Data Science Methods

Methods

We have extensive experience in the following AI, Pattern Recognition & Machine Learning methods. We use both supervised and unsupervised methods:

- Classification / Clustering
  - Difference Between Classification and Clustering
  - “A Comparative Study on Clustering and Classification Algorithms“, Jyotismita Goswami, June 2015
- Decision Tree Learning
  - Decision Trees – Lior Rokach, TAU
- Genetic Algorithms
- Graphs
- Lightweight Ontologies
- Nearest Neighbors
- Neural Networks
- Optimization (multidimensional functional minimization, quadratic forms)
- Random Forests

Logistic regression

It is a binomial probabilistic method, also called a binomial model method. It was the first method used in classification, especially in marketing for scoring or in epidemiology; it involves modeling a binary binomial variable (number of successes for n trials) or Bernouilli (with n = 1). It is used to describe the possession or not of a product, the description of a good or bad customer.

Partial Least Square

PLS regression is an old method widely used, especially in chemometrics in the food industry, in discrete spectral data analysis, so in large data, when the number of observations is less than the number of explanatory variables. Several variations are available:

- PLS1 variable Y-variable quantitative regression to be modeled by p X quantitative explanatory variables; it is a question of looking for a model of regression on orthogonal components constructed starting from the p variables of X centered;
- PLS2 regression: set of q target variables Y to model from a set of p quantitative explanatory variables; the goal here is to optimize a sum of squares of covariances between an orthogonal component and each of the response variables.

Support Vector Machines (SVM)

Support Vector Machines are a class of learning algorithms initially defined for the prediction of a binary qualitative variable. They were then generalized to the forecast of a quantitative variable. The principle is to look for the optimal margin hyperplane which should separate the data while being as far as possible from all the observations: therefore, the classifier must have the greatest possible generalization capacity.

Neural networks

A neural network is the association into a graph of elementary objects. The networks can be described by the architecture of the graph (the layers …), the level of complexity (number of neurons), the type of neurons (activation or transition functions) and by the goal: supervised learning or not . It is

therefore, in the case of supervised learning, by minimizing the quadratic loss, of estimating the parameters of the coefficients of the inputs of the neurons of all the layers, as well as the parameters of the activation function which calculates the output network.

Random forest and Ensemble methods

This is the aggregation of a collection of random decision trees. It’s a family of overall methods. The construction of a binary discrimination tree consists in determining a sequence of nodes, defined by the joint choice of a variable among the explanatory variables and a partition into two classes from this variable. The interest of the set-based methods is to generate several rules of prediction and then to put in common their different answers (by majority vote in the case of classification, or in the average in the case of a regression).

Adaptive Gradient or Gradient boosting

It is a family of algorithms based on the optimization of a loss function that is supposed to be differentiable; a sequence of models must be built and, at each stage, each added model must appear as a step towards a better solution. This step is taken in the direction of the gradient of the loss function.

Industries

We have applied our Machine Learning knowledge to the following industries

- Aerospace (turbulence prediction for airlines)

- Finance (algorithmic trading)
- Insurance (fraud detection)
- Marine vessel route monitoring (anomaly detection for vessels trajectory)
- Search (personalized search via Graph methods, search engine)
- Website Monitoring (courbe de charge pour des sites B2B)

General Tools and Frameworks

- Apache Giraph (iterative graph processing system built for high scalability),
- GraphX (Apache Spark's API for graphs and graph-parallel computation) ,
- Graphlab (extensible machine learning framework),
- Flink (streaming dataflow engine that provides data distribution, communication, and fault tolerance for distributed computations over data streams)
- Gelly (Apache Flink's *graph-processing API and library,* a suitable platform for large-scale graph analytics),
- Hadoop (HDFS, Cassandra, Zookeeper, Scalding)
- @Google
    - Internal Google frameworks (GFS, MapReduce, BigTable, ProtoBuffs)
    - Internal Java based data sketching library
- JBLAS (Linear Algebra Library)
- Mahout (Distributed ML library for Hadoop)
- Powergraph (framework for large-scale machine learning and graph computation),
- Spark (fast and general engine for large-scale data processing),
- Scikit-Learn, Numpy, Scipy,
- iPython (python libraries for prototypes)
- FloydHub (Docker container with multiple ML Libraries, "All-in-one Docker image for Deep Learning)

- Apache Storm (real-time processing engine)
- TensorFlow: an open source Machine Learning software library for numerical computation using data flow graphs. Nodes in the graph represent mathematical operations, while the graph edges represent the multidimensional data arrays (tensors) communicated between them.

Overall (used) Languages

- C# / .Net
- C++
- Javascript
- Java
- Python
- Scala

Case Studies

We have applied our Machine Learning knowledge to the following industries:

- Aerospace (turbulence prediction for airlines)
- Finance (algorithmic trading)
- Insurance (fraud detection)
- Marine vessel route monitoring (anomaly detection for vessels trajectory)
- Search (personalized search via Graph methods, search engine)
- Website Monitoring (courbe de charge pour des sites B2B)

Read more in the Case Studies Section.

Partners

We work closely with two other firms specialized in Machine Learning methods:

- Hadrian Advisors (FR): They are specialized in Data Science applied to Financial markets, and Insurance problems.
- Aware Technologies (Canada): They offer unique MLaaS (Machine Learning as a Service) infrastructure and know-how, with very high end supervised and unsupervised machine learning techniques.

Thanks to our close collaboration, we are able to bring extensive man and brain power, and diversify our Data Science work with innovative techniques. We can help you assess what mathematical & AI tools best fit your needs, regardless of the industry, and then accompany you in implementing some of these methods, and develop efficient & attractive web based tools to manage them.