FINAL PROJECT [ROBT407]

NAZARBAYEV UNIVERSITY
SCHOOL OF SCIENCE AND TECHNOLOGY

INTRODUCTION

This final project aims to provide you with an opportunity to apply the machine learning algorithms (not restricted to the ones covered in class) to interesting real-world learning problems. In a supervised learning framework, given $X := \{x_1, ..., x_m\}$ with corresponding labels $Y := \{y_1, ..., y_m\}$, where $y_i \in \{\pm 1\}$ for $i = 1, \ldots, m$. We seek to infer a function $g : \mathcal{X} \to \{\pm 1\}$ to predict accurately whether a new observation will belong to class +1 or −1. In general, g - could be any learning machine defined on any learning from data problem (supervised, semi-supervised, unsupervised or active, reinforcement and many other types), since the core-foundational knowledge we studied in this course via VC theory of generalization and regularization techniques should apply to any g. This project is a chance to you and your team to optimize a robust machine learning model (i.e. come up with a new angle on an old problem). Successful implementation on benchmark datasets has a potential to become full-fledged research papers. Everyone of you can go way above and beyond the state-of-the-art methods.

IMPORTANT DATES

- Announced date: 06/11/2018 [*Start brainstorming about your project -**NOW***]
- Project proposal date: 11/11/2018
- Project presentation date: 30/11/2018
- **Final report date:** 30/11/2018

METHOD OF DELIVERY

Assignment deliverables should be submitted via Moodle to the course instructor before the due date.

Deliverables

- **Project Proposal** - due date - **10/11/2018**
  A single page brief project proposal that includes the following information (submitted to Moodle)
  - Project Title.
  - The project idea (or problem definition).
  - Data set.
  - Software package that you will use.
  - Team members - up to three in each team, and their expected contributions.
  - Review and include relevant literature (1-3 papers).

1. **Report describing in detail the work of a team with the following sections (use the Latex TEMPLATE provided in the moodle; - length 4-8 pages long)**
   - Abstract
   - Introduction
   - Methods
   - Results
   - Conclusion
   - References
   - Contribution (what and how each member contributed to the project) .
2. **Ten minutes of Oral Presentation** (slides should be submitted to Moodle beforehand)
3. **Source Codes (Jupyter Notebooks) + Data sets** and a file named *README*, and include in it a short description of all codes files you are submitting.

LEVEL OF COLLABORATION ALLOWED

- Collaboration is allowed on this assignment – each group should consist of maximum of three students. Discussions on course materials and implementation of the project are encouraged.
- Each team should write the final solutions/reports separately and understand them fully. External resources can be consulted, but not copied from.
- You are expected to discuss and learn together on how to use a specific machine learning tool. Homework #3 have already prepared you to use a ScikitLearn toolbox which you can use to implement this project. In addition, there's bunch of tutorials both with videos, texts and other materials that you can refer to.

GRADING CRITERIA

- 40% - Implementation (well documented source code)
- 30% - Benchmark - Accuracy (i.e. comparison with the best performance in the literature)
- 15% - overall work and report quality
- 15% - discussion (for example of success/failure; limitations, etc.)

1 Machine learning tools

Since implementation of standard algorithms from scratch may take up significant amount of time, in this project, you are encouraged to use available machine learning tools/libraries and concentrate on model selection problem by applying an algorithm of your interest for a real-world problem. However, there is no restriction posed if you can manage your time and want to implement a novel algorithm from scratch. Only keep in mind that the deadline is the final one without any further extension.

There are many machine learning or data mining packages that provides very optimized implementation of learning algorithms including the ones listed below. It is highly recommended using one of them to achieve your final project goals:

1. **Scikit-Learn**: machine learning in Python (scikit-learn.org).
2. **Pytorch**: deep learning library in Python
3. **TensorFlow**: TensorFlow is an open source software library for numerical computation using data flow graphs for building deep learning models
   - **Tutorial**: https://www.youtube.com/watch?v=dYhrCUFN0eM.
4. **Google Colaboratory** https://colab.research.google.com

