

Application of Artificial Intelligence

Opportunities and limitations through life & Earth sciences examples

Clovis Galiez



Grenoble

Statistiques pour les sciences du Vivant et de l'Homme

March 27, 2019

Goal

- Discover and practice machine learning (ML) techniques
 - Linear regression
 - Logistic regression
 - Neural networks
- Experiment some limitations
 - Curse of dimensionality
 - Hidden overfitting
 - Sampling bias
- Towards autonomy with ML techniques
 - Design experiments
 - Organize the data
 - Evaluate performances

Today's outline

- AI? What for?
- Glance on the applications in these series
 - Microbiome and metagenomics
 - Glacier melting prediction
- Curse of dimensionality
- Regularization

AI? What is it? What for?

Scope of these series: machine learning



80's expert systems

Modern artificial intelligence is mainly based on data science.

We will focus on the *data science* part of artificial intelligence :
machine learning.

Some machine learning methods

What machine learning tool you already know?

Some machine learning methods

What machine learning tool you already know?

For classification tasks:

- Linear Discriminant Analysis (LDA)
- Logistic regression
- Support Vector Machine (SVM)
- Artificial neural networks

For regression tasks:

- Linear regression
- Regressive artificial neural networks

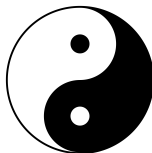
Controversies

In the media:

- AI solve all problems: ecology, unemployment, etc.
- AI is dangerous: big data is watching you

In the scientific community:

- AI solves everything: you can predict anything if you have the data
- AI does not explain anything: let's stick back to the usual models



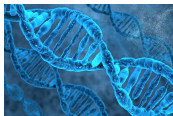
What is the right place of AI?

Studying biological function through DNA information

From an organism to its **genome**...



Organism



DNA



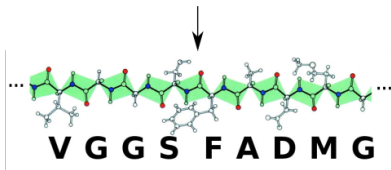
5kbp - 5Mbp

Bioinformatics: from genome to function

ACGATGTATTCAGCGATTACGATAAAGCTACGTAGTGGCA

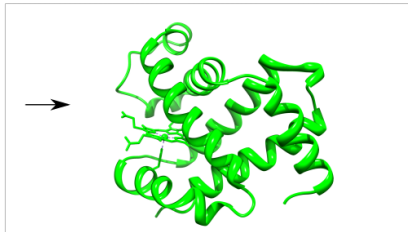
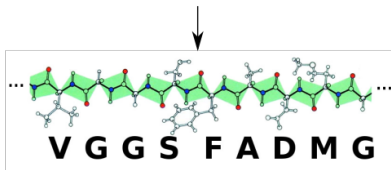
Bioinformatics: from genome to function

ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



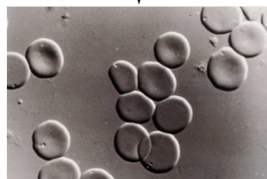
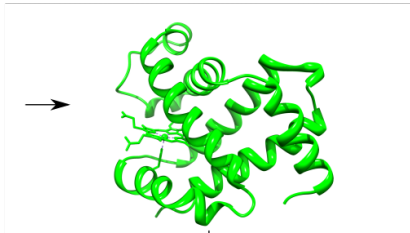
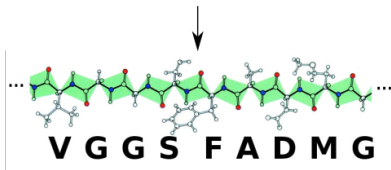
Bioinformatics: from genome to function

ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



Bioinformatics: from genome to function

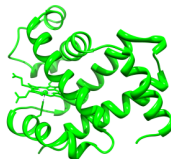
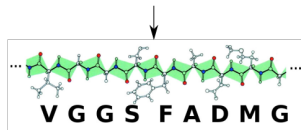
ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA



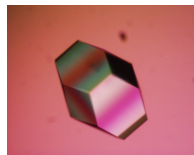
O₂ transport

Protein structure prediction

(cheap)
↓
ACG**ATGTATTCAGCGATTACGATAAAGCTACGTAGT**GGCA

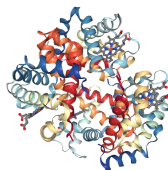


← (expensive)

