T–61.6020: Popular Algorithms in Data Mining and Machine Learning P

Introduction (23.1.2008)

Nikolaj Tatti

ntatti@cc.hut.fi

Introduction

- Arrangments, Requirements, Prerequisites.
- The content of the course.

Basic info

- Postgraduate seminar course (5 cr).
- Sessions on Wed. 14–16, T5.
- Language: English
- Homepage: `http://www.cis.hut.fi/Opinnot/T-61.6020/`
- e-mail: `t616020@cis.hut.fi`

Requirements

- Give a presentation on the selected topic (abt. 45 min).
- Complete all 10 assignments and write a raport.
- Participitate to other presentation (one absence is allowed).

Grade

- Course will be graded (1–5).
- Presentation (0–2).
- Active Participation (0–1).
- Assignments (0–3).

Prerequisites

- Probability theory.
- Linear Algebra.
- (Optimization theory).
- Experience with scientific papers.
- Programming (Matlab and Python).

Presentation

- Clear presentation.
- The goal is to teach the fellow students.
- Don’t put too much stuff.
- Use examples.
- Don’t put formulas that you are not going to explain.
- Practice the presentation couple of times (time it!)

Presentation

- Emphasis on the algorithm.
  - How does it work?
  - Why does it work?
- In ideal case, students should be able to implement the algorithm based on your presentation / slides.
- You should send the slides abt. week before the presentation to `t616020@cis.hut.fi`, so that we can comment them.

Extra topics

- This course is designed for 10 students.
- If we have more than 10 students, some of you will give talk on the topic that is not on the list.
- These presentations will be given after the main 10 algorithms are presented.

Participation

- Students are encouraged to ask questions during and after the presentation.
- Affects the grade.
- The presentation is your best shot to understand the algorithm.
- Less boring.

Assignments

- There will be 10 assignments, one for each algorithm.
- In each assignment you need to implement the algorithm and test it on some toy data.
- For each assignment you need to write a short report explaining the algorithm (abt. 1 page), your results (abt. 1 page). You should also attach the code into your report.
- Assignments will be separated into two parts (5 assignments per part).
- Deadlines will be given later.

Assignments

- We will provide stubs for the assignments either in Matlab or Python.
- We also provide the toy data sets.
- You don't have to use the stubs.
- If you have any problems you can always ask us (`t616020@cis.hut.fi`).

The Content

- This course is about algorithms!
- The list of the algorithms is based on `http://www.cs.uvm.edu/~icdm/algorithms/index.shtml`, Top 10 Algorithms in Data Mining from ICDM 2006.
- It is not *The* list but it is *a* list.

Algorithms

- Classification: Decision trees (id3), mixture models, kNN, SVM, AdaBoost.
- Global analysis: K-Means, EM, PageRank.
- Pattern searching: APriori, FP-Tree.

Classification

- Given input and output, the goal is to learn the function that can reproduce the output from the input.
- Decision trees.
- Mixture models.
- kNN.
- SVM.
- AdaBoost.

Global Analysis

- The goal is to summarise the data.
- K-Means.
- EM for clustering.
- PageRank (link analysis)

Pattern Search

- Searching for information explaining (small) portions of data.
- APriori.
- FP-Tree.