UNIVERSITY OF ONTARIO INSTITUTE OF TECHNOLOGY

Quickstart of Kaggle Competition

Electrical and Computer Engineering
Yu Huang

What is Kaggle?

Kaggle: https://www.kaggle.com/

Useful resources on Kaggle

- All Getting Start competitions: all provides tutorials
- All scripts of all competitions
- Playground competitions: first attempts to the competition

Preparation

Based on Python Language

1. Tools: Data processing + Machine Learning + Data visualization
    a. Data processing: Pandas, Numpy
    b. Machine Learning: sci-kit learn (clustering, classification, pre-processing, evaluation), Nltk (NLP), Keras and TensorFlow (Neural Network)
    c. Data visualization: Matploit, Seaborn
2. Set up your local environment
    a. Linux (VM based on Vagrant)
        i. http://datasciencetoolbox.org/
    b. Windows
        i. Install Anaconda directly

Kaggle Tutorial

Titanic

https://www.dataquest.io/course/kaggle-competitions

San Francisco Crime Classification

https://www.kaggle.com/blobs/download/forum-message-attachment-files/2808/crimeSF_NN_logodds.ipynb

A Glance at **Data Mining** and **Machine Learning**

Simple Examples (Book Clustering)

1. 3000 lines and paragraphs from three different books (1000 each)
2. 4087 features (stemmed but stopwords are not removed)
3. Task: cluster these 3000 data into three clusters (k=3) and submit a simple CSV file including one column cluster tag (1, 2, or 3).

Simple Example (Missing Link Prediction)

1. An 85 x 85 adjacency matrix of a co-authorship network is given (denoted by A)

   A(i,j)=1 i and j are connected (they co-authored some articles with each other)

   A(i,j)=0 i and j are not connected (they haven't collaborated yet)
2. 40 elements of (links) of the matrix are missing (with negative values)
3. A(i,j)=-1 the link is missing, it could be one or zero.
4. 20 out of 40 missing links were originally connected (values 1).
5. The task is to predict the original values of missing values.
6. Submit the completed matrix with no negative values (in CSV format).

Simple Example (Gender Prediction)

1. 1166 training data from a photo sharing website. The date belong to 1166 female and male users (735 male and 431 female).
2. In trainingclass.csv, “1” is female and “0” is male.
3. 120 test data (their gender is unknown)
4. 390 features or photo tags (the actual words are not given).
5. The task is to predict the gender of 120 users in test data.
6. Submit a csv file of 120 predicted class (0 or 1).

Reference

Prof Masoud Makrehchi‘s Slides