Data analysis and stochastic modeling

Lecture 4 – Machine learning and estimation theory

Guillaume Gravier

guillaume.gravier@irisa.fr

Why statistical modeling?

exploratory statistics $\longrightarrow$ inferential statistics

describe
analyze

generalize
decide

$\rightarrow$ summarize data using a (parametric) model

$\rightarrow$ estimate the parameters of a model

$\rightarrow$ decision, discrimination, classification

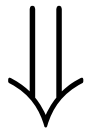
$\rightarrow$ prediction

Why statistical modeling?

Summarize all the data available in a model for which the number of parameters is small with respect to the amount of data

Use **all** the information available!

1. Prior knowledge of what we expect (and do not expect!) to see
2. Data, data, data, data... (the best data is more data)



STATISTICAL MACHINE LEARNING

NOTIONS OF MACHINE LEARNING

Slides courtesy of Samy Bengio.

Various problems to solve

- Let $Z_1, Z_2, \cdots, Z_n$ be an n-tuple random sample of an unknown distribution of density $p(z)$.
- All Z_i are independently and identically distributed (iid).

1. Classification: $Z = (X, Y) \in \mathbb{R}^d \times \{-1, 1\}$

   $\Rightarrow$ given a new x, estimate $P(Y|X = x)$

2. Regression: $Z = (X, Y) \in \mathbb{R}^d \times \mathbb{R}$

   $\Rightarrow$ given a new x, estimate $E[Y|X = x]$

3. Density estimation: $Z \in \mathbb{R}^d$

   $\Rightarrow$ given a new z, estimate $p(z)$

The function space

Learning = search for a good function in a function space $\mathcal{F}$

Examples of parametric functions:

- Regression

$$\hat{y} = f(x; a, b) = a \cdot x + b$$

- Classification

$$\hat{y} = f(x; a, b) = \text{sign}(a \cdot x + b)$$

- Density estimation

$$\hat{p}(z) = f(z; \mu, \Sigma) = \frac{1}{(2\pi)^{\frac{|z|}{2}} \sqrt{|\Sigma|}} \exp\left(-\frac{1}{2}(z-\mu)^T \Sigma^{-1} (z-\mu)\right)$$

The loss function

Learning = search for a good function in a function space $\mathcal{F}$

Examples of loss functions $L : \mathcal{Z} \times \mathcal{F}$

- Regression

$$L(z, f) = L((x, y), f) = (f(x) - y)^2$$

- Classification

$$L(z, f) = L((x, y), f) = \begin{cases} 0 & \text{if } f(x) = y \\ 1 & \text{otherwise} \end{cases}$$

- Density estimation

$$L(z, f) = -\log p(z)$$

Risk and empirical risk

- Minimize the expected risk on $\mathcal{F}$, defined for a given function f as

$$R(f) = E_Z[L(z, f)] = \int_Z L(z, f)p(z)dz$$

- Induction Principle
  - find $f \in \mathcal{F}$ which minimizes $R(f)$
  - problems: $p(z)$ is unknown, and we don't have access to all $L(z, f)$!!!
- Empirical Risk

$$\hat{R}(f, D_n) = \frac{1}{n}\sum_{i=1}^{n} L(z_i, f)$$

Risk and empricial risk (cont'd)

- The empirical risk:

$$\hat{R}(f, D_n) = \frac{1}{n} \sum_{i=1}^{n} L(z_i, f)$$

- The (expected) risk:

$$R(f) = E_Z[L(z, f)] = \int_Z L(z, f) p(z) dz$$

- The empirical risk is an unbiased estimate of the risk

The principle of empirical risk minimization (ERM):

$$f^\star(D_n) = \arg\min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

Risk and empricial risk (cont'd)

- Training error:

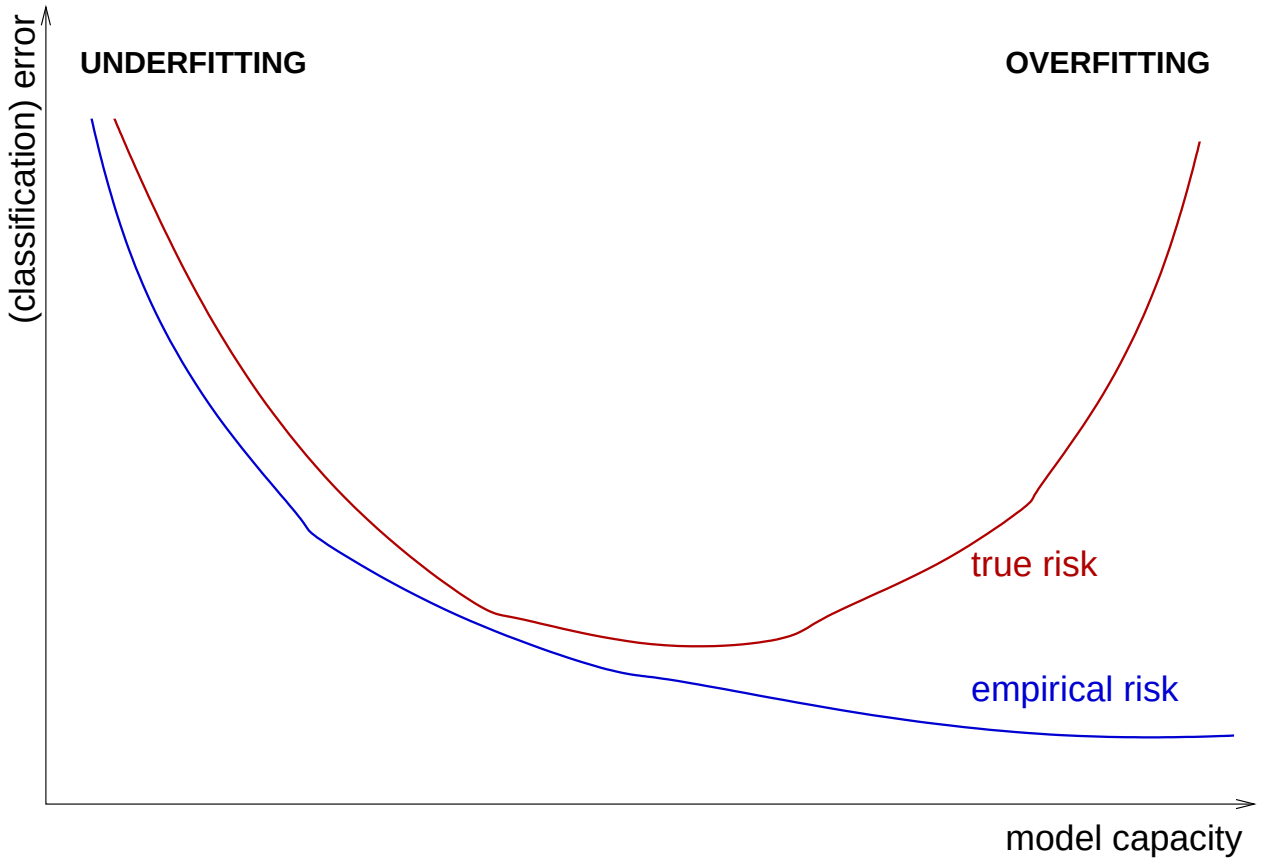
$$\hat{R}(f^\star(D_n), D_n) = \min_{f \in \mathcal{F}} \hat{R}(f, D_n)$$

