Application of Artificial Intelligence

Opportunities and limitations through life & Life sciences examples

Clovis Galiez

Grenoble
Statistiques pour les sciences du Vivant et de l'Homme

March 31, 2020

Disclaimer

- You should form teams of 2 persons on Teide.
- Answer the questions in the template at `https://clovisg.github.io/teaching/asdia/ctd2/quote2.Rmd` and post-it on teide.
- You can use the following Riot channel `https://riot.ensimag.fr/#/room/#ASDIA:ensimag.fr`, I'll be present to answer live questions during the lecture slots. Do not hesitate to post your understandings and mis-understandings out of the time slots, I won't judge it, I'll only judge your involvment and curiosity.
- You can send me emails (`clovis.galiez@grenoble-inp.fr`) for specific questions, and I'll answer publicly on the riot channel.

Goals

- Have a critical understanding of the place of AI in society
- Discover and practice machine learning (ML) techniques
  - Linear regression
  - Logistic regression
- 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

- Short summary of the last lecture
- Lasso regularization
- Experiment the curse of dimensionality
- Logistic regression

Last lecture

Remember

What do you remember from last lecture?

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality

Last lecture

Remember

What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality
  - Overfit can stem from too many features (capacity of description increases exponentially)
  - More data helps

Last lecture

Remember

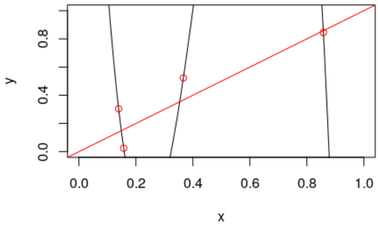
What do you remember from last lecture?

- Phantasm and opportunities of AI
- Microbiomes
  - Diverse
  - Still a lot to discover
  - Play key roles in global **geochemical cycles** and in **human health**
- Curse of dimensionality
  - Overfit can stem from too many features (capacity of description increases exponentially)
  - More data helps
  - Restricting the parameter space: regularization
    - Ridge

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:



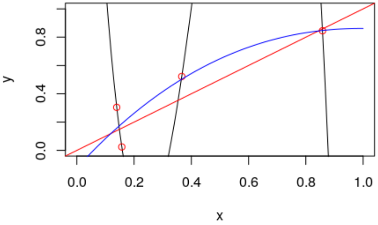
Maximum *likelihood* coefficients:

β_0	β_1	β_2	β_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):



Maximum *a posteriori* coefficients

β_0	β_1	β_2	β_3
-0.1279	2.2561	-1.5779	0.3180

Ridge regularization

Consider the linear model $Y = \vec{\beta}.\vec{X} + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2)$.

Facts

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

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2$$

2. Putting a Gaussian prior $\beta_i \sim \mathcal{N}(0, \eta^2)$ on the parameters is the same as solving the following optimization problem (ridge regularization):

$$\min_{\vec{\beta}} \sum_{i=0}^{N} (y_i - \vec{\beta}.\vec{x_i})^2 + \frac{\sigma^2}{\eta^2} ||\vec{\beta}||_2^2$$

3. It tells the model **to avoid high values** for the parameters. It is equivalent to introduce *fake* data at coordinates:

$$\vec{x} = (\frac{\sigma}{\eta}, \frac{\sigma}{\eta}, ..., \frac{\sigma}{\eta}), y = 0$$

From ridge to lasso

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

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

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

Question!

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

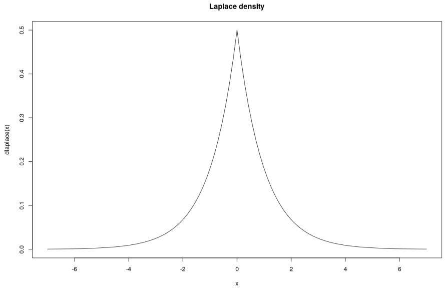
- Think individually (5')
- Vote

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 criterion is minimized when maximizing the posterior probability of the parameters.

Show that curse of dimensionality happens!

Design a simple experiment showing the curse of dimensionality in the linear regression setting.

