambda} \right\},$$

where \mathcal{K} is the Kullback-Leibler divergence defined as

$$\mathcal{K}(\rho, \pi) = \begin{cases} \int_{\mathcal{F}} \log \left(\frac{d\rho}{d\pi} \right) d\rho & \text{when } \rho \ll \pi, \\ +\infty & \text{otherwise.} \end{cases}$$

Typical quasi-Bayesian estimators

Typical quasi-Bayesian estimators

MAQP

$$\hat{\phi}_\lambda \in \arg \max_{\phi \in \mathcal{F}} \hat{\rho}_\lambda(\phi).$$

Typical quasi-Bayesian estimators

MAQP

$$\hat{\phi}_\lambda \in \arg \max_{\phi \in \mathcal{F}} \hat{\rho}_\lambda(\phi).$$

Mean

$$\hat{\phi}_\lambda = \mathbb{E}_{\hat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \hat{\rho}_\lambda(d\phi).$$

Typical quasi-Bayesian estimators

MAQP

$$\hat{\phi}_\lambda \in \arg \max_{\phi \in \mathcal{F}} \hat{\rho}_\lambda(\phi).$$

Mean

$$\hat{\phi}_\lambda = \mathbb{E}_{\hat{\rho}_\lambda} \phi = \int_{\mathcal{F}} \phi \hat{\rho}_\lambda(d\phi).$$

Realization

$$\hat{\phi}_\lambda \sim \hat{\rho}_\lambda.$$

And so on.

Statistical aggregation revisited

Statistical aggregation revisited

Assume that \mathcal{F} is finite.

Statistical aggregation revisited

Assume that \mathcal{F} is finite.

The mean of the quasi-posterior $\widehat{\rho}_\lambda$ amounts to the celebrated exponentially weighted aggregate (EWA)

$$\widehat{\phi}_\lambda = \mathbb{E}_{\widehat{\rho}_\lambda} \phi = \sum_{i=1}^{\#\mathcal{F}} \omega_{\lambda,i} \phi_i$$

where

$$\omega_{\lambda,i} = \frac{\exp(-\lambda R_n(\phi_i)) \pi(\phi_i)}{\sum_{j=1}^{\#\mathcal{F}} \exp(-\lambda R_n(\phi_j)) \pi(\phi_j)}.$$

 Guedj (2013). Agrégation d'estimateurs et de classificateurs : théorie et méthodes, *Ph.D. thesis, Université Pierre & Marie Curie*

Pour se remettre dans le bain

Pour se remettre dans le bain

1. Citer les quatre estimateurs bayésiens les plus couramment utilisés.
2. Qu'est-ce qu'une région de crédibilité ?
3. Comment prédire quand on est bayésien(ne) ?
4. Décrire l'approche quasi-bayésienne.
5. Pourquoi l'approche quasi-bayésienne peut-elle être vue comme une généralisation de l'apprentissage bayésien ?
6. Illustrer la provenance du quasi-posterior au moyen d'une formulation variationnelle.
7. Citer les quatre estimateurs quasi-bayésiens les plus couramment utilisés.
8. Rappeler le principe de l'ERM et de l'agrégation à poids convexe (EWA). Quel est le lien entre EWA et apprentissage quasi-bayésien ?

Assessing the performance: the oracle approach

Oracle:

$$\phi^* \in \arg \min_{\phi \in \mathcal{Y}^X} R(\phi).$$

Ultimate goal: do almost as well as the oracle.

Assessing the performance: the oracle approach

Oracle:

$$\phi^* \in \arg \min_{\phi \in \mathcal{Y}^X} R(\phi).$$

Ultimate goal: do almost as well as the oracle.

