Structural approach to the deep learning method

Leonid A. Sevastianov[1,3] Anton L. Sevastianov[1] Edik A. Ayrjan[2] Anna V. Korolkova[1]
Dmitry S. Kulyabov[1,2] Imrikh Pokorny[4]

NEC–2019, 30 September – 4 October, 2019 Budva, Montenegro

[1]RUDN University, Moscow, Russian Federation

[2]LIT JINR, Dubna, Russian Federation [3]BLTP JINR, Dubna, Russian Federation

[4]Technical University of Košice, Košice, Slovakia

Technologies

Machine learning problems

- Machine learning and neural network technologies are overrated.
- Machine learning methods reduce the culture of analytical thinking.
- To achieve results in projects involving data analysis, knowledge of the subject is more important than deep knowledge of Machine learning.
- The profession of Data Scientist is greatly overrated, generalists are gradually disappearing.

Technology overrated

- Most of the problems that can now be solved with the help of modern methods of data analysis and neural networks have been solved for a long time.
- The tasks are essentially not new. Analysts who understand the subject area participate in their decision.
- Often, machine learning algorithms in such systems are already in place.
- To do something fundamentally new and really applicable here is extremely difficult.
- “The apples that fell from the tree are already harvested.”

Analytical thinking

- You need to deeply understand the subject area:
  - what data is needed;
  - are any predictive algorithms needed;
  - is it possible to verify the prediction.
- Requires an analytical approach.
- Requires a culture of working with data.
- Requires the ability to put hypotheses.

More important knowledge of the subject

- The disadvantages of a typical Data Scientists include:
    - almost do not ask any questions;
    - data and so will tell about everything;
    - use some arbitrary data;
    - They say that they built some kind of model.
- The result cannot be verified.

Universal specialists will no longer be

IMHO, an effective Data Scientist

- can not be a generalist;
- must be an expert in the subject area.

Data science is not rocket science

Project structure

How the data analysis project works

- Project requirements
- Project data
- Development and implementation of the project

Requirements

- We initially do not know anything about what data we have.
- We need to understand the statement of the problem.
- We must understand what result is required to get from the project.
- We must decide by what method the problem can be solved.
- We need to set data requirements.

Data

- Search for data to solve the problem:
  - we will find out what sources are available to us;
  - we form a sample with which we will continue to work.
- Data research:
  - explore the central position and variability;
  - identify correlations between signs;
  - build distribution schedules.
- Data preparation.

Development and implementation

- Model development.
- Software implementation of the model.
- Run training set.
- Testing on a test sample.
- Verification of the result.
- Loop (you can start all over again).

Requirements

Understanding of the task

- It is necessary to clearly define the purpose of the study.
- What is the problem?
- What metrics will measure success?

The choice of analytical approach

- The choice of approach depends on what type of response you need to get as a result:
  - if you need a yes / no answer, a Bayesian classifier is suitable;
  - if you need an answer in the form of a numerical sign, then regression models are suitable;
  - if it is necessary to determine the probabilities of certain outcomes, it is necessary to use a predictive model;
  - if you need to identify relationships, a descriptive approach is used.

Data requirements

- What data will give the desired answer?
- Data requirements:
  - content;
  - data formats;
  - data sources.

Data

Data collection

- We collect data from available sources.
- We make sure that the sources:
  - available;
  - reliable;
  - can be used to obtain the required data in the required quality.
- It is necessary to understand whether we received the data we wanted.
- Revision of data requirements.
- Deciding on the need for additional data.
- Finding a replacement for missing data.

Data analysis

- Are the collected data representative of the problem?
- Descriptive statistics apply to all variables that will be used in the selected model:
    - the central position is studied (middle, median, mode);
    - emissions are searched for and variability is estimated (variance, standard deviation);
    - histograms of the distribution of variables are built;
    - other visualization tools are used (for example, boxes with a mustache).

Data analysis

- Correlations between variables are calculated.
- If there are significant correlations between the variables, some variables may be discarded as redundant.

Data preparation

Data collection and analysis + data preparation = 70%–90% of the project time.

Data preparation

- We process the data in such a way that it is convenient to work with them:
  - remove duplicates;
  - process missing or incorrect data;
  - we check and correct formatting errors.
- We are designing a set of factors that machine learning will work with in the next steps:
  - feature extraction;
  - feature selection.
- Errors at this stage can be critical.
  - Excessive number of characteristics = model retrained.
  - Insufficient number of signs = model is under-trained.

Development and implementation

Model construction

When the type of model is defined and there is a training sample, we develop the model and test it on a set of features.

Model application

- Calculations alternate with model setup.
- Does the constructed model meet the original task?

Steps

- Diagnostic measurements are taken to help determine if the model works as intended.
- The statistical significance of the hypothesis is checked.
- It is necessary to make sure that the data in the model are correctly used and interpreted and the result obtained does not go beyond the limits of statistical error.

Implementation

- Implementation is carried out in stages:
  - a limited group of users;
  - test environment.
- Feedback system.