- Individual reflexion (5')
- Then we decide on a common experimental plan

Experimental plan

- Simulate in R a dependence between a vector $\vec{X}$ and an output variable y.

Experimental plan

- Simulate in R a dependence between a vector $\vec{X}$ and an output variable y.
- Find the maximum likelihood of the parameters of a linear regression.

Experimental plan

- Simulate in R a dependence between a vector $\vec{X}$ and an output variable y.
- Find the maximum likelihood of the parameters of a linear regression.
- Add components to $\vec{X}$ that are not related to the output variable? Are the coefficients near to 0?

Experimental plan

- Simulate in R a dependence between a vector $\vec{X}$ and an output variable y.
- Find the maximum likelihood of the parameters of a linear regression.
- Add components to $\vec{X}$ that are not related to the output variable? Are the coefficients near to 0?
- Add regularization and check if the correct coefficient are recovered.

Logistic regression (classification)

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary $(0/1)$ random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \to [0,1]$ is a function chosen to minimize some *loss* in order to have

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \rightarrow [0, 1]$ is a function chosen to minimize some *loss* in order to have $f(\vec{x}) \approx z$ for realizations $\vec{x}, z$ of $\vec{X}, Z$.

Classification

Let:

- $\vec{X}$ be an M-dimensional random variable,
- and Z binary (0/1) random variable.

$\vec{X}$ and Z are linked by some unknown *joint* distribution.

A predictor $f : \mathbb{R}_+^M \to [0, 1]$ is a function chosen to minimize some *loss* in order to have $f(\vec{x}) \approx z$ for realizations $\vec{x}, z$ of $\vec{X}, Z$.

Which loss?

Logistic regression

A natural predictor is $f(\vec{x}) = p(Z = 1|\vec{x})$.

[1] This choice is theoretically sound, in particular when $\vec{x}|Z = i \sim \mathcal{N}(\vec{\mu_i}, \Sigma)$, or x_i's are discrete.

Logistic regression

A natural predictor is $f(\vec{x}) = p(Z = 1|\vec{x})$. Problem: $p(Z = 1|\vec{x})$ is unknown.

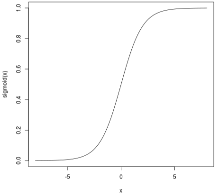
[1] This choice is theoretically sound, in particular when $\vec{x}|Z = i \sim \mathcal{N}(\vec{\mu_i}, \Sigma)$, or x_i's are discrete.

Logistic regression

A natural predictor is $f(\vec{x}) = p(Z=1|\vec{x})$. Problem: $p(Z=1|\vec{x})$ is unknown.
We **model** it by[1] :

$$f_w(\vec{x}) = \sigma(\vec{w}.\vec{x} + b)$$

where the function σ is the logistic sigmoid $\sigma : x \mapsto \frac{1}{1+e^{-x}}$



[1]This choice is theoretically sound, in particular when $\vec{x}|Z=i \sim \mathcal{N}(\vec{\mu_i}, \Sigma)$, or x_i's are discrete.

Conditional likelihood

Exercise

1. Show that it is not possible to find the parameters $\vec{w}$ by maximum likelihood if we don't know the distribution of $\vec{x}$.

2. Let $f(\vec{x}) = p(Z = 1|\vec{x}) = \sigma(\vec{w}.\vec{x} + b)$. Show that the *conditional* log-likelihood $LL = \log P(z_1, ..., z_N|\vec{x}_1, ..., \vec{x}_N, \vec{w}, b)$ writes:

$$LL(\vec{w}, b) = \sum_{i=1}^{N} [z_i. \log f(\vec{x}_i) + (1 - z_i). \log(1 - f(\vec{x}_i))]$$

3. To what well-known loss the optimization of this conditional likelihood corresponds?

4. Interpret geometrically the role of parameters $\vec{w}$ and b.

Curse of dimensionality in classification

From the previous exercise, if the k^{th} component of the feature vector $\vec{x}$ plays no role in the classification process, what should be the value of w_k?

What can you expect in practice?

> If you expect only few explanatory components in your vector of features $\vec{x}$, what shall you do?

Next week we will apply these methods on real data!