Excess risk:

$$\mathcal{E}(\cdot) = R(\cdot) - R^* \geq 0, \quad R^* = R(\phi^*).$$

PAC oracle inequalities

PAC oracle inequalities

Let R^* denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P} \left(R \left(\widehat{\phi}_\lambda \right) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

PAC oracle inequalities

Let R^* denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P} \left(R \left(\hat{\phi}_\lambda \right) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (e.g., Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

PAC oracle inequalities

Let R^* denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P} \left(R \left(\widehat{\phi}_\lambda \right) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (e.g., Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- ▶ $\alpha = \frac{1}{2}$ (slow rate)
- ▶ $\alpha = 1$ (fast rate)

PAC oracle inequalities

Let R^* denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P} \left(R \left(\widehat{\phi}_\lambda \right) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (e.g., Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- ▶ $\alpha = \frac{1}{2}$ (slow rate)
- ▶ $\alpha = 1$ (fast rate)

Let $d = \dim(\mathcal{X})$

- ▶ $\Delta(\phi, \epsilon) \propto d + \log \frac{1}{\epsilon}$
- ▶ $\Delta(\phi, \epsilon) \propto \log d + \log \frac{1}{\epsilon}$

PAC oracle inequalities

Let R^* denote the oracle risk. For any $\epsilon > 0$,

$$\mathbb{P} \left(R \left(\widehat{\phi}_\lambda \right) - R^* \leq \spadesuit \inf_{\phi \in \mathcal{F}} \left\{ R(\phi) - R^* + \frac{\Delta(\phi, \epsilon)}{n^\alpha} \right\} \right) \geq 1 - \epsilon,$$

where $\spadesuit \geq 1$ and $\lambda \propto n$. If $\spadesuit = 1$, the inequality is *exact* or *sharp*.

Key argument: concentration inequalities (e.g., Bernstein, Hoeffding) + Legendre transform of the Kullback-Leibler divergence.

Typical rates in the literature

- ▶ $\alpha = \frac{1}{2}$ (slow rate)
- ▶ $\alpha = 1$ (fast rate)

Let $d = \dim(\mathcal{X})$

- ▶ $\Delta(\phi, \epsilon) \propto d + \log \frac{1}{\epsilon}$
- ▶ $\Delta(\phi, \epsilon) \propto \log d + \log \frac{1}{\epsilon}$

The remainder term grows with d and the size of \mathcal{F} . It decreases with n .

Hoeffding inequality

Let V_1, \dots, V_n be independent real-valued random variables such that $a_i \leq V_i \leq b_i$ a.s. Let $\bar{V}_n = \frac{1}{n} \sum_{i=1}^n V_i$.

Hoeffding inequality

Let V_1, \dots, V_n be independent real-valued random variables such that $a_i \leq V_i \leq b_i$ a.s. Let $\bar{V}_n = \frac{1}{n} \sum_{i=1}^n V_i$. Then

$$\mathbb{P}(\bar{V}_n - \mathbb{E}\bar{V}_n > t) \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^n (b_i - a_i)^2}\right), \forall t > 0.$$

Lemma (Csiszar, 1975 ; Catoni, 2004)

Let (A, \mathcal{A}) be a measurable space. For any probability μ on (A, \mathcal{A}) and any measurable function $h : A \rightarrow \mathbb{R}$ such that

$$\int (\exp \circ h) d\mu < \infty,$$

$$\log \int (\exp \circ h) d\mu = \sup_{m \in \mathcal{M}_\pi(A, \mathcal{A})} \left\{ \int h dm - \mathcal{K}(m, \mu) \right\},$$

with the convention $\infty - \infty = -\infty$. Moreover, as soon as h is upper-bounded on the support of μ , the supremum with respect to m on the right-hand side is reached for the Gibbs distribution g given by

$$\frac{dg}{d\mu}(a) = \frac{\exp \circ h(a)}{\int (\exp \circ h) d\mu}, \quad a \in A.$$

The PAC-Bayesian theory

The PAC-Bayesian theory

...consists in producing PAC inequalities of quasi-Bayesian estimators.

The PAC-Bayesian theory

...consists in producing PAC inequalities of quasi-Bayesian estimators.

- Shawe-Taylor and Williamson (1997). A PAC analysis of a Bayes estimator, *COLT*
- McAllester (1998). Some PAC-Bayesian theorems, *COLT*
- McAllester (1999). PAC-Bayesian model averaging, *COLT*
- Catoni (2004). Statistical Learning Theory and Stochastic Optimization, Springer
- Audibert (2004). Une approche PAC-bayésienne de la théorie statistique de l'apprentissage, *Ph.D. thesis*,
Université Pierre & Marie Curie
- Catoni (2007). PAC-Bayesian Supervised Classification: The Thermodynamics of Statistical Learning, IMS
- Dalalyan and Tsybakov (2008). Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity,
Machine Learning
- Alquier and Lounici (2011). PAC-Bayesian theorems for sparse regression estimation with exponential weights,
Electronic Journal of Statistics

A flexible and powerful framework

A flexible and powerful framework

Numerous models addressed by the PAC-Bayesian literature.

- Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*
- Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

A flexible and powerful framework

Numerous models addressed by the PAC-Bayesian literature.

- Alquier and Wintenberger (2012). Model selection for weakly dependent time series forecasting, *Bernoulli*
- Seldin, Laviolette, Cesa-Bianchi, Shawe-Taylor and Auer (2012). PAC-Bayesian inequalities for martingales, *IEEE Transactions on Information Theory*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

Towards (almost) no assumptions to derive powerful results

- Bégin, Germain, Laviolette and Roy (2016). PAC-Bayesian bounds based on the Rényi divergence, *AISTATS*
- Alquier and Guedj (2017). Simpler PAC-Bayesian bounds for hostile data, *Machine Learning*

Sampling

Monte Carlo integration

Objective: approximation of an integral

$$\mathcal{J} = \int h(x)f(x)dx.$$

Monte Carlo integration

Objective: approximation of an integral

$$\mathcal{J} = \int h(x)f(x)dx.$$

Key idea: exploit the fact that $\mathcal{J} = \mathbb{E}_{X \sim f}[h(X)]$.

Monte Carlo principle

Monte Carlo principle

Sample a sequence $x_1, \dots, x_m \sim f$.

Monte Carlo principle

Sample a sequence $x_1, \dots, x_m \sim f$.

Then use

$$\hat{\mathcal{J}}_m = \frac{1}{m} \sum_{i=1}^m h(x_i)$$

as an estimator of \mathcal{J} .

Monte Carlo principle

Sample a sequence $x_1, \dots, x_m \sim f$.

Then use

$$\hat{\mathcal{J}}_m = \frac{1}{m} \sum_{i=1}^m h(x_i)$$

as an estimator of \mathcal{J} .

Justification: by the Strong Law of Large Numbers,

$$\hat{\mathcal{J}}_m \rightarrow \mathcal{J}.$$

Approximation evaluation

Approximation evaluation

Estimate the variance with

$$\nu_m = \frac{1}{m} \frac{1}{m-1} \sum_{i=1}^m (h(x_i) - \hat{d}_m)^2,$$

Approximation evaluation

Estimate the variance with

$$\nu_m = \frac{1}{m} \frac{1}{m-1} \sum_{i=1}^m (h(x_i) - \hat{\mathcal{J}}_m)^2,$$

and recall that for m large,

$$\frac{\hat{\mathcal{J}}_m - \mathbb{E}[h(X)]}{\sqrt{\nu_m}} \approx \mathcal{N}(0, 1).$$

Importance sampling

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

An alternative to direct sampling is importance sampling, with the following trick:

Importance sampling

Simulating from f might not be the best idea: difficult/impossible, not optimal, ...

An alternative to direct sampling is importance sampling, with the following trick:

$$\begin{aligned}\mathbb{E}_{X \sim f}[h(X)] &= \int h(x)f(x)dx = \int h(x)\frac{f(x)}{g(x)}g(x)dx \\ &= \mathbb{E}_{X \sim g}\left[h(X)\frac{f(X)}{g(X)}\right],\end{aligned}$$

which allows us to use other distributions.

Importance sampling

Importance sampling

Sample a sequence $x_1, \dots, x_m \sim g$.

Importance sampling

Sample a sequence $x_1, \dots, x_m \sim g$.

Then use

$$\hat{\mathcal{J}}_m = \frac{1}{m} \sum_{i=1}^m \frac{f(x_i)}{g(x_i)} h(x_i)$$

as an estimator of \mathcal{J} .

Justification: by the Strong Law of Large Numbers,

$$\hat{\mathcal{J}}_m \rightarrow \mathcal{J}.$$

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f .

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f .
2. Instrumental distribution g may be chosen among distributions easy to simulate.

Justification

1. Converges for any choice of the distribution g as long as the support of g contains the support of f .
2. Instrumental distribution g may be chosen among distributions easy to simulate.
3. The same sample generated from g can be used repeatedly, not only for different functions h but also for different densities f .