- Is the training error a biased estimate of the risk? YES.

$$E[R(f^\star(D_n)) - \hat{R}(f^\star(D_n), D_n)] \geq 0$$

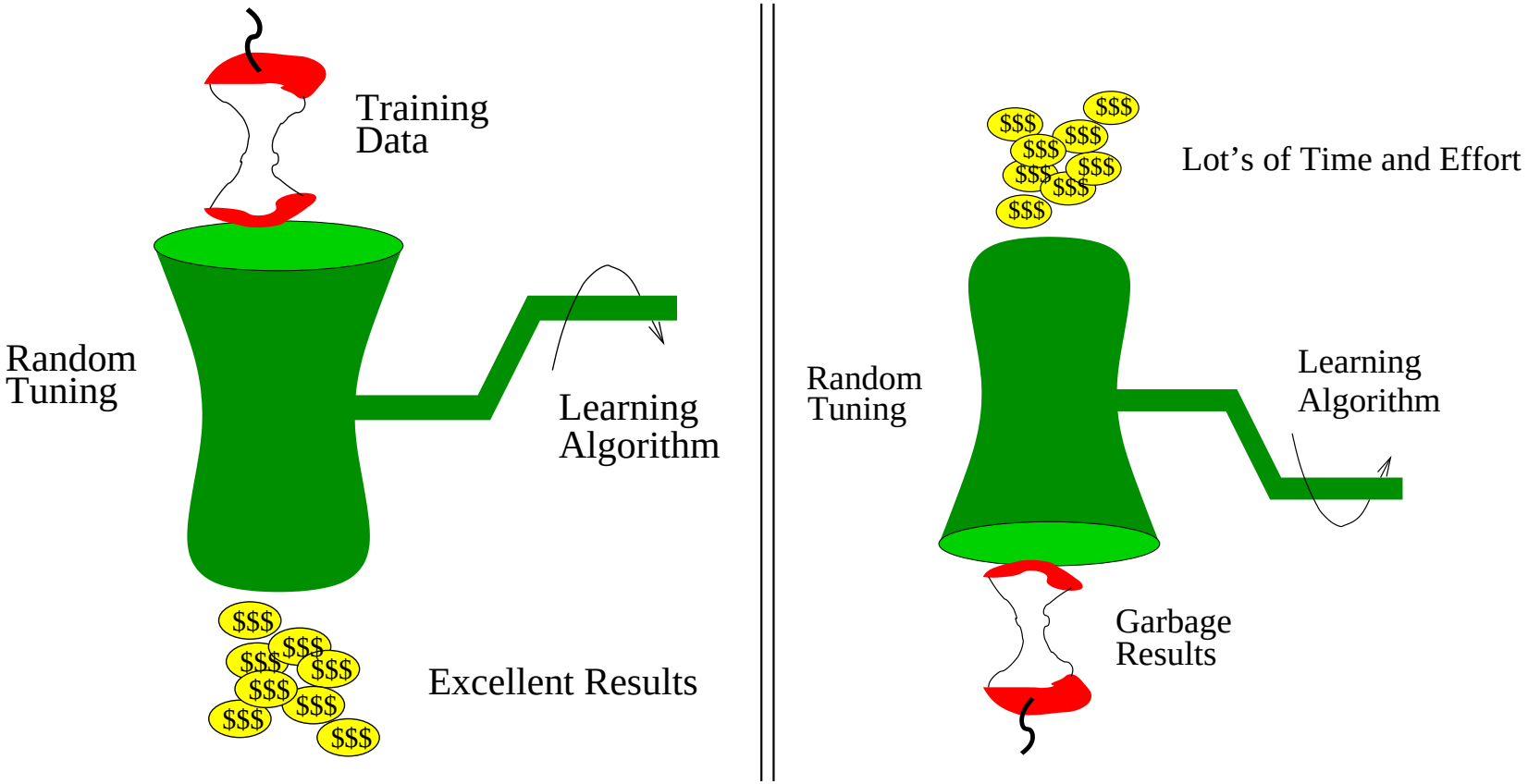
- The solution $f^\star(D_n)$ found by minimizing the training error is better on D_n than on any other set $D_n^{'}$ drawn from $p(z)$.