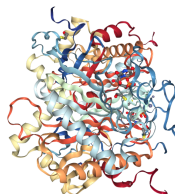


Predict the structure from sequence

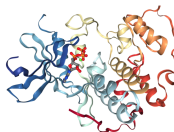
```
>1A3N:A|PDBID|CHAIN|SEQUENCE
VLSPADKTRNVKAAMGKVGGAHAGEYGAELER
MFLSPFTTXYFPHFDLSHGSAQVKGHGKVV
ADALTNAAVHVDDMPNALGALSDLHAHKLRV
DPVNFKLLSHCLLVTLAAHLPAEFTPAVHAS
LDKFLASVSTVLTSKYR
```



```
>1HXP:A|PDBID|CHAIN|SEQUENCE
MTQFNVPVDHPHRRYNPLTGQWILVSPHRAKRPW
EGAQETPAKQVLPAHDPCFLCAGNVRVTGDKN
PDYTGTYVPTNDPAALMSDTPDAESHDPIMRC
QSARGTSRVICFSPDHSKTLPELSVAALTEIVK
TWQEGTAEIGKTYPMVQVFENKGAAMGCSNPMP
HQIWMANSFLPNEAEREDRLQKEYFAHQKSPML
VDYVQRELADGSRVTVEIHLAVVPYWAANPF
ETLLLPKAHVLRITDLDQSRDLALAKKLTLS
RYDNLFCQCSFFPYSMGWHGAPFNGEENQHWQLHA
HFYFPFLRSATVRKFMVGYEMLAETQRDLTAEQ
AAERLRAVSDIHFPRESGV
```



```
>1HCK:A|PDBID|CHAIN|SEQUENCE
MENPQKVEKIGEGTYGVVYKARNKLTGEVVAL
KKIRLDTETEGVPSTAIRESLLKELNHPNIV
KLGDVIHTENKLYLVFEFLHQDLKKFMDASAL
TGIPFLPKSYLQQLGLAFCHSHRVLRHDL
KPNQLINTEGAIKLADPGLARAFVGPVRYTYT
HEVVTLMYRAPEILLGCKYYSTAVDIWSLGC
FAEMVTRRALFPDSEIDQLFRIFRTLGTDPDE
VWVPGVTSMPDYKPSFPKWARQDFSKVPPLD
EDGRSLLSQMLHYDPNKRISAKAALAHPPFDQ
VTKPVPHLRL
```



CASP competition

Blind competition. Simple principle:

- a sequence is given
- have to predict the structure.

CASP competition

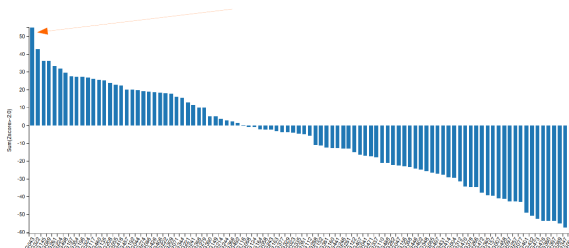
Blind competition. Simple principle:

- a sequence is given
- have to predict the structure.

13th CASP...

... AI wins !

Google's DeepMind



Problem not closed!

The negative view:

- We don't understand more what is the process of folding
- It's stupid
- It's very costly

Problem not closed!

The negative view:

- We don't understand more what is the process of folding
- It's stupid
- It's very costly

The positive view:

- It tells people that some information is still unexploited by other models
- Can be readily useful for applied science/technology

One has to see the right place of AI:

- Powerful **tool** to extract information from data
- Won't explain everything for you
- Will push forward further developments

Be thoughtful

Job offer

Looking for a young and highly motivated data scientist/engineer for optimization of marketing campaign in social network media.

Required skills:

- Machine learning methods
- Statistics
- Programming

Salary: 30k-50k depending on experience.

Be thoughtful

Job offer

Looking for a young and highly motivated data scientist/engineer for optimization of marketing campaign in social network media.

Required skills:

- Machine learning methods
- Statistics
- Programming

Salary: 30k-50k depending on experience.

Should you apply?

What would be the best reward in your life

You are good.

What would be the best reward in your life

You are good.

Choose where to put your energy to move forward the society.

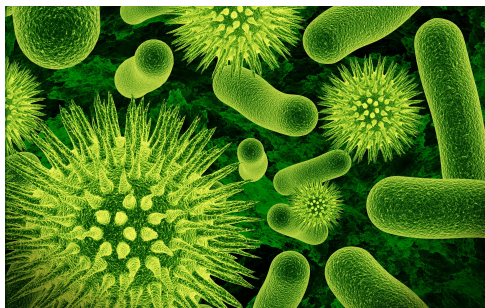
Machine learning for microbial bioinformatics

The microbial world

They are everywhere... they work hard 24h a day... they fight against each other... and they collaborate.

The microbial world

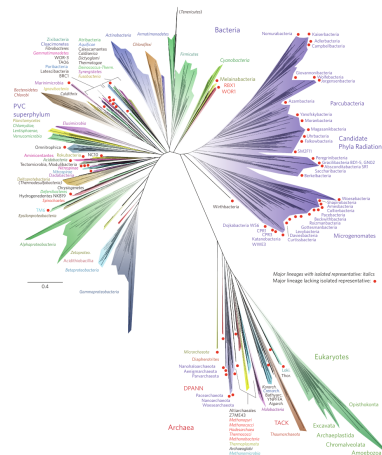
They are everywhere... they work hard 24h a day... they fight against each other... and they collaborate.



Bacteria, viruses, phages, picoeukaryotes, etc.

Origins and evolution of micro-organisms

Not a fixed knowledge: we continue to discover new branches of life.



[Hug et al. 2016]

Some facts are known, but the deep origins are still debated.

