Workshop 9

Learning From Big Data

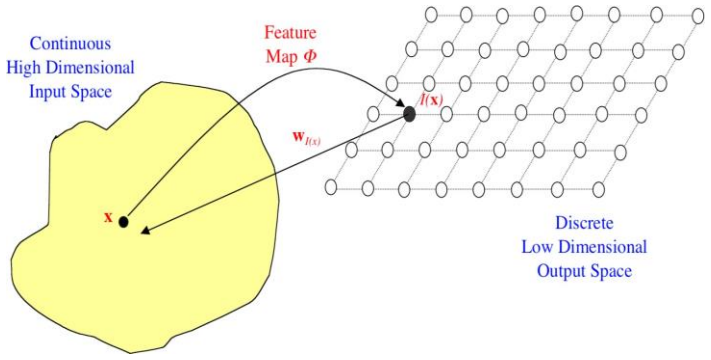
Self-Organizing Map (SOM)

*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".

As lectured, The SOM is based on the neurobiological studies: different sensory inputs (motor, visual, auditory, etc.) are mapped onto corresponding areas of the cerebral cortex in an **orderly fashion,** so called "the principle of topographic map formation."

In the area of deep learning, the applications of SOMs is to **transform** an input space of arbitrary dimension into a one or two dimensional discrete map (popularly in 2D).

A 2D map is constructed by a grid of neuros,



The spatial location of an output neuron in this topographic map corresponds to a particular domain or feature drawn from the input space, which may be used for classification.

Task 1: Building SOM

In this exercises, we are going to implement a recommendation system, by building a self-organizing map. It's a unsupervised deep learning project.

We will use the SOM to solve a business problem, so called fraud detection for a bank.

Suppose that we're given a data set, the information of customers applying for credit cards. Basically, these information are the data that customers had to provide when filling the application forms.

Our mission is to detect potential fraud within these applications or customers.

So, by the end of the mission we should give the explicit list of the customers who potentially cheated.

Actions:

- Open a notebook, som.ipynb, given in …/Lab9/Self_organizing_maps/, where we learn how to build a SOM, train the data, and find a list of potential cheaters.
- Study the codes in each cell of the notebook, including the comments, while you run them.
- Try to understand the codes and answer the questions I give in the comments.

Task 2: Mega-Case Study

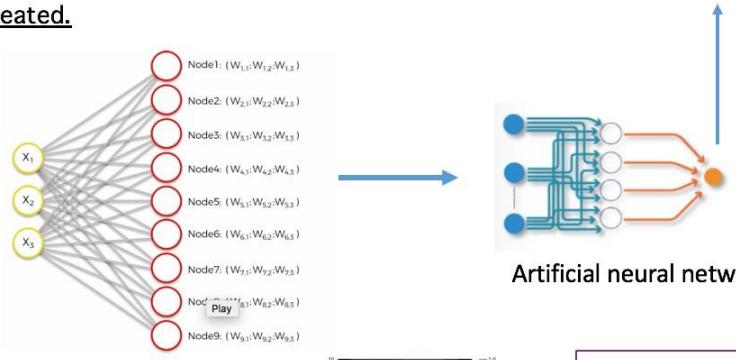
It's an extension to **Task 1** and do more work. The idea is to make a more advanced model, where we can predict the probability that each customer cheated.

To achieve this, we need to go from unsupervised to supervised learning, making a Hybrid Deep Learning Model.

The work consists of two parts:

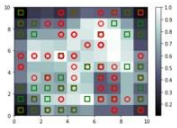
1) Simply to obtain the results from the previous SOM training for identifying potential cheaters.

2) Based on the list of customers, who are potential cheaters, we add a supervised model, an ANN, to predict the probability of each customer cheated.



Artificial neural network

Self-Organizing Map
(unsupervised learning)



Mega_case_study.ipynb, is presented in …/Labs/Lab9

Actions:

- Open a notebook, Mega_case_study.ipynb, given in …/Lab9, where we implement a hybrid model, which consists of SOM and ANN. by using this advanced model, we can predict the probability of each customer cheated.
- Study the codes in each cell of the notebook, including the comments, while you run them.
- Try to understand the codes and answer the questions I give in the comments.

Task 3: (Optional)

Do some research on the Internet or other source and write an outline or short paper for the comparison between SOM and K-means Clustering. At least, you should answer the following questions:

- What is the goal of each approach?
- What's the key difference of learning technique between them?
- Can you use SOM to achieve the goal of K-means (or vice versa) ?