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 17, 2019

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

Last lecture

Remember

What do you remember from last lecture?

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

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 dataset/sampling

Model and data

🌡, ☀, … $\xrightarrow{\approx\text{predicts}}$

features

mass balance

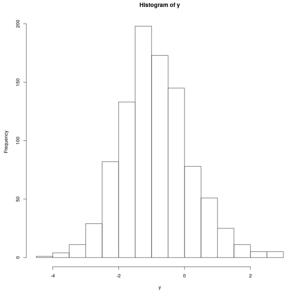
Goal: predict the future melt

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:

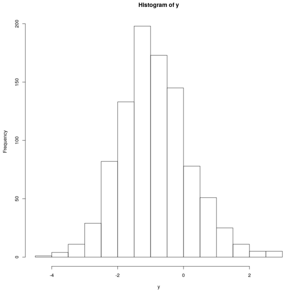


Skewed marginal distribution

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.

Hope you've learned some stuff during those lectures!