



Introduction to **Machine Learning and Data Mining** (Học máy và Khai phá dữ liệu)

Khoat Than

School of Information and Communication Technology
Hanoi University of Science and Technology

About the course

- Period: 15 weeks
 - Lectures: 11-12 weeks
 - Project report: 3-4 weeks
- Lecture directory: tiny.cc/046cjz
- Time & location:
 - 13:15 – 15:45 Tuesday at D9-102
- Question + advice:
 - Reserved by email: khoattq@soict.hust.edu.vn
 - DSLab, room 1002, building B1
- Join and discuss somethings with us:
<http://www.facebook.com/groups/1578056932500777/>



Contents

- Lecture 1: introduction to Machine Learning & Data Mining
- Lecture 2: data crawling and pre-processing
- Lecture 3: linear regression
- Lecture 4: clustering with K-means
- Lecture 5: classification and kNN
- Lecture 6: random forest
- Lecture 7: probabilistic models
- Lecture 8: support vector machines (SVM)
- Lecture 9: neural networks
- Lecture 10: model assessment & selection
- Lecture 11: frequent itemset mining
- Lecture 12: practical advices

Goals of the course

- Help students to have a good basic background on Machine Learning & Data Mining.
- Identify the main **advantages** and **limitations** of the methods/models in ML&DM.
- Be able to design & implement an ML/DM-based system, and evaluate its performance.

Some technologies/libraries



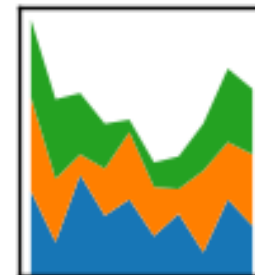
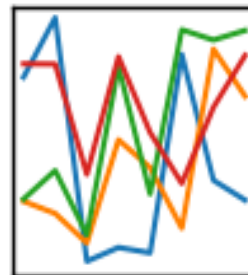
ANACONDA
Powered by Continuum Analytics®



TensorFlow

pandas

$$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$$



Evaluation (đánh giá)

- Attendance and activeness
- Midterm test: **Capstone Project or IBM Badges**
- Final exam
 - Paper-based test
- Overall: Midterm test (40%) + Final exam (60%)

IBM Badges

- Join and take two badges from the following online courses, provided by IBM:
 - Machine Learning with Python
(<https://cognitiveclass.ai/badges/machine-learning-python>)
 - Data Science for Business – Level 2
(<https://cognitiveclass.ai/badges/data-science-business-graduate>)



Capstone Project

- Students work in groups, each consists of 3-4 students.
- Each group choose a problem/topic to be solved, datasets to be used, algorithms in ML/DM.
- Each proposal should be precisely described
 - The problem: short description, input, output, data type, future application, ...
 - The algorithms or tools, planned to be used
 - Data sets to be used
- **Project registration: before 01/04/2020**
 - Via Google Form (TBA)

Capstone Project: requirements

- The result will be presented in the ending period of this subject. Every member is required to contribute to his/her project.
- Project report:
 - **Source code:** save your code into one zip file
 - **Readme.txt:** describes clearly how to setup, compile, and run your code
 - **Written report:**
 - Introduce the problem to be solved, the data sets were used
 - Details about the methods for analyzing data
 - Results of different evaluations, new conclusions/findings, ...
 - The main components of your code
 - The difficulties in this project, and your proposed solution
 - ...

Capstone Project: evaluation

- The evaluation of each project will be based on
 - The difficulty of the problem of interest
 - The appropriateness & quality of the chosen method/solution
 - The rigor of the empirical evaluation and assessment on the chosen method/solution
 - The quality of the presentation
 - The quality of the written report
- Each project will have 15' for slide presentation & demo
- **If you use some existing libraries/packages/codes, you have to clearly declare your usage in the written report and slide presentation**

Some references

- Lecture slides
- Reference books:
 - T. M. Mitchell. *Machine Learning*. McGraw-Hill, 1997.
 - Trevor Hastie, Robert Tibshirani, Jerome Friedman. *The Elements of Statistical Learning*. Springer, 2009.
 - Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT press, 2016.
 - E. Alpaydin. *Introduction to Machine Learning*. The MIT press, 2020.
 - Jiawei Han, Micheline Kamber, Jian Pei. *Data Mining: Concepts and Techniques* (3rd Edition). Morgan Kaufmann, 2011.
- Software:
 - WEKA (<http://www.cs.waikato.ac.nz/ml/weka/>)
- Data for experiments:
 - UCI repository: <http://archive.ics.uci.edu/ml/>