Machine Learning 2020 Course — Homework

Matteo Matteucci
matteo.matteucci@polimi.it

Marco Cannici
marco.cannici@polimi.it

Politecnico di Milano — June 14, 2020

Protein Expression in Mice with Down Syndrome

Background

«**Down syndrome (DS)** is a chromosomal abnormality associated with intellectual disability and affecting approximately one in 1000 live births worldwide. It is due to an extra copy of the long arm of human chromosome 21 (Hsa21) and the consequent increased level of expression, due to dosage, of some subset of the genes it encodes. The overexpression of genes encoded by the extra copy of a normal chromosome in DS is believed to be sufficient to perturb normal pathways and normal responses to stimulation, causing learning and memory deficits. The **memantine drug** is currently in use for treatment of moderate to severe Alzheimer's Disease (AD) and has been proposed for treatment of learning deficits in DS. While memantine is known to modulate excitatory neurotransmission through antagonizing activity of N-methyl-D-aspartate (NMDA) receptors, little is known about its effects on protein expression, either alone or with learning paradigms.

In this dataset, the effect of memantine on protein responses is studied in the partial trisomy mouse model of DS, named **Ts65Dn**. Untreated Ts65Dn mice fail to learn in context fear conditioning CFC but if they are first injected with memantine, they learn successfully, i.e., learning is rescued. Protein lysates were prepared from brains of 3 month old male Ts65Dn Down syndrome model mice and their male littermate wild type controls, after training in context fear conditioning (CFC) with and without injection with the drug memantine. The context-shock (CS) group of mice are placed in a novel cage, allowed to explore for several minutes and then given a brief electric shock; normal, wildtype mice learn to associate the novel context with the aversive stimulus and will freeze upon re-exposure to the same cage. To control for the effects of the shock alone, a second group of mice, the shock-context (SC) group, are placed in the novel cage, immediately given the electric shock, and then allowed to explore; with these conditions, normal, wild type mice do not learn to associate the novel cage with the shock and do not freeze upon re-exposure to the same cage. Unlike their wild type littermates, the Ts65Dn CS group of mice fail to learn and do not freeze; this learning impairment can be corrected, however, if the Ts65Dn are injected with memantine prior to training. To control for the effects of injection alone, an additional CS group is also injected with saline (no drug). »

— Higuera, Clara, Katheleen J. Gardiner, and Krzysztof J. Cios. "Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome." PloS one 10.6 (2015).

Dataset Composition

The data set consists of the expression levels of 77 proteins/protein modifications that produced detectable signals in the nuclear fraction of the cortex. Eight classes of mice are described based on features such as genotype, behavior and treatment. According to genotype, mice can be control or trisomic. According to behavior, some mice have been stimulated to learn (context-shock) and others have not (shock-context) and in order to assess the effect of the drug memantine in recovering the ability to learn in trisomic mice, some mice have been injected with the drug and others have not.

Classes

The goal is to classify mice into 8 different classes, based on protein expression. For each sample, an additional attribute `class` identifies the mice category:

- `c-CS-s` : control mice, stimulated to learn, injected with saline
- `c-CS-m` : control mice, stimulated to learn, injected with memantine
- `c-SC-s` : control mice, not stimulated to learn, injected with saline
- `c-SC-m` : control mice, not stimulated to learn, injected with memantine
- `t-CS-s` : trisomy mice, stimulated to learn, injected with saline
- `t-CS-m` : trisomy mice, stimulated to learn, injected with memantine
- `t-SC-s` : trisomy mice, not stimulated to learn, injected with saline
- `t-SC-m` : trisomy mice, not stimulated to learn, injected with memantine

Train/Test splits

The dataset is provided already split into train (367 samples) and test (245 samples) sets. While the training set comes with all the 77 protein measurements for each sample, a malfunctioning in the analysis procedure caused the measurements of the `SOD1_N` protein expression to be lost in all test samples. Features are all real values and no data is missing, except from the `SOD1_N` feature on test samples.

- `train.csv` : 367×77 dataset
- `test.csv` : 245×76 dataset

Requests

1. Perform a **preliminary analysis** on the data using the techniques presented during the course. For instance, but not limited to, visualize features, identify if features (i.e., protein expressions) are correlated and determine which features are most correlated with the target class. Using **clustering**, study if there are structures in the data that allow samples from different classes to be easily identified. Compare the performance of different clustering algorithms on this data using the metrics presented during the course. You can limit this preliminary analysis only to training data.
2. Perform **classification** in order to classify mice in the 8 different classes. Perform features selection, compare different algorithms and identify the one that works the best on this dataset. Finally test the performance of the best algorithm on the provided test set. Since the `SOD1_N` feature is missing in test samples, remove it from training samples before training.
3. We want to recover from the data loss of `SOD1_N` features on test samples. Train a **regressor** which is able to predict the value of the `SOD1_N` feature given the remaining ones. Compare different regression algorithms for this task. Since `SOD1_N` features are missing in test samples, use only the training data for this step and make use of robust evaluation techniques to compare algorithms.
4. Use the regression model trained at the previous step to recover the `SOD1_N` column, by predicting the `SOD1_N` value of each test sample. Determine if the test performance of the best model found at step (2.) improves if the `SOD1_N` feature is also used for prediction (and training).

Submission

Use this form to submit your homework: `http://tiny.cc/ML2020Submission`
The deadline is July 23 at 23:59 (UTC+1)

- Create a Jupyter notebook to answer all the requests, using the libraries presented during the laboratory classes.
- Include Name, Surname and Student ID in the notebook.
- You are free to use the structure that you prefer within the notebook. However, please use markdown cells (`Cell > Cell Type > Markdown`) to insert section titles and clearly identify the different requests. You are free to add subsections to make the notebook more readable.
- Add text cells (markdown) to briefly explain what you did and why, and to help you answer the requests.
- Please, check that the notebook can execute correctly before submitting your work.

  Press `Kernel > Restart & Run All` and check that all cells execute correctly without errors.