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

April 15, 2020

Today's outline

- Short summary of the last lecture
- Continue IBD experiment
- Sampling biases
    - Redundancy
    - Imbalanced data

Last lecture

Remember

What do you remember from last lecture?

Last lecture

Remember

What do you remember from last lecture?

- Logistic regression

Last lecture

Remember

What do you remember from last lecture?

- Logistic regression
- Microbiome
    - Plays a key role in human health
    - 1000's of species in one human gut

Last lecture

Remember

What do you remember from last lecture?

- Logistic regression
- Microbiome
  - Plays a key role in human health
  - 1000's of species in one human gut
- Need for regularization

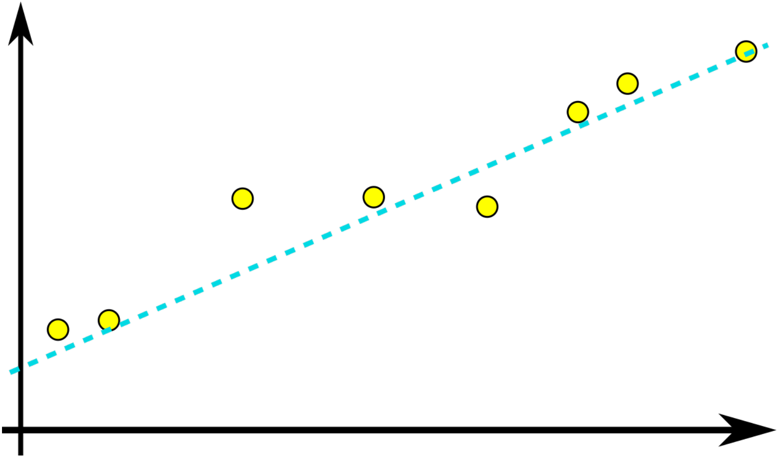
IBD experiment

Microbial species abundances have been computed for 396 individuals (148 with IBD, 248 healthy).

More than 1000's of species.