Choice of importance function

Choice of importance function

The optimal choice is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx},$$

but the integral is unknown (obviously).

Choice of importance function

The optimal choice is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Choice of importance function

The optimal choice is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

Choice of importance function

The optimal choice is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

If $\sup f/g = M < +\infty$, the accept-reject algorithm may be used.

Choice of importance function

The optimal choice is

$$g^*(x) = \frac{|h(x)|f(x)}{\int |h(x)|f(x)dx},$$

but the integral is unknown (obviously).

In practice, pick a density g which is close enough to $|h|f$ and for which $|h|f/g$ is bounded.

Beware: importance sampling suffers from the curse of dimensionality.

If $\sup f/g = M < +\infty$, the accept-reject algorithm may be used.

The instrumental function may be π (the prior). But often inefficient if data informative, and impossible if π is improper...

Sampling random variables

The fundamental theorem of sampling:

Sampling random variables

The fundamental theorem of sampling: if $U \sim \mathcal{U}(0, 1)$, let F denote a CDF, then $F^-(U)$ is a random variable with distribution dF .

Sampling random variables

The fundamental theorem of sampling: if $U \sim \mathcal{U}(0, 1)$, let F denote a CDF, then $F^-(U)$ is a random variable with distribution dF .

In practice, this theorem has a very limited scope since the pseudo-inverse F^- is usually unknown/not analytically tractable.

Accept-reject

Goal: sample $x \sim f$.

Accept-reject

Goal: sample $x \sim f$.

Ingredients: a density g such that

- ▶ f/g is upper-bounded by M (known)
- ▶ sampling from g is easy

Accept-reject

Goal: sample $x \sim f$.

Ingredients: a density g such that

- ▶ f/g is upper-bounded by M (known)
- ▶ sampling from g is easy

The algorithm:

1. Sample $z \sim g$ and $u \sim \mathcal{U}(0, Mg(z))$
2. If $u \leq f(z)$, take $x = z$, otherwise go back to 1.

How should we choose g ?

The mean number of samples to obtain 1 realization is M !
Efficient algorithm = M close to 1.

How should we choose g ?

The mean number of samples to obtain 1 realization is M !
Efficient algorithm = M close to 1.

In practice we pick a g similar to f .

How should we choose g ?

The mean number of samples to obtain 1 realization is M !
Efficient algorithm = M close to 1.

In practice we pick a g similar to f .

Nice fact: no need to know the normalizing constant of f !

MCMC

MCMC

MCMC stands for Monte Carlo Markov Chain.

MCMC

MCMC stands for Monte Carlo Markov Chain.

Key idea: easier to sample a Markov chain with stationary distribution f rather than independent variables with distribution f .

MCMC

MCMC stands for Monte Carlo Markov Chain.

Key idea: easier to sample a Markov chain with stationary distribution f rather than independent variables with distribution f .

Reminder: a Markov chain is a sequence of variables (X_t) such that the conditional law of X_t only depends on X_{t-1} .

MCMC

MCMC stands for Monte Carlo Markov Chain.

Key idea: easier to sample a Markov chain with stationary distribution f rather than independent variables with distribution f .

Reminder: a Markov chain is a sequence of variables (X_t) such that the conditional law of X_t only depends on X_{t-1} .

Examples: let (u_t) be a sequence of i.i.d variables

MCMC

MCMC stands for Monte Carlo Markov Chain.

Key idea: easier to sample a Markov chain with stationary distribution f rather than independent variables with distribution f .

Reminder: a Markov chain is a sequence of variables (X_t) such that the conditional law of X_t only depends on X_{t-1} .