Risk and empirical risk



$\Longrightarrow$ Don't fit too much on a (limited) training set!

Is it magic?


Training
Data
Random
Tuning
Learning
Algorithm
$$$
Excellent Results
Lot's of Time and Effort
Random
Tuning
Learning
Algorithm
Garbage
Results

Statistical models and machine learning

Statistical modeling and decision theory can be seen as a particular “subset” of the general machine learning theory

- function space limited to probability mass/density functions
- only classical decision rules for classification problems
- boils down to density estimation

with nice properties concerning the quality of the estimated functions!

Introductory example

[borrowed from A. Rakotomamonjy and G. Gasso, INSA Rouen]

We wish to classify a pixel into class 1 (dark) and class 2 (light).

If we **know nothing on the pixel**:

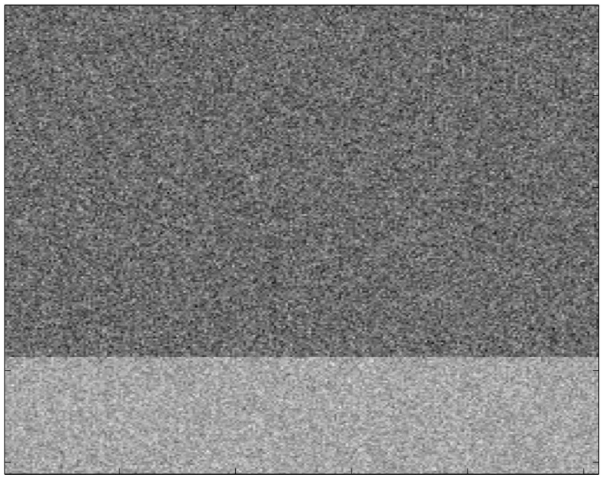
$\rightarrow$ $\max_k P[C_k]$

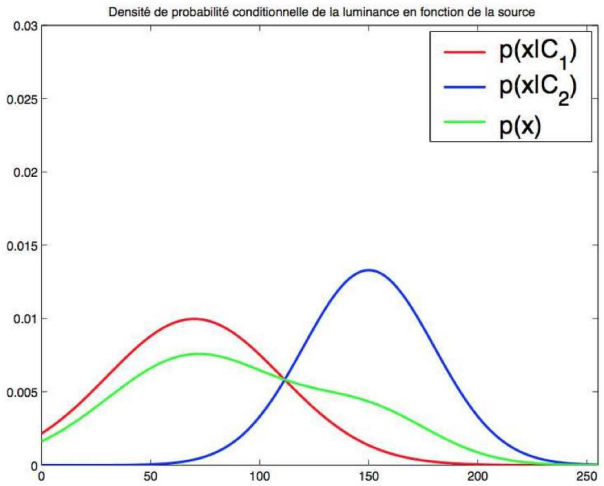
$\rightarrow$ It's dark!

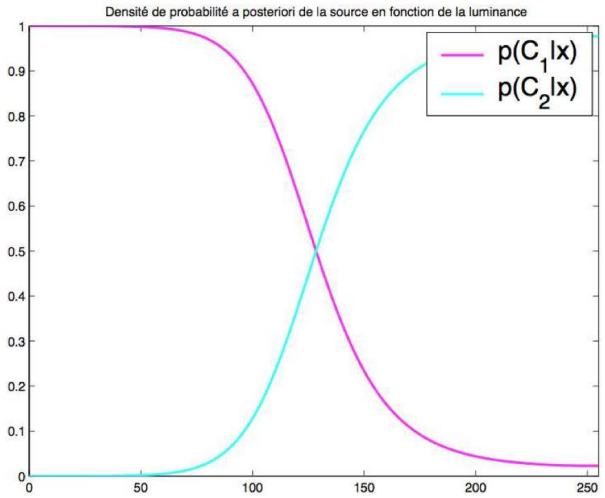
If we **know the pixel value**:

$\rightarrow$ make use of $P[x|C_k]$

$\rightarrow$ choose according to $P[C_k|x]$







Optimal decision rules

- Maximum likelihood model esimtation

$$\widehat{\theta} = \arg\max_{\theta} p(x;\theta)$$

- Classification

  Maximum a posteriori $\quad \widehat{c} = \arg\max_c p(c|x) = \arg\max_c p(x|c)p(c)$

  Maximum likelihood $\quad \widehat{c} = \arg\max_c p_c(x)$

  Maximum likelihood is a particular case of maximum a posterior with $p(c) \rightsquigarrow \mathcal{U}$.

- Hypothesis testing

$$\frac{p(x;H_0)}{p(x;H_1)} \begin{array}{c} H_0 \\ > \\ < \\ H_1 \end{array} \beta$$

Why maximum a posteriori?

$$D : x \in X \longrightarrow y = D(x) \in \{1, \ldots, K\}$$

For a cost function l_{jk} (cost of deciding class j when it's in fact class k), the (conditional) risk of deciding class j after observing x is given by

$$R(D(x) = j|x) = \sum_{k=1}^{K} l_{jk} P[\text{class}(x) = k]$$

and leads to the theoretical (average) risk defined as

$$E[R(D(x))] = \int_X R(D(x)|x)p(x)dx$$

The MAP (aka Bayes) decision rule corresponds to choosing the class i for a sample x such that

$$R(D(x) = i|x) < R(D(x) = j|x) \quad \forall j \neq i$$

and minimizes $E[R(D(x))]$.

