UEF Summer School 2017

Machine Learning Applied to Bioinformatics and Speech Technology

Project tasks

General info

- Larger machine learning + bioinfo tasks that will be worked on individually
- Three available options for the project (next slides)
- Returnables:
  - Final report:
    - Introduction, methodology, practical implementations, results
  - Source code
  - Returned to Moodle (link will be added on “Tasks” page on website)
  - (Also remember to do fill in learning diary)
- Deadline: 8.9.2017, submission to Moodle
- Daily tutoring sessions 21.8-25.8 10:00 - 12:00.
  - Chat running on the website during the second week (See “Chat” section on website)
  - Synapse discussion board for this summer school (See “Task” section on website)

General info

- Implementations in Python or R preferred
  - Read: These are the languages we TAs actually are familiar with.
  - Other languages are allowed too, but we suggest to use pre-made implementations as often as possible.
- Unsolved problems: No answers and “correct ways” to do things, yet!
  - As such, no need to achieve superb performance.
  - We still expect clever ideas and executable code that solves the task on some level.
- For computing resources: CSC Taito and Taito-GPU
  - Account can be created based on UEF account here: https://sui.csc.fi
  - We have limited amount of temporary accounts available you are not able to create an account.
  - https://github.com/CSCfi/machine-learning-scripts/tree/master/courses/uefml2017

Project 1: Detecting Parkinson's disease from speech samples

- Thousands of samples of "aaaa" spoken to a microphone, from PD patients and control group.
- **Task:** Classify samples to PD and non-PD. Additionally analyse how PD samples differ from control group.
- Pre-processed dataset + evaluation scripts.
  - Dataset is already split into training and evaluation set.
  - Evaluation with EER (Equal-error-rate) and AUROC (Area Under ROC)
- Requires **validated** Synapse account and **access to mPower data**.
- Main TA: Anssi Kanervisto

Project 2: RNASeq

- The aim is to predict the proportion of different cell types from bulk RNASeq using clever approach utilizing data from single cell RNASeq -> cell type deconvolution problem
- **Task:** Can you come up with a clever way to utilize scRNAseq to deconvolve the signal in the Alzheimer brain tissue samples?
- Presentation on data by Thanneer Perumal available on “Course Notes” section.
- Data
  - Single-cell RNA-seq data: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67835
  - Bulk RNA-seq data: https://www.synapse.org/#!Synapse:syn3163039
- Main TA: Juha Mehtonen

Project 3: Brain imaging data

- **Task:** Associate differentially expressed genes in AD with images from the brain. Which brain region/cell expresses which gene?
- Data
  - Allen Brain Atlas