Examples: let (u_t) be a sequence of i.i.d variables

- ▶ $X_0 \sim \pi_0$ and $X_t = \rho X_{t-1} + u_t$
- ▶ More generally $X_t = g(X_{t-1}, u_t)$ for any measurable function g

Objectives

Goal: sample from a target distribution ρ

Objectives

Goal: sample from a target distribution ρ

- ▶ to generate a sequence $(x^t)_{t=1}^T$ of realizations,

Objectives

Goal: sample from a target distribution ρ

- ▶ to generate a sequence $(x^t)_{t=1}^T$ of realizations,
- ▶ to compute expectations $\mathbb{E}[f(x)] \simeq \frac{1}{T} \sum_{t=1}^T f(x^t)$,

Objectives

Goal: sample from a target distribution ρ

- ▶ to generate a sequence $(x^t)_{t=1}^T$ of realizations,
- ▶ to compute expectations $\mathbb{E}[f(x)] \simeq \frac{1}{T} \sum_{t=1}^T f(x^t)$,
- ▶ to maximize functions $\arg \max_x \phi(x)$,

Objectives

Goal: sample from a target distribution ρ

- ▶ to generate a sequence $(x^t)_{t=1}^T$ of realizations,
- ▶ to compute expectations $\mathbb{E}[f(x)] \simeq \frac{1}{T} \sum_{t=1}^T f(x^t)$,
- ▶ to maximize functions $\arg \max_x \phi(x)$,
- ▶ ...

Objectives

Goal: sample from a target distribution ρ

- ▶ to generate a sequence $(x^t)_{t=1}^T$ of realizations,
- ▶ to compute expectations $\mathbb{E}[f(x)] \simeq \frac{1}{T} \sum_{t=1}^T f(x^t)$,
- ▶ to maximize functions $\arg \max_x \phi(x)$,
- ▶ ...

Burn-in period to reach convergence and stabilize the algorithm.

Construction and key properties

Construction and key properties

A Markov chain is fully defined by

- ▶ The distribution of X_0
- ▶ The distribution of X_t conditionally to X_{t-1} (transition dynamic)

Construction and key properties

A Markov chain is fully defined by

- ▶ The distribution of X_0
- ▶ The distribution of X_t conditionally to X_{t-1} (transition dynamic)

Irreducible chain: every region of the state may be reached.

- ▶ Transient: the mean number of passages is finite
- ▶ Recurrent: coming back is assured

Invariant distribution: the chain admits an invariant distribution f if there exists a density f such that

$$x_t \sim f \implies x_{t+1} \sim f.$$

The chains built by MCMC algorithms admit a unique invariant distribution.

Convergence

Convergence

Given a density f , we are interested in transition dynamics such that

- ▶ The invariant distribution is unique (f)
- ▶ The distribution of x_t is "close" to the invariant density whenever t is large enough (total variation norm).
- ▶ Ergodic theorem

$$\frac{1}{T} \sum_{t=1}^T h(x_t) \xrightarrow{T \rightarrow \infty} \int h(x)f(x)dx$$

Metropolis-Hastings (MH) algorithm

Metropolis-Hastings (MH) algorithm

Goal: sample $x \sim f$.

Metropolis-Hastings (MH) algorithm

Goal: sample $x \sim f$.

Ingredient: a density $q(\cdot|x_t)$ such that $\text{supp}(f) \subset \text{supp}(q \cdot |x)$

Metropolis-Hastings (MH) algorithm

Goal: sample $x \sim f$.

Ingredient: a density $q(\cdot|x_t)$ such that $\text{supp}(f) \subset \text{supp}(q \cdot |x)$

Algorithm: at time $t + 1$,

Metropolis-Hastings (MH) algorithm

Goal: sample $x \sim f$.

Ingredient: a density $q(\cdot|x_t)$ such that $\text{supp}(f) \subset \text{supp}(q \cdot |x)$

Algorithm: at time $t + 1$,

$$x_{t+1} = \begin{cases} z \sim q(\cdot|x_t) & \text{with probability } \rho, \\ x_t & \text{otherwise} \end{cases}$$

where

$$\rho = \min \left\{ 1, \frac{f(z)q(x_t|z)}{f(x_t)q(z|x_t)} \right\}.$$

Key properties

The chain produced by the MH algorithm

Key properties

The chain produced by the MH algorithm

1. is irreducible
2. is ergodic
3. admits f as an invariant distribution

Examples of instrumental / proposal distribution

Examples of instrumental / proposal distribution

1. Independent distributions

$$q(z|x) = g(z).$$

This generalizes the accept-reject algorithm.

Examples of instrumental / proposal distribution

1. Independent distributions

$$q(z|x) = g(z).$$

This generalizes the accept-reject algorithm.