Microbiome importance in biogeochemical cycles



Nitrogen cycle [Canfield et al., Science 2010]

CO₂ turnover: viruses kill 20% of the living biomass in the ocean every day! [Suttle, Nat. Microbiol. 2007]



Microbiome importance in human health

The bright side:



Health status highly correlated with the diversity of the gut microbiome [Valdes et al. 2018]

Germany: Ten die from E.coli-infected cucumbers

🕒 28 May 2011



The death toll in Germany from an outbreak of E.coli caused by infected cucumbers has risen to at least 10.

The cucumbers, believed to have been imported from Spain, were contaminated with E.coli which left people ill with hemolytic-uremic syndrome (HUS).

Hundreds of people are said to have fallen sick.



It is unclear whether the cucumbers were infected at source or in transit

The dark side:

[Karch et al. EMBO Mol. Med. 2012]

The human gut microbiome

2000's
Human genome



\approx 20k protein-coding genes

2010's
Gut metagenomes



The human gut microbiome

2000's
Human genome



$\approx 20\text{k}$ protein-coding genes $\xrightarrow{\times 100}$ $\approx 2\text{M}$ protein-coding genes

2010's
Gut metagenomes



Human gut microbiome is rich!

Gut microbiota and higher order diseases

- **Autism**
spectrum disorder (ASD), but the underlying mechanisms are unknown. Many studies have shown alterations in the composition of the fecal flora and metabolic products of the gut microbiome in patients with ASD. The gut microbiota influences brain development and behaviors through the neuroendocrine, neuroimmune and autonomic nervous systems. In addition, an abnormal gut microbiota is associated with several diseases, [Li et al. *Front. in Cell. Neur.* 2017]
- **Type II diabetes** (50 microbial genes → AUC ROC 0.81)
[Qin et al. *Nature* 2012]
- **Parkinson's differential abundance of gut microbial species**
[Heintz-Buschart et al. *Mov. Disord.* 2018]

Gut microbiota and higher order diseases

- **Autism**
spectrum disorder (ASD), but the underlying mechanisms are unknown. Many studies have shown alterations in the composition of the fecal flora and metabolic products of the gut microbiome in patients with ASD. The gut microbiota influences brain development and behaviors through the neuroendocrine, neuroimmune and autonomic nervous systems. In addition, an abnormal gut microbiota is associated with several diseases, [Li et al. *Front. in Cell. Neur.* 2017]
- Type II diabetes (50 microbial genes \rightarrow AUC ROC 0.81)
[Qin et al. *Nature* 2012]
- Parkinson's differential abundance of gut microbial species
[Heintz-Buschart et al. *Mov. Disord.* 2018]

Can we associate the presence of microbes to a phenotype?

You may ask yourself

What all of this has to do with machine learning?!

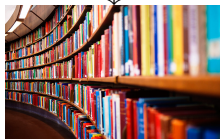
Metagenomics: the (very) big picture

sample



sequencing
→
metagenomic

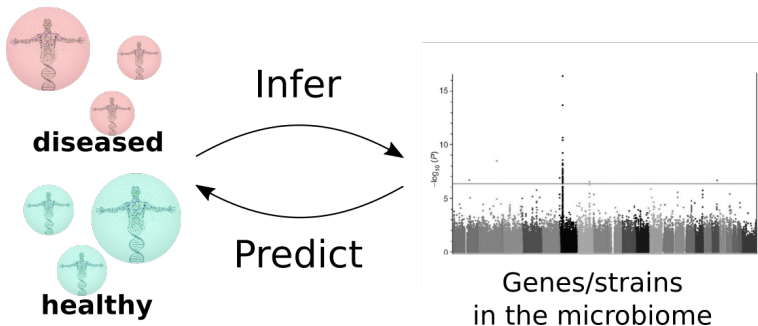
metagenome



catalog of species

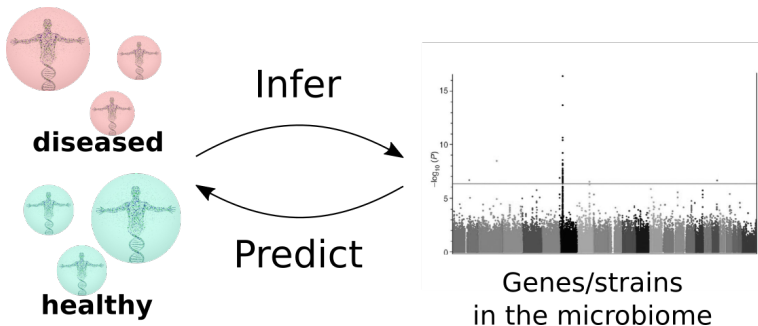
MWAS: metagenome-wide association studies

Relates the variation of the microbiome to the phenotype.



MWAS: metagenome-wide association studies

Relates the variation of the microbiome to the phenotype.



Seance 2

Can we diagnosis **Inflammatory Bowel Disease** through the structure of the gut microbial community?

Techniques involved: logistic regression, lasso regularization.

MWAS as a classification problem

Let:

- \vec{X} be an M -dimensional random vector of abundance of species,
- and Z binary (0/1) random variable describing the disease state of a human.

