MPIMG

MAX-PLANCK-GESELLSCHAFT

Freie Universität Berlin

Seminar: Applied Machine Learning

Annalisa Marsico

OWL RNA Bionformatics group

Max Planck Institute for Molecular Genetics

Free University of Berlin

SoSe 2015

Goals of this course - I

- Soft skills
  - Learn how to evaluate a research paper
  - Learn what makes a paper good
  - Learn how to get your paper published
  - Learn how to give a scientific talk
  - Learn to be critical / evaluate

Goals of this course - II

- Hard skills
  - Get an overview of the Methods in Machine Learning field & the applications to Biology/Bioinformatics
  - Learn how basic concepts / algorithms/ statistical methods are applied and extended in this field
  - Learn how to ask the right biological question and choose the right Machine Learning method ‚to solve it‘

Course design

- Today -> overview on the topics, assignment of papers
- Student presentations
  - Each student will choose a paper and will give a presentation
  - Two presentations per term (30-40 minutes + 15 minutes questions)
  - Discussion: questions, critical assessmnet

Presentation guidelines

Compression with minimal loss of information

1. Understand the context & data used
2. Identify the important question/motivation
3. Focus on the method
4. Summarize shortly the main findings
   - Forget about unimportant details
5. Evaluate and think about possible future directions

Advices / Help

- Read your paper twice before saying ‚I don‘t understand it‘
- Read the supplementary material
- Do not try to understand every detail but the general idea has to be clear
- Main objective: lively interesting talk that promotes discussion
- Come anytime to me with questions (write me 3-4 days before)
  marsico@molgen.mpg.de
  Tel: +49 30 8413 1843
  where: MPI for Molecular Genetics, Ihnestrasse 63-73, Room 1.3.07
- send me your presentation one week before your talk
- Get feedback and give feedback (also to me ☺)

Practical information

Day	First talk	Second talk
April 16	Introduction	Introduction
~~April 23~~	~~Girls‘ Day~~	~~Girls‘ Day~~
April 30		
May 07		
May 14		
May 21		
May 28		
June 04		
June 11		
June 18		
June 25		
July 07		
July 09	backup	backup

Link to the Seminar webpage

http://www.molgen.mpg.de/3415218/Seminar-Applied-Machine-Learning

Topics

General machine learning papers

1. Assessing the accuracy of prediction algorithms for classification: an overview
2. An introduction to ROC analysis
3. A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection

Feature selection

1. A review of feature selection techniques in bioinformatics
2. Novel unsupervised feature filtering of biological data

Topics

Unsupervised Learning (applications to Bioinformatics)

1. Cluster analysis of gene expression data: A Survey
2. Biclustering algorithms for Biological data analysis: A Survey

Random Forests (applications to Bioinformatics)

1. Simple decision rules for classifying human cancers from gene expression profiles
2. Prediction of protein - protein interactions using random decision forest framework
3. Detection and interpretation of expression quantitative trait loci (eQTL).
4. RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State

Topics

Classification with Support Vector Machines (SVMs)

1. The spectrum kernel: A string Kernel for SVM protein classification
2. Kernel-based machine learning protocol for predicting DNA-binding proteins
3. A boosting approach for motif modeling using ChIP-chip data

Neural Networks and deep Learning

1. Gene prediction in metagenomic fragments: A large scale machine learning approach
2. Beyond the 'best' match: machine learning annotation of protein sequences by integration of different sources of information
3. Deep learning of the tissue-regulated splicing code

Topics

Active Learning (applications to Bioinformatics)

1. Predicting positive p53 cancer rescue regions using Most Informative Positive (MIP) active learning
2. Active Learning with Support Vector Machine applied to Gene Expression Data for Cancer Classification

Semi-supervised Learning (applications to Bioinformatics)

1. Semi-supervised learning improves gene expression-based prediction of cancer recurrence
2. Matching experiments across species using expression values and textual information

Multi-task Learning (applications to Bioinformatics)

1. Probabilistic multi-class multi-kernel learning: on protein fold recognition and remote homology detection
2. Integrating sequence, expression and interaction data to determine condition-specific miRNA regulation.