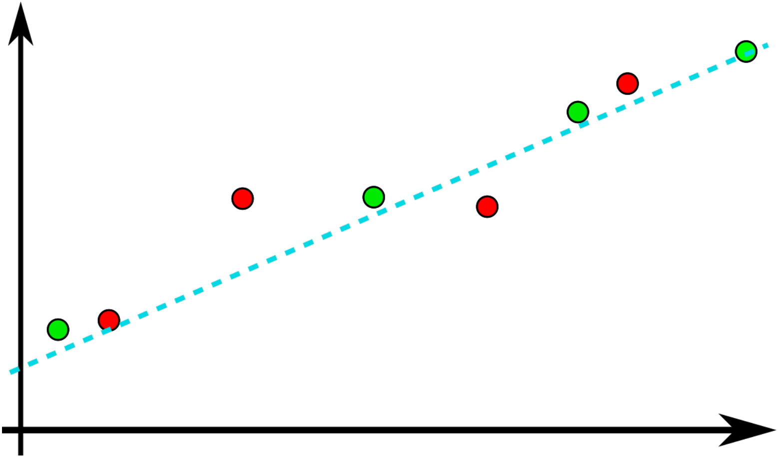
Hidden overfitting

Overfitting and cross-validation



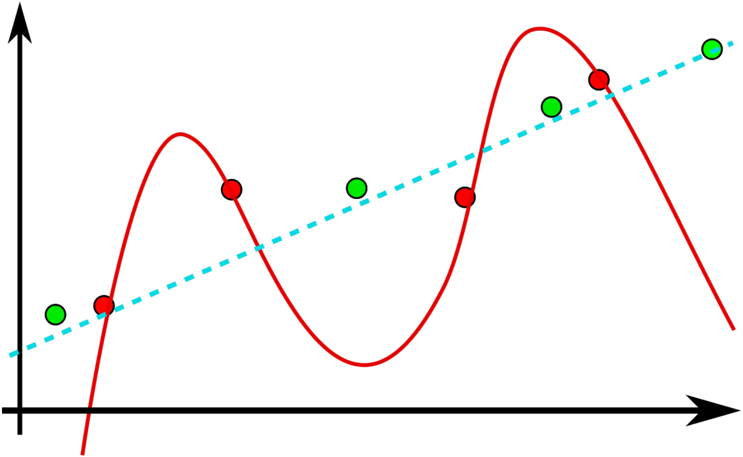
Data linearly distributed

Overfitting and cross-validation



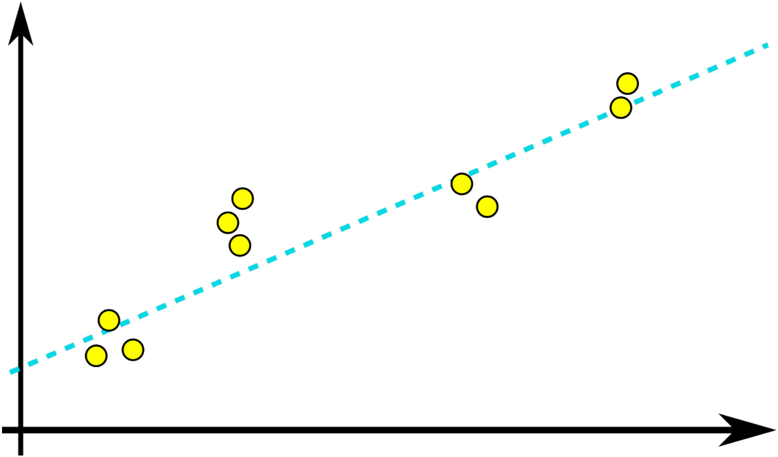
We randomly choose a training set

Overfitting and cross-validation



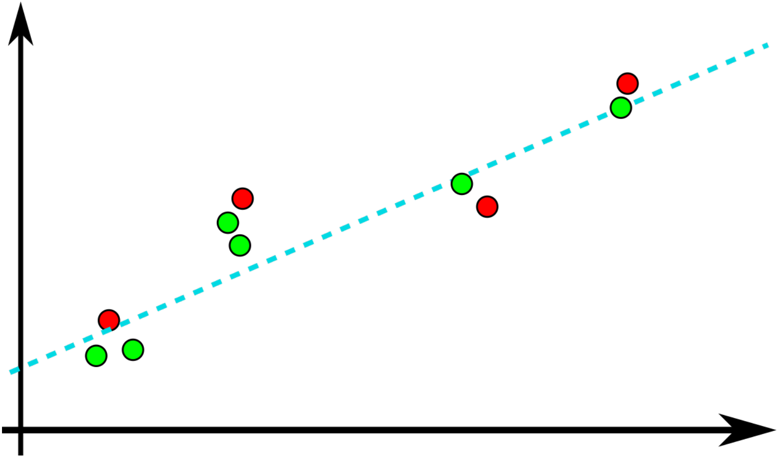
After training, we can measure the overfit on test set

Overfitting and cross-validation



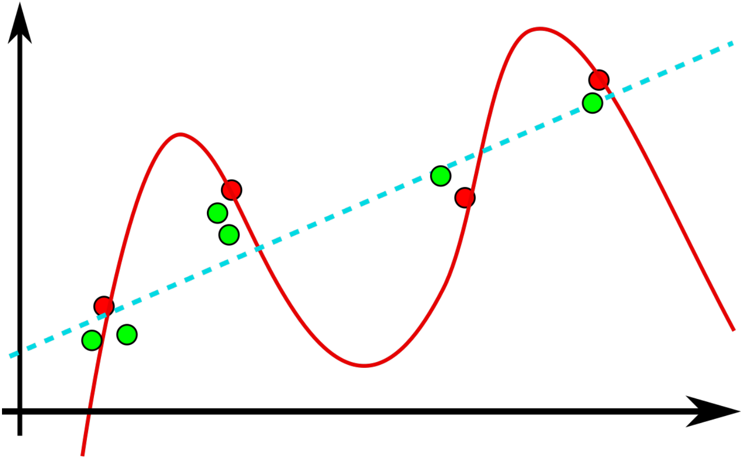
Redundant data linearly distributed

Overfitting and cross-validation



We randomly choose a training set

Overfitting and cross-validation



Hidden overfit here!

Redundancy in datasets

Cross-validation is a method (supposedly) providing a way to optimize parameters so that the model **generalizes** as much as possible.

Exercise

Design an experiment proving experimentally that cross-validation can have good performances across folds, but poor generalization/real poor performance.

Propose and implement a method reducing this effect.

Imbalanced data

Imbalanced dataset/sampling

Model and data

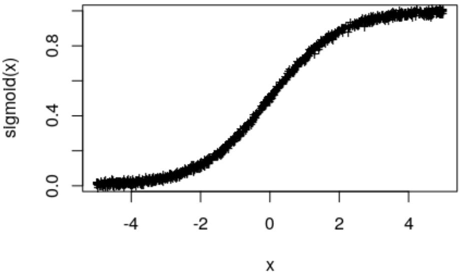
 ,  ,... $\xrightarrow{\approx\text{predicts}}$ 

features mass balance

Goal: predict the future melt

Glacier melting as a function of temperature

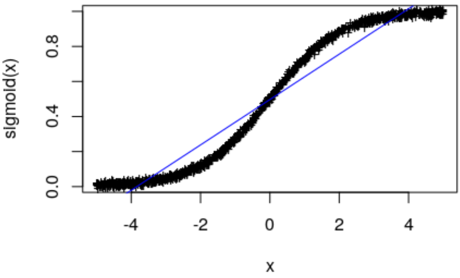
Consider that the response to temperature (x-axis) of melting of a glacier (y-axis) take the following form :