2 Specific Tasks

- First, you must **identify** data from domain of your interest, e.g., natural language processing, computer vision problems, DNA sequence analysis, text information retrieval, brain data analysis etc.
- **Perform model selection** to estimate the model with optimal hyper/ parameters that solves the problem of your chosen domain. We can regroup under **model selection** a number of problems, including:
  1. selecting the best features,
  2. selecting the best preprocessing (data normalization, mathematical transformations of feature space)
  3. selecting the best learning machine (e.g., neural network, deep learning architecture, linear model, kernel method, classification or regression algorithms etc.),
  4. selecting the best set of hyperparameters (number of layers and hidden units in a (deep) neural network, kernel type and regularization parameter in kernel methods, etc.).
  5. Implementing a cross validation (CV) strategy (standard 10-fold CV or nested CV, leave-one out CV and/or others)
- Different learning machines exist in literature such as linear models, neural networks, deep learning, convolutional networks and/or kernel methods etc, **you should adapt and apply** an algorithm of your interest and compare with other at least three methods try to outperform them in terms of a generalization performance. Explain what have you done to combat overfitting due to deterministic and stochastic noise.
  - For instance, you decided to work on a handwritten digit recognition task and you chose as your main algorithms a deep neural network (DNN) model.
  - Then, you are expected to optimize the DNN via your model selection strategy and compare its performance with the other standard algorithms (at least three) i.e linear soft-margin SVMs, Kernel soft-margin SVMs, Regularized Logistic Regression, Regularized Linear Regression or Gaussian Mixture Models.

Try to follow the following steps in your project:

1. Consider a dataset D (from any domain of your interest)
2. **Apply an algorithm of your choice on** D
3. Estimate its generalization error (E_{test})
4. **If**: generalization error smaller than what exists in the literature for the same dataset:
   - **End of the process: Outcome → Grade A:)**
5. **Else**:
   - Go back to step 2 with another algorithm or change the model selection strategy.

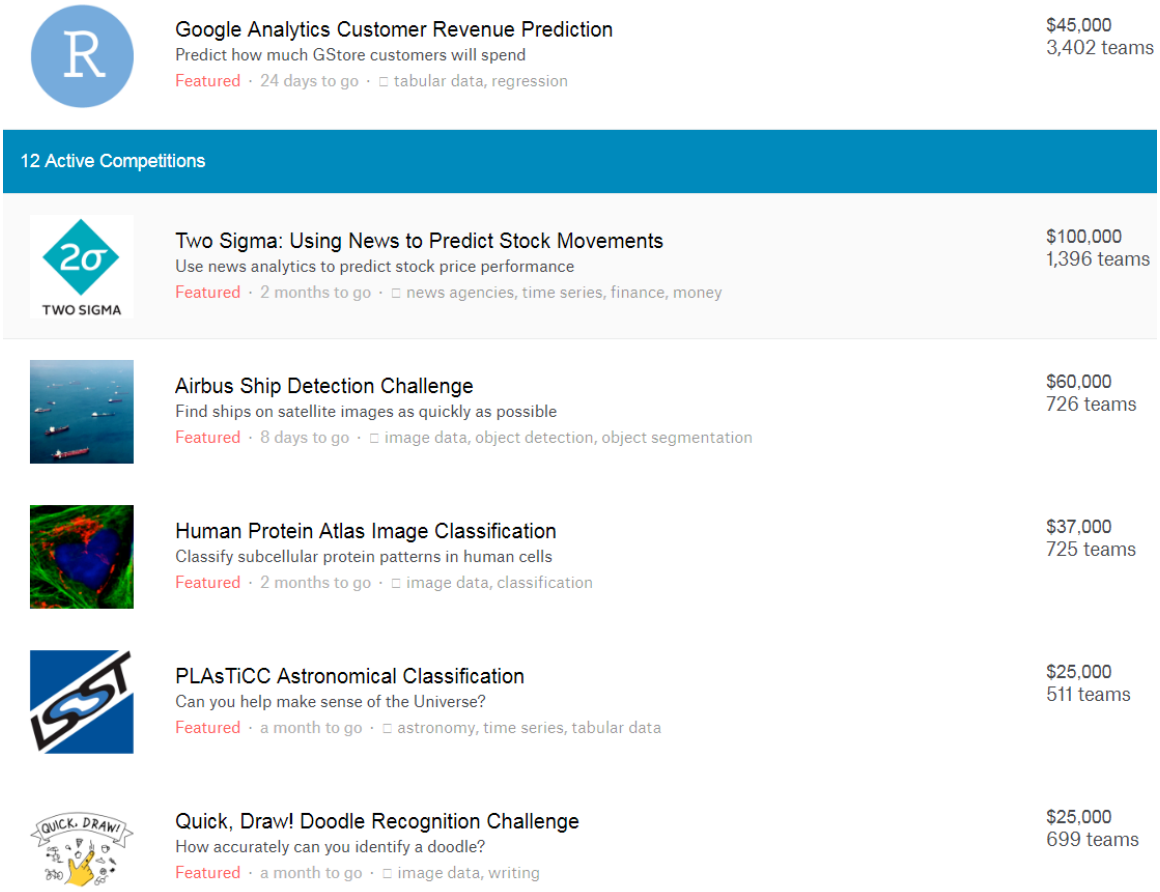
Project Databases and Ideas

Determining a domain of interest is one of the important task in this project. You have to spend sometime with your team members to explore internet databases dedicated for machine learning problems. For instance,