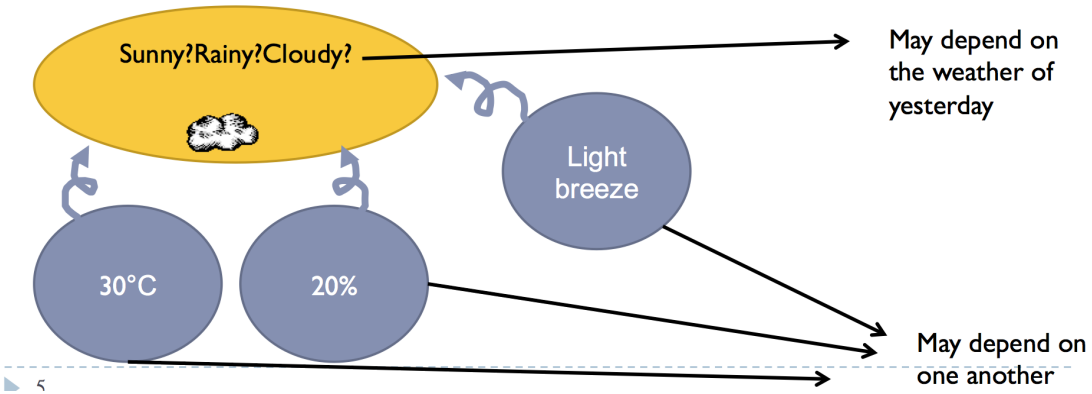
When Bayes gets naive

Given observed data x, we wish to predict a (discrete) label y. Bayesian (optimal) decision says that

$$\hat{y} = \arg\max_y p(y|x) = \arg\max_y p(x,y) = \arg\max_y p(x|y)p(y)$$

Example:

- y = weather of the day
- x = temprature, humidity, etc.

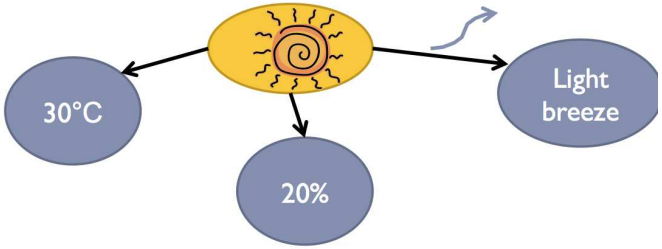


$\Rightarrow$ what model/law for $P(y)$ and $P(x|y)$?

When Bayes gets naive (cont'd)

The *naive Bayes* classifier assumes all observations to be (conditionally) independent

$$p(\mathbf{x}, y) = p(y) \prod_{i=1}^{n} p(x_i|y)$$



$$\arg\max_c P[Y = c|x_1, x_2, x_3] = \arg\max_c P[Y = c] \prod_{i=1}^{3} P[X_i = x_i|Y = c]$$

Parameter estimation

Statistical inference

Given a limited amount of samples from a population, infer the properties of the entire population.

Requires independent samples representative of the population

- sampling strategies exists (when sampling can be controled)
- assume this to be true in pattern recognition problems

Two inference strategies :

1. estimate basic characteristics, *e.g.* mean, variance or median, of the population from the samples
2. estimate the parameters of a model which has been selected from expert knowledge

Note : for simple models, estimating the mean and/or the variance is equivalent to estimating the parameters of the model.

Definition of a statistic

Each sample can be seen as a random variable X_i whose observed value is x_i, where all the variables X_i have the same distribution.

Example : suppose we extract n bulbs from a production line and measure their lifetime x_i. Assuming there has been no changes in the fabrication process, the values x_i can be considered as observations of a single random variable X. The model considers X_i the random variable corresponding to the lifetime of the i'th bulb, whose value is x_i. All the variables X_i follow the same distribution, that of X.

Definition

A statistic is a random variable which is a measurable function of $X_1, X_2, \ldots, X_n$, denoted $T = f(X_1, X_2, \ldots, X_n)$.

EMPIRICAL ESTIMATORS

Some well known statistics

Some well known statistics over a set of observations $X_1, \ldots, X_n$:

Empirical frequency $$F_k = \frac{1}{n} \sum_{i=1}^{n} \delta(X_i = k)$$

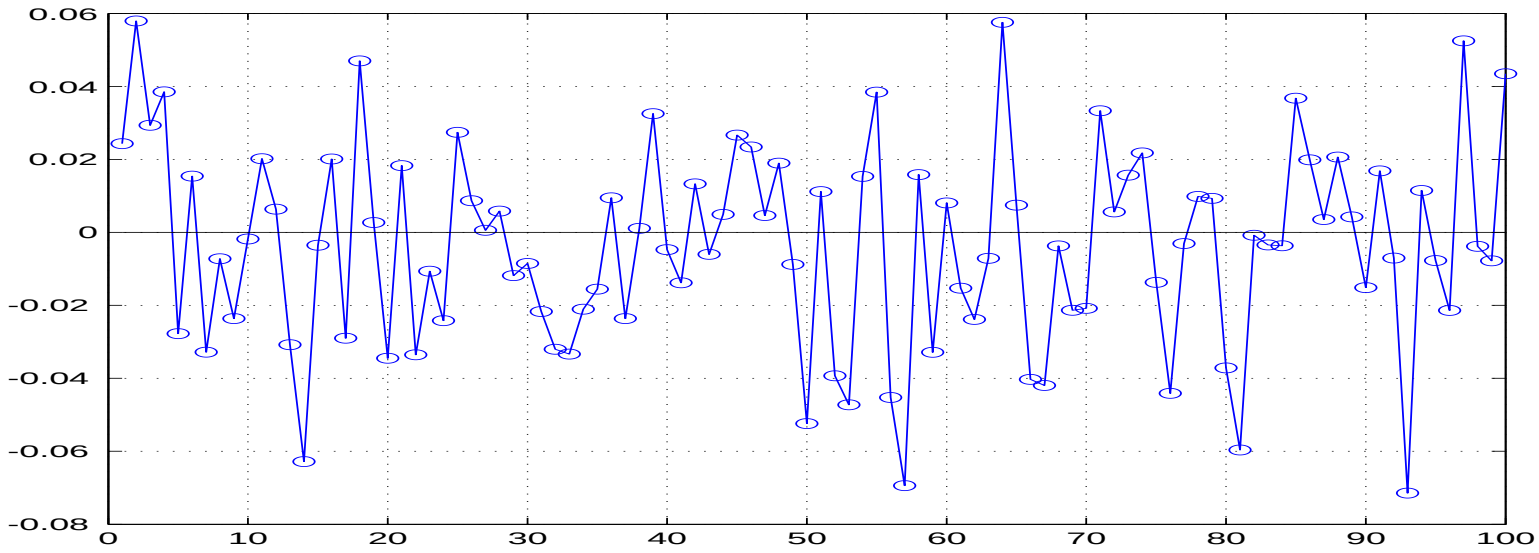
Empirical mean estimator $$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$$

