Workshop 2

Learning From Big Data

Spark ML, Scikit-Learn for ML, and Linear Regression

- Over the past ten years **Machine Learning (ML)** has conquered the industry: it is now at the heart of much in today's high-tech products, ranking your web search results, powering your smartphone's speech recognition, and recommending videos, beating the world champion at the game of Go. Before you know it, it will be driving your car."

- This is so called "*The Machine Learning Tsunami*" ,mentioned in "Hands-on Machine Learning with Scikit-Learn & Tensorflow", by Aurelien Geron , 2017.

- After discussed the concepts around Machine Learning and Linear Regression in my lecture, let us have the following Lab work with **Spark** and **Scikit-Learn**, the two platforms for developing ML based data processing and analysis.

* Please do the exercises or tasks without "**Optional**" mark at first. After that, if you still have some time, please try the tasks with "Optional".

Part 1: Working with pyspark.ml

- Before you carry out two tasks, briefly look at Spark ML APIs, particularly on pyspark.ml:

  https://spark.apache.org/docs/latest/api/python/pyspark.ml.html#

Task 1: Linear Regression

1) Given a *dataset* with ecommerce customer data for a company's website and mobile app, let us see a sample, which demonstrates a real application of ML linear regression.

   I have added helpful comments inside of the codes, presented in /Linear_Regression_Sample.ipynb, together with its dataset Ecommerce_Customers.csv.

   You may find them in directory …/Labs/Lab2/.

   Actions:

   - Study its codes with comments, while you run cell one by one in the notebook and see what happens.

2) The specification of a project for learning "ML Linear Regression modeling and applications", Linear_Regression_Project.ipynb, is presented in the folder…/Labs/lab2.

- Review this specification. Try to understand the project's goal and think about how to achieve it.

- As you are a beginner of this subject, it's hard to develop a solution by yourself at this stage. So, you may look at the project's solution, given in

  Linear_Regression_Project_SOLUTION.ipynb,

  following my instructions to study its implementation.

Part 2: Working with Scikit-Learn

1) **Scikit-Learn** is an excellent tool for Machine Learning based projects in Python, which provides various modules and functions for data scientists. Briefly have a look at this link, http://scikit-learn.org/stable/index.html, and get a big picture about this tool.

2) Besides ***Pandas***, ***NumBy*** and ***Matplotlib*** are also popular in use in Python programming for data scientists.

3) Let us have a tutorial for Matplotlib, by using tools_matplotlib.ipynb, that is given in …/Labs/Lab2/Matplotlib directory. Simply follow the notebook to learn its basics.

   [Check if you have installed the three tools in your system].

4) As our working time is limited, you might not finish the learning and practice on **Pandas** tutorial, given in the …/Labs/Lab1. Please continue to study them, simply following the two notebooks:

   …Lab_1/Pandas_and_NumPy/tools_pandas.ipynb

5) Working on data processing and analytics, you have to frequently deal with non-convenient data that is non-numerical, such as customer names, or zipcodes, country names, etc... So, you have to transform them. ***Panda*** plays an important role in this data processing or transformation.

Part 3: Exercises (Optional)

- **If** you still have a time in this Lab class, you may have the following exercises:

  i) Read a PDF text, named as "An Example for Model_based_ML" in …/Labs/Lab2/Model_based_ML, and understand what this example is for.

  ii) Have a look at the notebook, Model_based_ML_example.ipynb, located in the same directory. This notebook presents a detailed coding work for implementing this project in Python with Scitik-Learn.

References:

The following references **are optional.** If you would like to have a review for Linear Algebra and Linear Regression **in detailed math**, they would be helpful.

Linear Algebra

1. A notebook, math_linear_algebra.ipynb, is given in …/References/ipynb_files/

Linear Algebra is the branch of mathematics that studies vector spaces and linear transformations between vector spaces, such as rotating a shape, scaling it up or down, translating it (ie. moving it), etc.

Machine Learning relies heavily on Linear Algebra, so it is essential to understand what vectors and matrices are, what operations you can perform with them, and how they can be useful.

Linear Regression

2. See its theory in details from Chapter 2 and 3 of the e-book "Introduction to statistical learning" by Gareth James, et al., from directory…/References/PDF_files.

- *Edited by Chunshan Li..*