Data mining course

Erinija Pranckevičienė
(erinija.pranckeviciene@mf.vu.lt)
Vita Tomkutė
(vita.tomkute@gmc.vu.lt)

2019 November-December

Getting to know each other questions

- Name, surname, nickname, describe yourself a bit
- What program graduated?
- How much are you familiar with R?
  - Rstudio environment, R data structures, scripting, functions, cycles
  - Tidyverse suite, e.g. ggplot2, dplyr
  - linear models: glm, gam, caret
- Linear algebra
  - Do you know how to multiply matrices?
  - Do you know what is probability distribution? e.g. gaussian, binomial
  - Do you know how to compute correlation?
- Any data mining project you have worked with? short description.
- Do you know any data mining competition?
- Experience in Python, matlab, STAT, SPSS, WEKA?
- Are you familiar with version control systems? e.g. github
- What do you expect from this course?

Course structure

- Week 1 [VT]
  - T: Intro (github link, explanation); R — data manipulation, visualization, functions, getting data from/submitting data to github
  - W: cross validation; overfitting/underfitting, model complexity (bias-variance tradeoff); classification and regression result interpretation (error, confusion table); feature selection/extraction/design; dimensionality reduction
    Give project dataset
- Week 2 [EP]
  - T: Supervised learning: bayes rule, linear discriminant analysis LDA, quadratic DA(mention), K-nearest neighbor classifier (KNN), classification and regression trees (CART)
  - W: Supervised learning continuation: support vector machine (SVM), multiple classifiers systems
    Practice of each method with IRIS dataset

Course structure

- Week 3
  - T: review supervised learning from week 2, assembly methods [VT], Deep learning introduction: artificial neural networks (ANN), Hebb rule (reinforcement learning) [EP]
  - W: Unsupervised learning, K-means, Hierarchical clustering, PCA, multidimensional scaling [VT] Assign articles for next week article discussion
- Week 4
  - T: Important biomedical applications: article discussion (NOTE: benefit of article discussion is that you can find ideas for your project/project presentation)
  - W: Project presentation
- Week 5
  - Exam? TBD

Course evaluation

- Exam:
  - 35 % of the final evaluation
  - 2 h
  - 1 open question (25 %)
  - 10 multiple choice (75 %)
  - Date to be determined
- Project
  - 50 % of the final evaluation
  - 2 pages report (submit to github until 12.03 midnight)
  - 15 min presentation on 12.04
- Article discussion
  - 15 % of the final evaluation
  - 1 h / per article
- Exercises (3 times); answers and scripts submitted to github
  - Extra points

Project

- Same dataset for everyone
- Select 4 methods from all the methods we talked about & apply them to your dataset
- Compare results from different methods, select the best method
- Write a report (max 2 pages) about what you did (pdf)
- Submit your code to github page — include a link in your report (till 12.03 midnight — There will be penalty for late submission !)
- Short presentation (15 min) at 12.04
- Criteria; the following must be included in the report and presentation:
  - Figure describing dataset
  - 4 methods, why you choose them
  - Describe cross-validation strategy
  - Report train and test errors/confusion table
  - Describe evidence supporting the best method

Article discussion

- 5 articles (4 groups of 2 students, and 1 student alone)
- 20 min for article presentation (10 min each)
- ~ 40 min for discussion and questions

Must be included in the presentation:

- The main problem of the article
- Dataset presentation
- Methods in depth (highlight the ones that we presented in the course)
- Results and their interpretation

Exercises for extra points

First week:

- Dataset manipulation tasks, mainly R

Second week:

- Apply R script/function of each method on CARS dataset to classify observations into two or more classes, determined by the labels of a selected variable (TBD)

Third week:

- Clustering methods on other dataset (IRIS or TBD)
- apply multilayer perceptron to classify IRIS dataset into 3 classes

Literature

- J. Han. Data mining concepts and techniques (Third edition) (2011)
- T. Hastie et al. The Elements of Statistical Learning (2011)
- Manning et al. Introduction to Information Retrieval (2008)
- Duda et al. Pattern Classification (2000)
- Goodfellow et al. Deep learning (2016)

Our main github repository for this course:

https://github.com/VitaT/DM-course-2019