SD701 Big Data Mining

Apache Spark Session Lab 3

Albert Bifet & Jacob Montiel

October 25, 2017

Kaggle competition

- **Kaggle** is a platform for predictive modelling and analytics competitions on which companies and researchers post their data and statisticians and data miners from all over the world compete to produce the best models.

- **Kaggle** also hosts recruiting competitions in which data scientists compete for a chance to interview at leading data science companies like Facebook, Winton Capital, and Walmart.

Kaggle competition

- In this lab, we are going to participate in a Kaggle competition:
  - Register at Kaggle and read:
    - `https://www.kaggle.com/c/titanic`
  - Participate in the Kaggle competition for this lab:
    - `https://www.kaggle.com/c/forest-prediction/`
- The grade for this course will depend on the results submitted to the Kaggle competition. You will need to send the source code used with a brief explanation.

Kaggle competition

- Login in the Community Edition of Databricks:
  - `http://community.cloud.databricks.com/`
- Create a cluster that uses Spark 2.0

- Import from Workspace this notebook as an example of preparing a submission to Kaggle
  - `https://drive.google.com/file/d/0Bz5RPwpp2VWxcjlDT0RHYms2WG8/view?usp=sharing`
- Read and run the notebook
- Previous Labs:
  - `https://drive.google.com/file/d/0Bz5RPwpp2VWxcVpXdTM0WlNkbFE/view?usp=sharing`