2. Symmetric distributions

$$q(z|x_t) = h(|z - x_t|).$$

The acceptance ratio does not depend on q :

$$\rho = \min \left(1, \frac{f(z)}{f(x_t)} \right).$$

Random walks

Given x_t , the transition dynamic writes

$$z = x_t + \epsilon$$

where (ϵ_t) is a sequence of i.i.d variables, independent from (x_t) , and the distribution of ϵ is symmetric with respect to 0. Examples:

- ▶ normal distributions $\mathcal{N}(0, \sigma^2)$
- ▶ uniform distribution on symmetric intervals $\mathcal{U}(-a, a)$
- ▶ ...

Gibbs sampler

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Key idea: decomposition of $x \in \mathbb{R}^p$ in blocks such that the conditional distributions $f_j(x_j|x_i, i \neq j)$, $j = 1, \dots, p$ are easy to sample.

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Key idea: decomposition of $x \in \mathbb{R}^p$ in blocks such that the conditional distributions $f_j(x_j|x_i, i \neq j)$, $j = 1, \dots, p$ are easy to sample.

Algorithm: the chain transition dynamic is given by

1. Initialization: (x_1^1, \dots, x_d^1)

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Key idea: decomposition of $x \in \mathbb{R}^p$ in blocks such that the conditional distributions $f_j(x_j|x_i, i \neq j)$, $j = 1, \dots, p$ are easy to sample.

Algorithm: the chain transition dynamic is given by

1. Initialization: (x_1^1, \dots, x_d^1)
2. $x_j^{t+1} \sim f(x_j|x_1^{t+1}, \dots, x_{j-1}^{t+1}, x_{j+1}^t, \dots, x_d^t)$

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Key idea: decomposition of $x \in \mathbb{R}^p$ in blocks such that the conditional distributions $f_j(x_j|x_i, i \neq j)$, $j = 1, \dots, p$ are easy to sample.

Algorithm: the chain transition dynamic is given by

1. Initialization: (x_1^1, \dots, x_d^1)
2. $x_j^{t+1} \sim f(x_j|x_1^{t+1}, \dots, x_{j-1}^{t+1}, x_{j+1}^t, \dots, x_d^t)$
3. Iterate until convergence

Gibbs sampler

Goal: sample $x \sim f$ in dimension strictly larger than 1.

Key idea: decomposition of $x \in \mathbb{R}^p$ in blocks such that the conditional distributions $f_j(x_j|x_i, i \neq j)$, $j = 1, \dots, p$ are easy to sample.

Algorithm: the chain transition dynamic is given by

1. Initialization: (x_1^1, \dots, x_d^1)
2. $x_j^{t+1} \sim f(x_j|x_1^{t+1}, \dots, x_{j-1}^{t+1}, x_{j+1}^t, \dots, x_d^t)$
3. Iterate until convergence

The chain is

1. ergodic
2. with f as invariant distribution

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg \min_x f(x)$).

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg \min_x f(x)$).

Input: tolerance ϵ , initialization x_0 , step size α .

Gradient Descent

Goal: minimize a differentiable function f (compute $\arg \min_x f(x)$).

Input: tolerance ϵ , initialization x_0 , step size α .

While $f'(x_k) \geq \epsilon$ $x_{k+1} = x_k - \alpha f'(x_k)$

PAC-Bayes in practice

Previous instantiations of $\widehat{\phi}_\lambda$ are not tractable.

PAC-Bayes in practice

Previous instantiations of $\widehat{\phi}_\lambda$ are not tractable.

Instead of an infinite-dimensional functional space \mathcal{F} , we often resort to some projection onto \mathbb{R}^d .

PAC-Bayes in practice

Previous instantiations of $\widehat{\phi}_\lambda$ are not tractable.

Instead of an infinite-dimensional functional space \mathcal{F} , we often resort to some projection onto \mathbb{R}^d .

Sampling from a d -dimensional non-standard distribution is still an algorithmic challenge.