Empirical variance estimator $$S^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

A statistic is a random variable since any function of random variables is a random variable.

Empirical mean estimation

$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is a *statistic* estimating the mean value of a population from a finite set of samples.



zero-mean unit-variance Gaussian, 1000 samples per trial

empirical mean = -0.004, empirical standard deviation = 0.0306

Distribution of the empirical mean

It can easily be shown that

$$E[\overline{X}] = m \qquad \text{and} \qquad V[\overline{X}] = \frac{\sigma^2}{n}$$

The estimator $\overline{X}$ converges in quadratic mean toward m when $n \to \infty$, since $E[(\overline{X} - m)^2] \to 0$.

Moreover, the central limit theorem states that

$$\frac{\overline{X} - m}{\sigma/\sqrt{n}} \xrightarrow{\mathcal{L}} \mathcal{N}(0,1)$$

Note on Gaussian variables: for Gaussian variables, the convergence is in fact an equality, *i.e.* if $X \rightsquigarrow \mathcal{N}(0,1)$, then $\overline{X} \rightsquigarrow \mathcal{N}(m, \frac{\sigma}{\sqrt{n}})$.

Empirical variance estimation

The empirical variance estimator is given by

$$S^2 = \frac{\sum (X_i - \overline{X})^2}{n} \qquad S^2 = \frac{\sum X_i^2}{n} - \overline{X}^2$$

$$S^2 = \frac{\sum (X_i - m)^2}{n} - (\overline{X} - m)^2$$

which leads to

$$E[S^2] = \frac{n-1}{n}\sigma^2$$

The estimator is biased, since $E[S^2] \neq \sigma^2$, but does converge when $n \to \infty$!

Empirical proportion estimation

Given the discrete random variables X_i whose values are in $[1, K]$, the proportion estimation for the event k (or equivalently the probability $p_k = P[X = k]$) is estimated by the statistic

$$F_k = \frac{\sum \delta(X_i = k)}{n}$$

where $\delta(X_i = k) = 1$ if $X_i = k$ and 0 otherwise.

It can easily be shown that

$$E[F_k] = p_k \qquad \text{and} \qquad V[F_k] = \frac{p_k(1-p_k)}{n}$$

According to the central-limit theorem, $F \to \mathcal{N}(p, \sqrt{\frac{p(1-p)}{n}})$.

A practical example

Problem: Imagine a production line where objects are manufactured with a given length, where we know from previous study that the (real) distribution of the length is a Gaussian of mean 10 and standard deviation 2. For quality control purposes, a set of 25 samples are taken from the production line. What is the range of values for which we have 9 chances over 10 of observing the empirical mean $\overline{X}$?

Solution: We have seen so far that $\overline{X} \to \mathcal{N}(10; \frac{2}{\sqrt{25}})$. Moreover, for a zero mean unit variance Gaussian variable U, $P(-1.64 < U < 1.64) = 0.9$. Hence, with a probability of 0.9, we have

$$10 - 1.64\frac{2}{\sqrt{25}} < \overline{X} < 10 + 1.64\frac{2}{\sqrt{25}}$$

and the range of value for $\overline{X}$ is $[9.34, 10.66]$.

INTRODUCTION TO THE THEORY OF ESTIMATION

Quality of an estimator

We have seen that the *empirical statistics* $\overline{X}$, S^2 and F are *estimators* of respectively the mean, the variance and the (discrete) probability, since they almost surely converge towards the true quantity (resp. m, σ^2 and p_k).

But other estimators can be used, *e.g.* for the mean

- α-truncated mean where the αn biggest and smallest values are discarded
- median value ($\alpha = 50\,\%$)
- mean extrema values ($(\max(X_i) + \min(X_i))/2$)
- a randomly chosen sample
- a constant value, *e.g.* 0

$\Longrightarrow$ need for a quality measure!

Expected qualities

Let us consider an estimator T of a parameter θ, obtained from a set of samples X_i. The expected quality of an estimator are:

- convergence: $T \to \theta$ when $n \to \infty$
- convergence speed: some estimators converges more rapidly than others
- risk: the risk is defined as the mean quadratic error

$$E_\theta[(T-\theta)^2] = (\underbrace{E[T]-\theta}_{\text{bias}})^2 + \underbrace{V[T]}_{\text{variance}}$$

$\Rightarrow$ Given two non biased estimators, the best one is the one with the smallest variance.