Define a predictor $f : \mathbb{R}_+^M \rightarrow [0, 1]$ such that it minimizes a *loss* on a training set $(\vec{x}_1, z_1), \dots, (\vec{x}_N, z_N)$:

MWAS as a classification problem

Let:

- \vec{X} be an M -dimensional random vector of abundance of species,
- and Z binary (0/1) random variable describing the disease state of a human.

Define a predictor $f : \mathbb{R}_+^M \rightarrow [0, 1]$ such that it minimizes a *loss* on a training set $(\vec{x}_1, z_1), \dots, (\vec{x}_N, z_N)$:

$$\min_f - \sum_{i=1}^N z_i \cdot \log f(\vec{x}_i) + (1 - z_i) \cdot \log(1 - f(\vec{x}_i))$$

MWAS as a regression problem

Obesity has been shown to be linked to the gut microbiome.

Let:

- \vec{X} be an M -dimensional random vector of abundance of species,
- and Y a quantitative variable (e.g. BMI in \mathbb{R}_+) random variable describing the disease state of a human.

Define a predictor $f : \mathbb{R}_+^M \rightarrow \mathbb{R}$ such that it minimizes the *loss* on a training set $(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)$:

MWAS as a regression problem

Obesity has been shown to be linked to the gut microbiome.

Let:

- \vec{X} be an M -dimensional random vector of abundance of species,
- and Y a quantitative variable (e.g. BMI in \mathbb{R}_+) random variable describing the disease state of a human.

Define a predictor $f : \mathbb{R}_+^M \rightarrow \mathbb{R}$ such that it minimizes the *loss* on a training set $(\vec{x}_1, y_1), \dots, (\vec{x}_N, y_N)$:

$$\min_f \sum_{i=1}^N (f(\vec{x}_i) - y_i)^2$$

Seance 3

We will try to relate the **Body Mass Index** through the structure of the microbial community.

Techniques involved: linear regression with Lasso penalization.



AI for the prediction of evolution of glaciers

Glaciers grow and melt



AI and glacier evolution prediction

The causes influencing the evolution of glacier are complex:

- temperature 
- solar radiation 
- albedo of the glacier
- wind
- ...

Yearly mass balance can be estimated with physical models involving all these parameters.

Physical parameters are hard to get

Measuring the physical parameters can be cumbersome.



...sometimes hard to evaluate (e.g. measuring properties of the ice).

Even with the fanciest physical model (that can also be wrong), the results can't be totally accurate.

Use unbiased and easy to measure *proxy* parameters

Seance 4

We want to see if we can model glacier evolution by a handful of well-chosen and parameters that are easy to measure and that are supposed to be responsible for the main source of variance of the mass of the glacier.



features

\approx predicts



mass balance

Techniques involved: linear regression, deep neural networks, dropout regularization.

ML traps:

I. The curse of dimensionality

A model predicts unknown outcomes

≈ Definition

We will define a model as a function depending on parameters $\vec{\theta}$ and features \vec{x} describing a target variable \vec{y} .

The role of **machine learning** is

A model predicts unknown outcomes

≈ Definition

We will define a model as a function depending on parameters $\vec{\theta}$ and features \vec{x} describing a target variable \vec{y} .

The role of **machine learning** is to **infer** the parameters $\vec{\theta}$ from a **training** set $\{(\vec{x}, \vec{y})_i, i \in 1, ..N\}$ of known relations in order to have $f(\vec{x}_i) \approx \vec{y}_i$.

A model predicts unknown outcomes

≈ Definition

We will define a model as a function depending on parameters $\vec{\theta}$ and features \vec{x} describing a target variable \vec{y} .

The role of **machine learning** is to **infer** the parameters $\vec{\theta}$ from a **training** set $\{(\vec{x}, \vec{y})_i, i \in 1, ..N\}$ of known relations in order to have $f(\vec{x}_i) \approx \vec{y}_i$.

The **high hope** is that $f(\vec{x}) \approx \vec{y}$ for yet unknown \vec{x}, \vec{y} couples.

A model predicts unknown outcomes

≈ Definition

We will define a model as a function depending on parameters $\vec{\theta}$ and features \vec{x} describing a target variable \vec{y} .

The role of **machine learning** is to **infer** the parameters $\vec{\theta}$ from a **training** set $\{(\vec{x}, \vec{y})_i, i \in 1, ..N\}$ of known relations in order to have $f(\vec{x}_i) \approx \vec{y}_i$.

The **high hope** is that $f(\vec{x}) \approx \vec{y}$ for yet unknown \vec{x}, \vec{y} couples.

We can check that on a training set, but will it generalize?

Overfitting

One of the main source of overfitting can be **model hyperparametrization**.

Overfitting

One of the main source of overfitting can be **model hyperparametrization**.

Exercise

Suppose you have a model with one binary parameter θ . Given the input, how many outputs can your model describe?

Overfitting

One of the main source of overfitting can be **model hyperparametrization**.

Exercise

Suppose you have a model with one binary parameter θ . Given the input, how many outputs can your model describe?

Suppose you have a model with N binary parameters θ_i . Given the input, how many outputs can your model describe?

Overfitting

One of the main source of overfitting can be **model hyperparametrization**.

Exercise

Suppose you have a model with one binary parameter θ . Given the input, how many outputs can your model describe?