Existing implementation

► (Transdimensional) MCMC

- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

Existing implementation

► (Transdimensional) MCMC

- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models,
Electronic Journal of Statistics
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

► Stochastic optimization

- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization,
Mathematical Methods of Statistics

Existing implementation

► (Transdimensional) MCMC

- Guedj and Alquier (2013). PAC-Bayesian Estimation and Prediction in Sparse Additive Models, *Electronic Journal of Statistics*
- Alquier and Biau (2013). Sparse Single-Index Model, *Journal of Machine Learning Research*
- Guedj and Robbiano (2017). PAC-Bayesian High Dimensional Bipartite Ranking, *Journal of Statistical Planning and Inference*
- Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

► Stochastic optimization

- Alquier and Guedj (2017). An Oracle Inequality for Quasi-Bayesian Non-Negative Matrix Factorization, *Mathematical Methods of Statistics*

► Variational Bayes

- Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors, *Journal of Machine Learning Research*

Bridging the gap between theory and algorithms

Goal: PAC oracle inequalities for approximations of $\hat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Bridging the gap between theory and algorithms

Goal: PAC oracle inequalities for approximations of $\hat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Let $\tilde{\rho}_\lambda$ denote a VB approximation of $\hat{\rho}_\lambda$. The rate of convergence in PAC inequalities is of analogous magnitude for $\tilde{\rho}_\lambda$ and $\hat{\rho}_\lambda$.

■ Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors,

Journal of Machine Learning Research

Bridging the gap between theory and algorithms

Goal: PAC oracle inequalities for approximations of $\hat{\rho}_\lambda$ (echoes the celebrated statistical / computational tradeoff).

Let $\tilde{\rho}_\lambda$ denote a VB approximation of $\hat{\rho}_\lambda$. The rate of convergence in PAC inequalities is of analogous magnitude for $\tilde{\rho}_\lambda$ and $\hat{\rho}_\lambda$.

■ Alquier, Ridgway and Chopin (2016). On the properties of variational approximations of Gibbs posteriors,
Journal of Machine Learning Research

MCMC for online (sequential) quasi-Bayesian learning: the stationary distribution of the Markov Chain is indeed $\hat{\rho}_\lambda$.

■ Li, Guedj and Loustau (2018). A Quasi-Bayesian Perspective to Online Clustering, *arXiv preprint*

Conclusion

Take-home messages

- ▶ Sound mathematical foundations for statistical learning
(machine learning with a probabilistic perspective)

Take-home messages

- ▶ Sound mathematical foundations for statistical learning
(machine learning with a probabilistic perspective)
- ▶ Frequentist vs. Bayesian paradigms: uncertainty matters!

Take-home messages

- ▶ Sound mathematical foundations for statistical learning
(machine learning with a probabilistic perspective)
- ▶ Frequentist vs. Bayesian paradigms: uncertainty matters!
- ▶ Bayesian models vs. quasi-Bayesian prediction

Take-home messages

- ▶ Sound mathematical foundations for statistical learning
(machine learning with a probabilistic perspective)
- ▶ Frequentist vs. Bayesian paradigms: uncertainty matters!
- ▶ Bayesian models vs. quasi-Bayesian prediction
- ▶ Algorithmic challenges: tractable methods which scale up to modern (massive and complex) data

Take-home messages

- ▶ Sound mathematical foundations for statistical learning (machine learning with a probabilistic perspective)
- ▶ Frequentist vs. Bayesian paradigms: uncertainty matters!
- ▶ Bayesian models vs. quasi-Bayesian prediction
- ▶ Algorithmic challenges: tractable methods which scale up to modern (massive and complex) data
- ▶ A very exciting field to work in!



Dernière séance : TP

En Python ou dans un notebook Jupyter.

1. Calculer numériquement l'intégrale

$\sqrt{\pi} \int \exp(-3x) \log(1 + x^3) dx$ par Monte Carlo, et fournir un intervalle de confiance.

2. Calculer numériquement l'intégrale

$\sqrt{\pi} \int \exp(-3x) \log(1 + x^3) dx$ par échantillonnage d'importance, et comparer cette approximation à celle obtenue par simple Monte Carlo.

3. Echantillonner la distribution $1/3\mathcal{N}(-1, 1) + 2/3\mathcal{N}(2, 3/2)$ par un algorithme d'acceptation-rejet.

4. Echantillonner la distribution

$1/4\mathcal{N}(-1, 2) + 1/2\mathcal{N}(2, 3/2) + 1/4\mathcal{N}(4, 1)$ par l'algorithme de Metropolis-Hastings.



Benjamin Guedj, Ph.D.

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