Generalized risk

The risk can be defined in a more general way using a loss/error funcion $l(t, \theta)$, where $l(\theta, t) = 0$ iif $t = \theta$,

$$R(T, \theta) = E_\theta[l(T(X), \theta)]$$

Example of error functions are

quadratic error	$l(a, b) = (a - b)^2$
absolute error	$l(a, b) = \lvert a - b \rvert$
ϵ-loss	$l(a, b) = 0$ if $\lvert a - b \rvert < \epsilon$

Warning: unfortunately, directly minimizing the risk is possible only in some very particular cases...

Comparing the risk of estimators

- It is possible to compare estimators based on the risk, even though the risk function is rarely easily defined
- An estimator T is *better than* an estimator T' if

$$R(T, \theta) < R(T', \theta) \qquad \forall \theta \in \Theta$$

- It is usually impossible to find an estimator T which is better than any other for all the values of θ – think of a constant estimator
- Except in some very special cases, there seldom is an estimator which is uniformly better than all the others

About biased estimators

A biased estimator is not necessarily a bad estimator!

- Let's consider the following biased and unbiased variance estimators assuming the mean m is known

$$T = \frac{1}{n}\sum(X_i - m)^2 \qquad S_0^2 = \frac{1}{n-1}\sum(X_i - m)^2$$

  It can be shown that T is better that S_0^2 since $V[T] < V[S_0^2]$.

- Let's consider the following biased and unbiased estimators of the autocorrelation $r(p) = E[X_i X_{i-p}]$ (assuming m is known and null)

$$R_0 = \frac{1}{n-p}\sum_{i=1}^{n-p} X_i X_{i+p} \qquad R_1 = \frac{1}{n}\sum_{i=1}^{n-p} X_i X_{i+p}$$

  R_0 is unbiased but with a huge variance when $p \to n$, in which case R_1 is often prefered.

Unbiased minimum variance estimators

The problem of finding the best estimator cannot be solved in all generality and we therefore put limits to the problem at hand.

- class of estimators
- still cannot solve the risk minimization problem in most of the cases

$\Rightarrow$ search for a given law family $f(x, \theta)$ the unbiased estimator of θ with the minimal variance, the search of which is related to the notion of *sufficient statistic*.

Sufficient statistics

A sufficient statistic is a statistic which contains all the information carried by the samples X_i on θ.

Let us denote

- $L(x_1, x_2, \ldots, x_n; \theta)$ the density or mass function of $(X_1, \ldots, X_n)$,
- T a statistic whose density or mass function is given by $g(t; \theta)$

Fisher's factorization theorem

> *T is a sufficient statistic if* $L(\mathbf{x}, \theta) = g(t, \theta)h(\mathbf{x})$*, or, in other words, if the density of* $\mathbf{x}$ *conditionnaly to* T *is independant of* θ*.*

The idea of the definition is the following: if, when T is known, the conditional density of $(X_1, \ldots, X_n)$ no longer depends on θ, then T carries all the information concerning θ.

Examples of sufficient statistics

- Gaussian law, m known, σ unknown

$$L(\mathbf{x}, \theta) = \frac{1}{\sigma^n \sqrt{n} 2\pi} \exp\left(-\frac{1}{2}\left(\frac{x_i - m}{\sigma}\right)\right)$$

For the statistic $T = \sum (X_i - m)^2$, it can be shown that $T/\sigma^2 \rightsquigarrow \chi^2_n$ and hence that $L(\mathbf{x}, \theta) = g(t, \theta) h(\mathbf{x})$.

- Poisson with λ unknown

$$L(\mathbf{x}, \theta) = exp(-n\lambda) \frac{\lambda \sum x_i}{\prod x_i!}$$

The statistic $S = \sum X_i$ is exhaustive and $S \rightsquigarrow \mathcal{P}(n\lambda)$.

$\Rightarrow$ can tell if a statistic is exhaustive but does not tell how to find one if ever there exists one!

Theorem of Darmois

Theorem of Darmois

A necessary and sufficent condition for a sample $(X_1, \ldots, X_n)$ to admit a sufficient statistic is that the density be from the exponential family, i.e.

$$f(\mathbf{x}, \theta) = \exp\left(a(x)\alpha(\theta) + b(x) + \beta(\theta)\right)$$

Under certain conditions on the function a, the statistic $T = \sum a(X_i)$ is sufficient.

- Note that the theorem applies only if the definition domain of X does not depend on θ
- In fact, there exists efficient (*i.e.* unbiased minimum variance) estimators only for the exponential family
- Most common laws are from the exponential family (except those with a term of the form x^θ)

More sufficient statistics

The density of the law γ_θ is given by

$$\ln f(x,\theta) = -x + (\theta - 1)\ln(x) - \ln(\Gamma(\theta))$$

and the statistic $\sum \ln(X_i)$ is sufficient according to the previous theorem.

More examples of sufficient statistics:

Bernoulli with parameter p	$\sum X_i$
Gaussian, m unknown, σ known	$\sum X_i$
Gaussian, m known, σ unknown	$\sum (X_i - m)^2$
Gaussian, m and σ unknown	$(\overline{X}, S^2)$
exponential law	$\sum X_i$

The role of sufficient statistics

Theorem of Rao-Blackwell

If T is an unbiased estimator of θ and U a sufficient statistic for θ, then $T^* = E[T|U]$ is an unbiased estimator of θ at least as good as T.

Theorem

If there exist a sufficient statistic U for θ, then the unique unbiased minimum variance estimator T of θ only depends on U.

Theorem of Lehmann-Scheffe

If T^* is an unbiased estimator of θ depending of a complete sufficient statistics U, then T^* is the unique unbiased minimum variance estimator of θ. In particular, if T is an unbiased estimator of θ, then $T^* = E[T|U]$.