Suppose you have a model with N binary parameters θ_i . Given the input, how many outputs can your model describe?

It means that with more parameters, it is easier to get accurate predictions on the training set... But will generalize well?

Example: polynomial regression



Suppose you measure the fuel stream Y and the car speed x .
How could you simply model the dependency between x and Y ?

Example: polynomial regression



Suppose you measure the fuel stream Y and the car speed x .
How could you simply model the dependency between x and Y ?

$$Y = \beta_0 + \beta_1 \cdot x + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Example: polynomial regression



Suppose you measure the fuel stream Y and the car speed x .
How could you simply model the dependency between x and Y ?

$$Y = \sum_{i=0}^3 \beta_i x^i + \epsilon \text{ with } \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Fitting the parameters

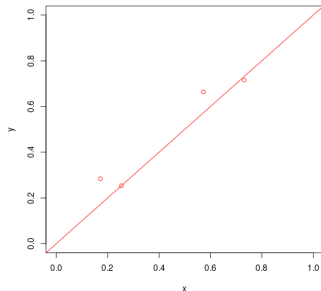


Your neighbor, gives you her home-made measurements. You, computer scientist, you fit the parameters of your model.

Fitting the parameters



Your neighbor, gives you her home-made measurements. You, computer scientist, you fit the parameters of your model.

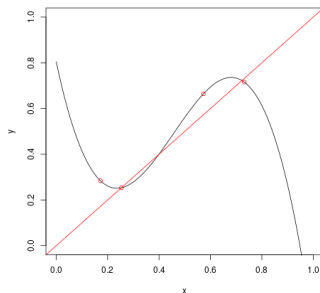


$$N = 4$$

Fitting the parameters

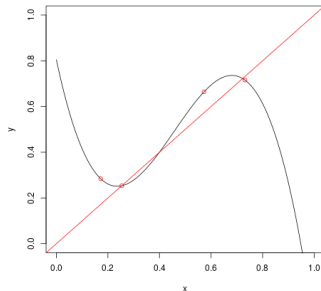


Your neighbor, gives you her home-made measurements. You, computer scientist, you fit the parameters of your model.



$$N = 4$$

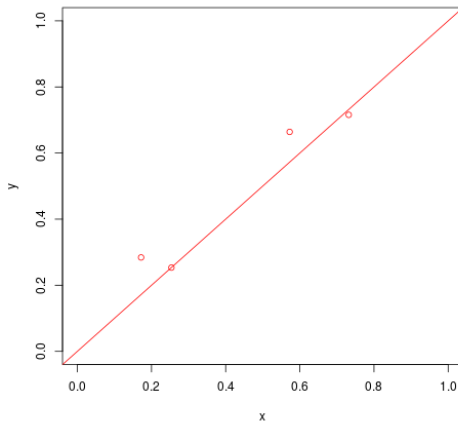
Fitting the parameters



What is the problem here? How to solve it?

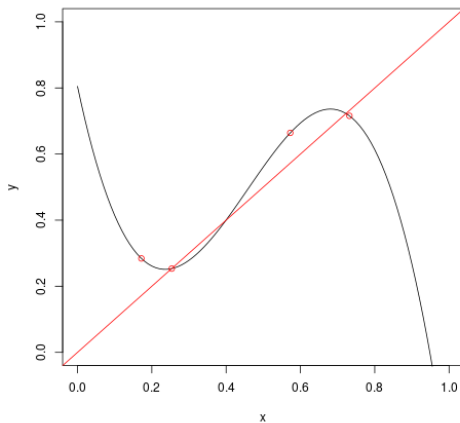
Adding more data helps!

$$N = 4$$



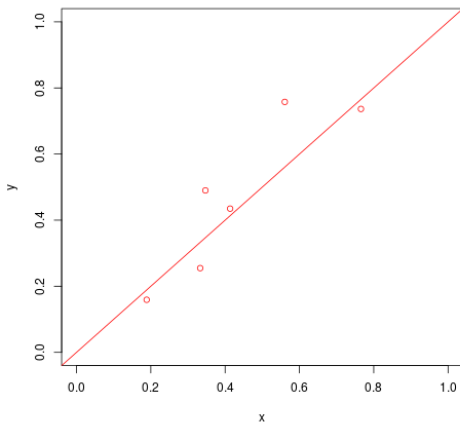
Adding more data helps!

$$N = 4$$



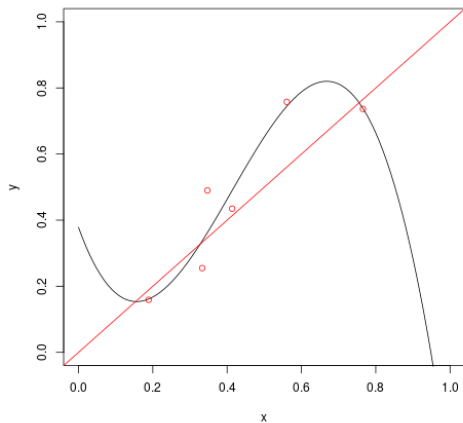
Adding more data helps!