- For computer vision related datasets refer to : http://www.cvpapers.com/datasets.html
- Another popular and diverse data sets can be accessed in UC Irvine Machine Learning Repository : http://archive.ics.uci.edu/ml/
- Similarly, check Amazon- dataset repository https://aws.amazon.com/datasets/
- 19 Free Public Data Sets For Your First Data Science Project https://www.springboard.com/blog/free-public-data-sets-data-science-project/
- Review machine learning competitions and dataset at **www.kaggle.com**

→ ☻♫ **BONUS POINTS**: If you try to participate at any current competition posted at kaggle.com and use deep learning algorithms. For instance take a look at the competitions:

  * https://www.kaggle.com/competitions



  - Decoding the Human Brain
    * https://www.kaggle.com/c/decoding-the-human-brain

    - https://www.kaggle.com/c/grasp-and-lift-eeg-detection
  - Develop a Gesture Recognizer for Microsoft Kinect
    - https://www.kaggle.com/c/GestureChallenge

3 Some more project ideas are provides below

1) The text classification database

The task of Dexter is to filter texts about "corporate acquisitions". This is a two-class classification problem with sparse continuous input variables.

**See attachment for details on data description*

2) Netflix Prize Dataset

The Netflix Prize data set gives 100 million records of the form "user X rated movie Y a 4.0 on 2/12/05". The data is available here: **Netflix Prize**

Project idea:

- Can you predict the rating a user will give on a movie from the movies that user has rated in the past, as well as the ratings similar users have given similar movies?
- Can you discover clusters of similar movies or users?

Can you predict which users rated which movies in 2006? In other words, your task is to predict the probability that each pair was rated in 2006. Note that the actual rating is irrelevant, and we just want whether the movie was rated by that user sometime in 2006. The date in 2006 when the rating was given is also irrelevant. The test data can be found at this website.

- Reference

  https://www.kaggle.com/netflix-inc/netflix-prize-data

3) Brain-Computer Interfaces

Description: This data set was acquired during the real-time control of a computer cursor by the instructor of the course. Using a cap with 32 integrated electrodes, EEG data were collected from ten subjects while they performed three activities: imagining moving their left hand, imagining moving their right hand, and thinking of foot movements. As well as the raw EEG signals, the data set provides pre-computed features obtained by spatially filtering these signals and calculating the power spectral density.

- **Reference:**

1) Reference video: https://www.youtube.com/watch?v=4eJlkvwNT2U

- **Task:** Perform exploratory data analysis to get a good feel for the data and prepare the data for machine learning. Train at least two different classifiers to assign class labels to the test data to indicate which activity the subject was performing while the data were collected.
- **Challenges:** This data set represents time series of EEG readings. A baseline approach could be based on the given precomputed features. It might also be possible to train a classifier on a window of some size around each time step.

If you have any questions regarding this project/data please contact me.

4) RGB (2D) and RGB-D (3D) Datasets of Sign Language Alphabet Used for Fingerspelling in Kazakhstan

Two types of RGB datasets were collected: single-instance and multi-instance datasets. The multi-instance relational dataset was obtained by the method of appending consecutive frames of features of the same letter and the same participant to represent one class. In a single-instance dataset a class is represented by only a single frame of RGB data, letter and participant IDs. For multi-instance dataset the motion data was saved.The entire dataset of segmented hands consists of 73688 images for relational (i.e. multi-instance) and 2518 images for propositional (i.e. single-instance) sign language alphabet. Videos were recorded using the Logitech c310 web camera (HD 720p).

Similarly, two representations of depth dataset were collected using the Leap Motion sensor. A depth dataset was collected for 30 people aged between 20 and 30 years old.

- **Reference:**
- Tazhigaliyeva, N., Kalidolda, N., Imashev, A., Islam, S., Aitpayev, K., Parisi, G. I., & Sandygulova, A. (2017, May). Cyrillic manual alphabet recognition in RGB and RGB-D data for sign language interpreting robotic system (SLIRS). In Robotics and Automation (ICRA), 2017 IEEE International Conference on (pp. 4531-4536). IEEE

5) 3D Motion Dataset of Children

Depth dataset of children's skeleton motion data was collected from around 100 children (aged between 5 and 16 years old). This dataset was utilized to estimate age and gender of children based on their motion data.

- **Reference:**
- Sandygulova, A., Dragone, M., & O'Hare, G. M. (2014, August). Real-time adaptive child-robot interaction: Age and gender determination of children based on 3d body metrics. In Robot and Human Interactive Communication, 2014 RO-MAN: The 23rd IEEE International Symposium on(pp. 826-831). IEEE.

For questions, please email anara.sandygulova@nu.edu.kz

Credits & Used resources

Note following external resources were used in preparation of the project ideas which you can refer further:

- www.kaggle.com
- http://www.chalearn.org
- www.stanford.edu/~piech/cs221
- www.cs.cmu.edu/~10701/projects.html