(saturation of the melting speed at high temperatures)

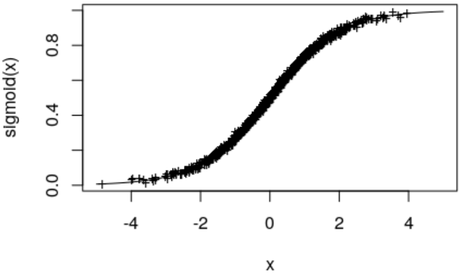
Linear modeling of the melt

Optimizing the MSE of a linear model should have the following form (blue line):



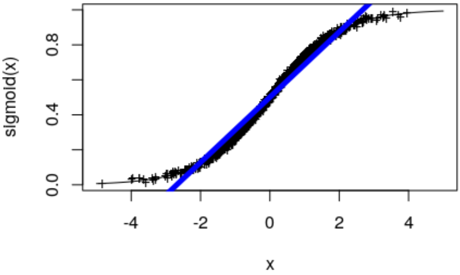
Sampling bias

But in reality, we seldom observe the extreme values, so that the data points are distributed as follow:



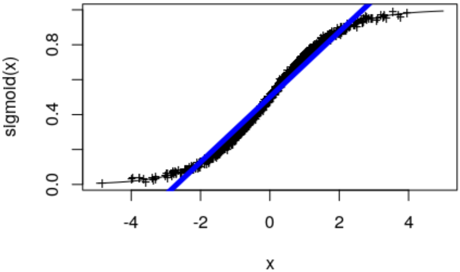
Effect of sampling bias on the model

So that when fitting a linear model, we get the following regression:



Effect of sampling bias on the model

So that when fitting a linear model, we get the following regression:



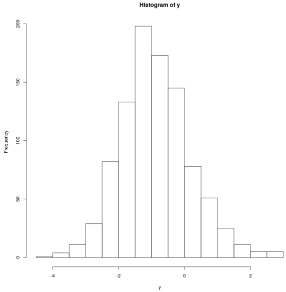
... so that computing the real MSE (with even sampling of the temperature range) is around 0.03.

Skewed marginal distribution

The loss is computed on **average** on the dataset:

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

Distribution of the y_i's:

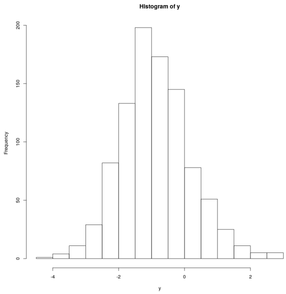


Skewed marginal distribution

The loss is computed on **average** on the dataset:

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

Distribution of the y_i's:



What could be an issue here?

Dealing with imbalanced data

Exercise

1. In a (linear) regression setting, design an experiment to prove empirically that imbalanced data can be a problem.

2. How could you change the following loss function in order to reduce the effect of the imbalance?

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

3. Look up the options of the `lm` R command that implements the solution you have found in 2. and show that you can reduce the impact of imbalance.

Weight the data

One trick is to give the data samples with a weight that is inversely proportional to the density. We want to optimize the following loss:

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

where w_i give corrects for the sampling density.

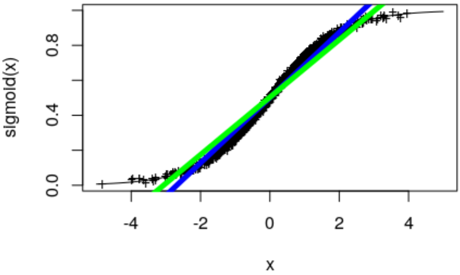
```
# estimate the density function
sampleDensity = density(data$y)

# compute the weights at data points
w = 1/approx(sampleDensity$x,sampleDensity$y,data$y)$y

# fit using the density the weights
linRegCorrected = lm(y~x,data,weights=w)
```

Corrected model

With weights, we get the green model:



Does not look a huge improvement, but **reduces the MSE[1] by a half** (0.015 instead of 0.03)!

[1] MSE computed with evenly distributed temperature over the whole range

Imbalanced data is common!

This effect actually applies to many cases, in particular with classification tasks (imbalance of 0 and 1 labels).

Beware!

Hope you've learned some stuff during those lectures!