$$N = 6$$



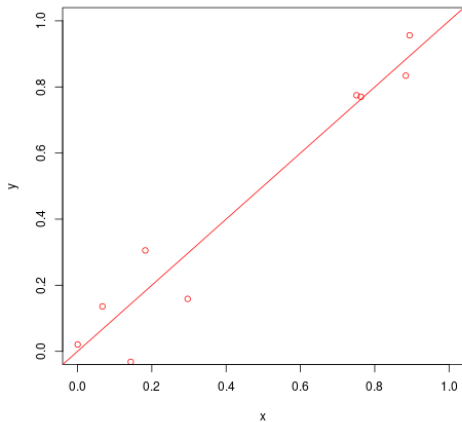
Adding more data helps!

$$N = 6$$



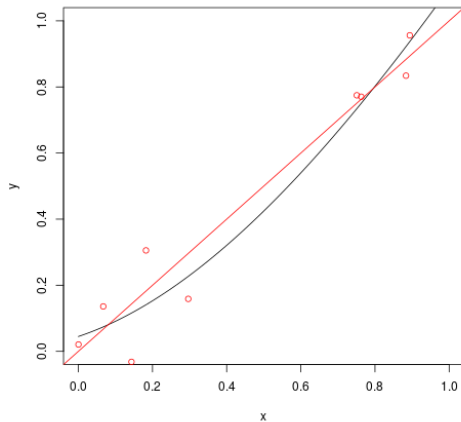
Adding more data helps!

$$N = 10$$



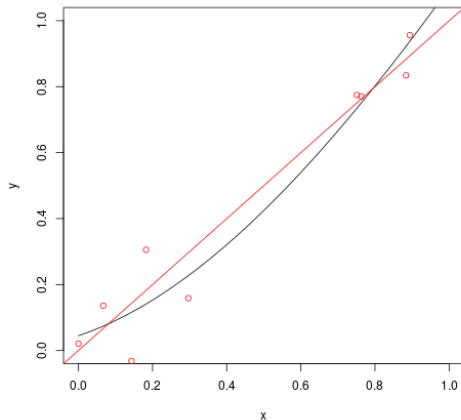
Adding more data helps!

$$N = 10$$



Adding more data helps!

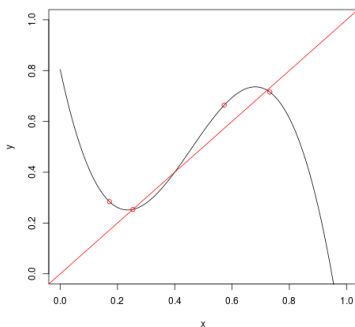
$$N = 10$$



What shall we do if we cannot get more data points?

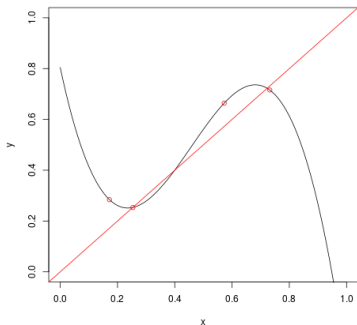
Toward regularization

What is making you deeply think that this model is wrong?



Toward regularization

What is making you deeply think that this model is wrong?



Some range of values for the parameters are unrealistic!

Regularization

The idea of regularization

Definition (well...)

Regularization is a set of methods for avoiding "unrealistic zones" in your parameter space.

Along the tutorials we will use:

- Ridge penalization (avoids high values of parameters)
- Lasso penalization (favor not using some parameters)
- Dropout (favor independence in the responsibilities of the parameters)

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.
So we can integrate *a priori* knowledge in our model :)

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.

So we can integrate *a priori* knowledge in our model :)

We consider β_0, \dots, β_3 as random variables (i.e. quantity having uncertainties).

We *model them*, for example with normal distributions centered on likely values (e.g. $\mu_0 = 0.1$, $\mu_1 = \dots$) with some likely variability (e.g. $\eta_0 = 0.005$, etc.).

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.

So we can integrate *a priori* knowledge in our model :)

We consider β_0, \dots, β_3 as random variables (i.e. quantity having uncertainties).

We *model them*, for example with normal distributions centered on likely values (e.g. $\mu_0 = 0.1$, $\mu_1 = \dots$) with some likely variability (e.g. $\eta_0 = 0.005$, etc.).

The model becomes:

$$\begin{aligned}\epsilon &\sim \mathcal{N}(0, \sigma^2) \\ \beta_i &\sim \mathcal{N}(\mu_i, \eta_i^2) \\ Y &= \sum \beta_i \cdot x^i + \epsilon\end{aligned}$$

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.

So we can integrate *a priori* knowledge in our model :)

We consider β_0, \dots, β_3 as random variables (i.e. quantity having uncertainties).

We *model them*, for example with normal distributions centered on likely values (e.g. $\mu_0 = 0.1$, $\mu_1 = \dots$) with some likely variability (e.g. $\eta_0 = 0.005$, etc.).

The model becomes:

$$\begin{aligned}\epsilon &\sim \mathcal{N}(0, \sigma^2) \\ \beta_i &\sim \mathcal{N}(\mu_i, \eta_i^2) \\ Y &= \sum \beta_i \cdot x^i + \epsilon\end{aligned}$$

What is "random" here?