In other words, an unbiased estimator function of a complete sufficient statistic is the best possible estimator.

ESTIMATION TECHNIQUES

Moment based methods

Principle: express analytically the moments as a function of the parameter and estimate the parameter value based on the empirical estimates.

Let g_i be functions such that $\forall \theta \in \Theta, E_\theta[g_i(X)] < \infty$. Typical such functions are $g_i(x) = x^i$ or $g_i(x) = I(x \in \Delta_i)$.

Moment estimates are solutions to the equation system given by

$$E_\theta[g_i(x)] = \overline{\mu}_i$$

Moment based methods: an example

- Lifetime of a component represented by the distribution

$$f(x;\alpha,\lambda) = [\lambda^{\alpha}/\Gamma(\alpha)]x^{\alpha-1}\exp(-\lambda x)$$

- Let's consider the two moments

$$\mu_1(\theta) = E_\theta[X] = \alpha/\lambda \quad \text{and} \quad \mu_2(\theta) = E_\theta[X^2] = \alpha(1+\alpha)/\lambda^2$$

- This equation system has a single solution

$$\alpha = (\mu_1(\theta)/\sigma(\theta))^2 \quad \text{and} \quad \lambda = \mu_1(\theta)/\sigma^2(\theta)$$

where $\sigma^2(\theta) = \mu_2(\theta) - \mu_1^2(\theta)$.

- By replacing the moment by their empirical estimates, the moment estimates are obtained.

Maximum likelihood estimators

Let $X = (X_1, \ldots, X_n)$ be random variables in $\mathbb{R}^d$, and $p(x;\theta)$ the density. The likelihood is the joint density of the observations, seen as a function of $\theta \to p(x;\theta)$

The maximum likelihood estimator $\widehat{\theta}(X)$ is such that

$$p(x;\widehat{\theta}(X)) \geq \max_{\theta\in\Theta} p(x;\theta)$$

and, if $p(x;\theta)$ is differentiable, it is given by

$$\frac{\partial p(x;\theta)}{\partial\theta} = 0 \ .$$

Note: the maximum likelihood estimator looks for the best fit of the (training) samples, assuming that the observations were the most probable.

Maximum likelihood estimators (cont'd)

In practice, we often use

$$\frac{\partial \ln p(x;\theta)}{\partial \theta} = 0 \ .$$

Moreover, if the X_i's are iid, then

$$\ln p(x;\theta) = \sum \ln p(x_i;\theta)$$

Maximum likelihood estimators: an example

Assume $X_i \rightsquigarrow \mathcal{N}(\mu, \sigma^2)$. The log-likelihood is given by

$$\ln p(x; \mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2$$

The maximum likelihood equations are given by

$$\frac{\partial \ln p(x;\mu,\sigma^2)}{\partial \mu} = 0 \qquad \text{and} \qquad \frac{\partial \ln p(x;\mu,\sigma^2)}{\partial \sigma^2} = 0$$

for which the solutions are given by

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widehat{\mu})^2$$

Theory of maximum likelihood estimation

Maximum likelihood estimation is related to unbiased minimum variance estimation and moment estimators.

- If there exists a sufficient statistics U, the maximum likelihood estimator depends on it.

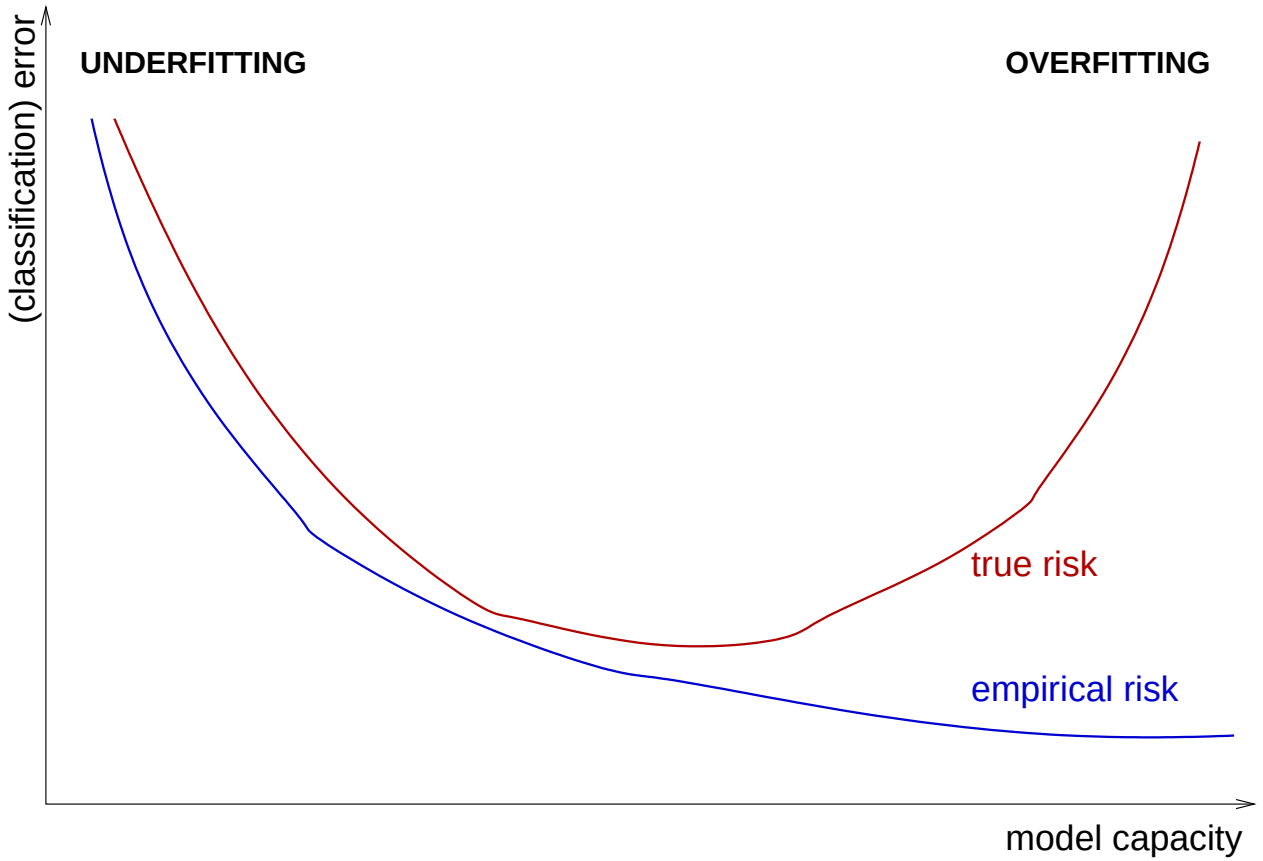
$$p(x,\theta) = g(u,\theta)h(x) \quad \text{and} \quad \frac{\partial \ln p(x,\theta)}{\partial \theta} = \frac{\partial \ln g(u,\theta)}{\partial \theta} \quad \text{hence} \quad \widehat{\theta} = f(u)$$