Prior distributions

In the bayesian world, probabilities represent the degree of knowledge.

So we can integrate *a priori* knowledge in our model :)

We consider β_0, \dots, β_3 as random variables (i.e. quantity having uncertainties).

We *model them*, for example with normal distributions centered on likely values (e.g. $\mu_0 = 0.1$, $\mu_1 = \dots$) with some likely variability (e.g. $\eta_0 = 0.005$, etc.).

The model becomes:

$$\begin{aligned}\epsilon &\sim \mathcal{N}(0, \sigma^2) \\ \beta_i &\sim \mathcal{N}(\mu_i, \eta_i^2) \\ Y &= \sum \beta_i \cdot x^i + \epsilon\end{aligned}$$

What is "random" here?

The β_i are model **parameters** (inferred from the training data).

The μ_i and η_i are **hyperparameters** (not inferred from the training).

Worked out example

Consider a simple model:

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

$$\beta \sim \mathcal{N}(5, \eta^2)$$

$$Y = \beta x + \epsilon$$

Exercise

1. Write the likelihood of β for observing $(y_1, x_1), \dots, (y_N, x_N)$. Deduce for which β it reaches its maximum.
2. For which β is the *posterior* probability distribution $p(\beta|Y_1 = y_1, \dots, Y_N = y_N) = \frac{p(Y_1=y_1, \dots, Y_N=y_N|\beta) \cdot p(\beta)}{p(Y_1=y_1, \dots, Y_N=y_N)}$ maximal?
3. Interpret what is the effect of putting a prior distribution on the β .

Toward ridge regularization

Consider the linear model $Y = \sum \vec{\beta} \cdot \vec{x}_i + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2$$

Toward ridge regularization

Consider the linear model $Y = \sum \vec{\beta} \cdot \vec{x}_i + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2$$

2. Show that putting a Gaussian prior centered on zero on the parameters is the same as solving the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2 + \lambda \|\vec{\beta}\|_2^2$$

Toward ridge regularization

Consider the linear model $Y = \sum \vec{\beta} \cdot \vec{x}_i + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2$$

2. Show that putting a Gaussian prior centered on zero on the parameters is the same as solving the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2 + \lambda \|\vec{\beta}\|_2^2$$

This is called **ridge regularization**. What is it enforcing?

Toward ridge regularization

Consider the linear model $Y = \sum \vec{\beta} \cdot \vec{x}_i + \epsilon$.

Exercise

1. Show that the maximum likelihood solution is the same as the solution of the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2$$

2. Show that putting a Gaussian prior centered on zero on the parameters is the same as solving the following optimization problem:

$$\min_{\vec{\beta}} \sum_{i=0}^N (y_i - \vec{\beta} \cdot \vec{x}_i)^2 + \lambda \|\vec{\beta}\|_2^2$$

This is called **ridge regularization**. What is it enforcing?
It tells the model **to avoid high values** for the parameters.

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe ? outputs.

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Having a model with N parameters θ_i that live in $\{1, \dots, K\}$. Given an input, the model can describe ? outputs.

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Having a model with N parameters θ_i that live in $\{1, \dots, K\}$. Given an input, the model can describe K^N outputs.

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Having a model with N parameters θ_i that live in $\{1, \dots, K\}$. Given an input, the model can describe K^N outputs.

How would you measure that for continuous parameters?

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Having a model with N parameters θ_i that live in $\{1, \dots, K\}$. Given an input, the model can describe K^N outputs.

How would you measure that for continuous parameters?

With the volume:

$$V_N(r) = K_N \cdot r^N$$

Further justification of ridge regularization

Having a model with N binary parameters θ_i . Given an input, the model can describe 2^N outputs.

Having a model with N parameters θ_i that live in $\{1, \dots, K\}$. Given an input, the model can describe K^N outputs.

How would you measure that for continuous parameters?

With the volume:

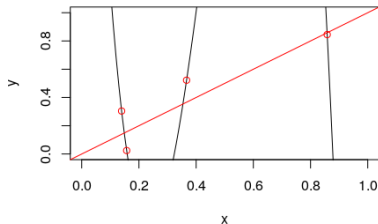
$$V_N(r) = K_N \cdot r^N \xrightarrow{r \rightarrow \infty} \infty$$

Thus, there are "more" possible model outputs when parameters have high values.

Ridge regularization example

Let's come back to the model $Y = \sum_{i=0}^3 \beta_i x^i + \epsilon$.

The maximum likelihood with 4 points will give a $\vec{\beta}$ fitting perfectly the points:



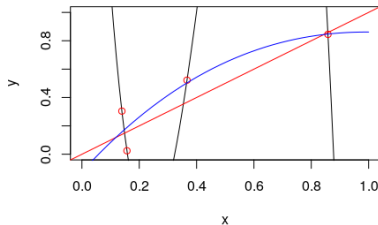
Coefficients:

| x_0 | x_1 | x_2 | x_3 |
|-------|---------|---------|----------|
| 5.169 | -54.388 | 155.755 | -114.487 |

Ridge regularization example

Let's come back to the model $Y = \sum_{i=0}^3 \beta_i x^i + \epsilon$.

With a prior $\mathcal{N}(0, \eta^2)$ the maximum a posteriori of the vector $\vec{\beta}$ corresponds to (blue curve):



Coefficients:

| x0 | x1 | x2 | x3 |
|---------|--------|---------|--------|
| -0.1279 | 2.2561 | -1.5779 | 0.3180 |

From ridge to lasso

Suppose you model a variable Y depending on some explanatory variables x with a linear model:

$$Y = \beta_0 + \sum_{i=1}^N \beta_i . x_i + \epsilon$$

Imagine now that you know that actually few variables actually explain your target variable.

Exercise

Gaussian priors on β_i centered on 0 avoid high values of β_i . Will it push the non-explanatory variables down to 0?

From ridge to lasso

Suppose you model a variable Y depending on some explanatory variables x with a linear model:

$$Y = \beta_0 + \sum_{i=1}^N \beta_i . x_i + \epsilon$$

Imagine now that you know that actually few variables actually explain your target variable.

Exercise

Gaussian priors on β_i centered on 0 avoid high values of β_i . Will it push the non-explanatory variables down to 0?

- Think individually (5')
- Vote

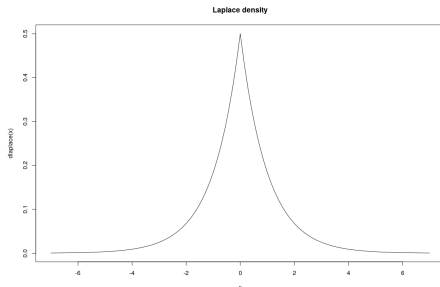
Lasso penalization

What should be the shape around 0 of the prior distribution if we want to use less parameters?

Lasso penalization

What should be the shape around 0 of the prior distribution if we want to use less parameters?

Something like:



$$f(x) = \frac{1}{2}\lambda e^{-\lambda|x|}$$

Exercise

Work out the formula to see what the model will minimize.

Overfitting depends on:

- Size of the training set
- Complexity of the problem
- The parametrization of the model
- The type of the model

See you next week to work with
real-world data!

Let's put manually some information!

Consider the following problem: a car can drive Y km with x liters of fuel. Since there are some variability in the consumption (temperature, uphill, etc.) we model it by:

Let's put manually some information!

Consider the following problem: a car can drive Y km with x liters of fuel. Since there are some variability in the consumption (temperature, uphill, etc.) we model it by:

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

Let's put manually some information!

Consider the following problem: a car can drive Y km with x liters of fuel. Since there are some variability in the consumption (temperature, uphill, etc.) we model it by:

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

What represents β ?

Let's put manually some information!

Consider the following problem: a car can drive Y km with x liters of fuel. Since there are some variability in the consumption (temperature, uphill, etc.) we model it by:

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

What represents β ?

Your neighbor tells you that $\beta = 5000$ according to its linear regression on its own home-cooked data.

What would you think ?

Let's put manually some information!

Consider the following problem: a car can drive Y km with x liters of fuel. Since there are some variability in the consumption (temperature, uphill, etc.) we model it by:

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

What represents β ?

Your neighbor tells you that $\beta = 5000$ according to its linear regression on its own home-cooked data.

What would you think ?

This guy is bullshitting me!

Bayesian approach

In the bayesian world, probabilities represent the degree of knowledge. What is important is not the probability of a given outcome, but its complete distribution.

Let's take an example.

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

We know that:

- $\arg \max_y p(Y = y) =$

Bayesian approach

In the bayesian world, probabilities represent the degree of knowledge. What is important is not the probability of a given outcome, but its complete distribution.

Let's take an example.

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

We know that:

- $\arg \max_y p(Y = y) = \beta x$
- and $\mathbb{E}(Y) =$

Bayesian approach

In the bayesian world, probabilities represent the degree of knowledge. What is important is not the probability of a given outcome, but its complete distribution.

Let's take an example.

$$Y = \beta x + \epsilon \text{ where } \epsilon \sim \mathcal{N}(0, \sigma^2)$$

We know that:

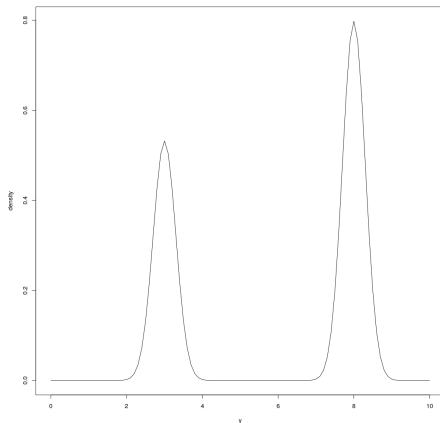
- $\arg \max_y p(Y = y) = \beta x$
- and $\mathbb{E}(Y) = \beta x$

But what may be more important is that $Y \sim \mathcal{N}(\beta x, \sigma^2)$.

Meaning that we know that Y lies roughly around βx with decreasing probability as far as it gets away from it.

Why is it important?

Importance of the full distribution



Importance of the full distribution