- If $\widehat{\theta}$ is the ML estimator of θ, $f(\widehat{\theta})$ is the ML estimator of $f(\theta)$.
- the ML estimate is asymptotically efficient, *i.e.*

$$V[\widehat{\theta}_n] \to \frac{1}{I_n(\theta)}$$

- For the exponential family, the ML estimates are equal to the moment estimates.

Dangers of data fitting

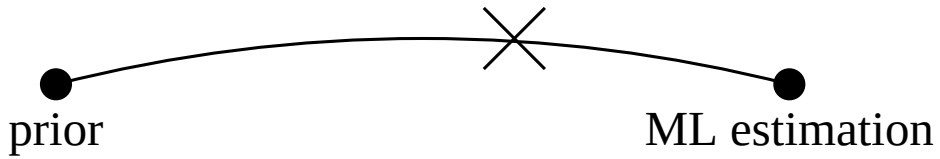


Maximum a posteriori

- The ML estimator might lead to bad solutions: *e.g.*, very small variances for Gaussians when the amount of training data is small
- The maximum a posteriori (MAP) estimator is given by

$$\widehat{\theta} = \arg\max_{\theta} p(\theta|x) = \arg\max_{\theta} p(x|\theta)p(\theta)$$

- The MAP estimator acts as a regularized ML estimator.



Other approaches

Some other fancy criteria for parameter estimation:

- (Bayesian) information criterion [BIC]
- minimum classification errors
  - explicit minimization [MCE]
  - neural networks based estimation
- maximum entropy (models) [Maxent]
  - choose the model with the largest entropy possible!
- maximum mutual information [MMI]
- etc.

Other approaches: examples

- Bayesian information criterion

$$\mathrm{BIC}(x_1^n, \theta) = \ln p(x_1^n; \theta_{\mathsf{ML}}) - \frac{1}{2} \#\theta \ln(n)$$

- minimum classification error

$$\begin{aligned} d(i) &= -\ln p(x_i, c_i; \theta) + \ln \left(\frac{1}{N} \sum_{j \neq i} \exp(\eta \ln p(x_i, c_j; \theta)) \right)^{\eta} \\ e(i) &= \frac{1}{1 + \exp(-\alpha d(l) + \beta)} \end{aligned}$$

$\Rightarrow$ minimize $\sum_i e(i)$ using a GPD algorithm.

Other approaches: examples

MCE algorithm:

1. initialize parameter θ_0
2. while not converged
   (a) compute log-likelihoods $p(x_i, c_i; \theta_k)$
   (b) update θ_{k+1} so as to maximize $\sum_i e_i$

What's the goal?

The goal can be

1. give the best model you can on a training set
2. give the expected performance of a model obtained by empirical risk minimization given a training set
3. give the best model and its expected performance

   - if the goal is (1) $\longrightarrow$ model selection
   - if the goal is (2) $\longrightarrow$ risk estimation
   - if the goal is (2) $\longrightarrow$ both!

Two popular protocoles are used either for model selection methods, namely validation and cross-validation.

Validation methodology

Principle: divide the data into two separate sets

- For **risk estimation**
  - training set = used for model selection (eventually with divided into training/validation sets)
  - test set = compute the empirical risk as an estimation of the risk
- For **model selection**
  - training set = estimate the parameters with some given hyper-parameters
  - validation set = estimate the empirical risk for the model obtained on the training set
  - select the hyper-parameters with the best empirical risk on the validation set
  - estimate the model parameters on the complete data set for the optimal hyper-parameters (optionnal)

  $\Rightarrow$ **the risk on the validation set is a very bad estimator of the risk!**

Cross-validation methodology

Principle: divide the data into N separate sets D_n

- For **risk estimation**
  - foreach seg D_i
    - model selection using the training set $\{D_{j\neq i}\}$
    - estimate the risk with the empirical risk on the test set D_i
  - average the risk estimators
- For **model selection**
  - foreach set D_i
    - select the best model on the training set $\{D_{j\neq i}\}$ for some given hyper-parameters
    - compute the empirical risk on the validation set D_i
  - select the hyper-parameters with the best average empirical risk over all the validation sets
  - estimate the model parameters on the complete data set given the optimal parameters