Aalto University
School of Science

Department of
Computer Science

Applications of Machine Learning

Professor Juha Karhunen

Personnel of the AML group

Group Head

Juha Karhunen

Docents

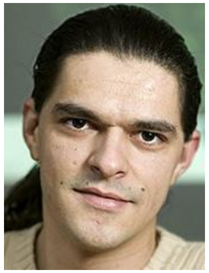
Miki Sirola

Francesco Corona

Postdocs

Mark van Heeswijk

Doctoral Students


Matthieu Molinier

Alexander Grigorievskiy

Luiza Sayfullina

Students

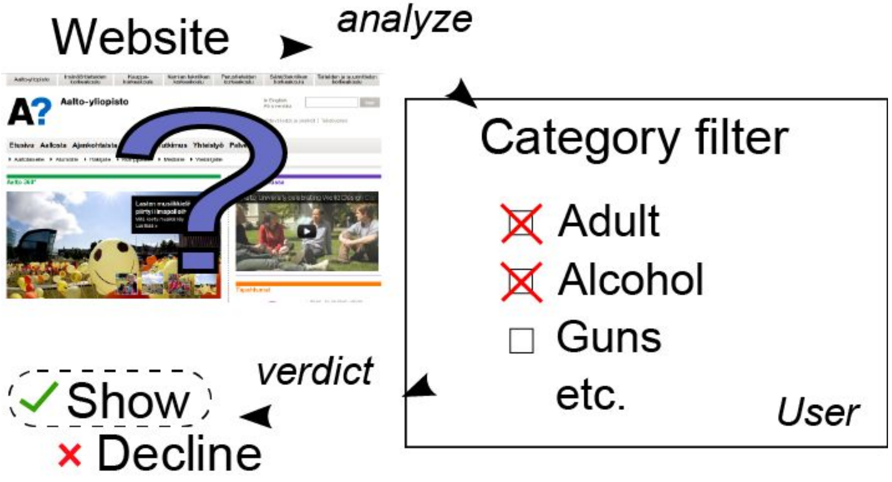
Zhao Chen

Scope and Goals

The AML group carries out both theoretical and experimental work on developing and applying new machine learning techniques for solving various application problems.

More specific research topics:

- Time series analysis and prediction;
- Dimensionality reduction;
- Extreme learning machines;
- Environmental applications;
- Industrial applications;
- Classification of web sites based on images;
- Detection of malicious Android software.



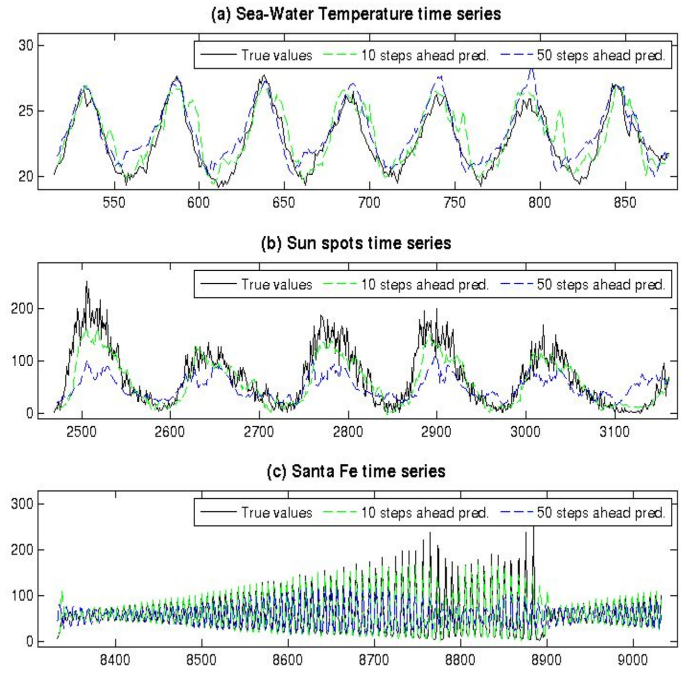
see also http://research.ics.aalto.fi/eiml/publications.shtml

Time series analysis and prediction

- Formerly the main research topic of the group.
- Often one predicts only one step ahead.
- We have studied prediction farther away.
- Linear methods for time series prediction and analysis are well-known.
- We have used nonlinear neural network and machine learning methods.
- Lots of possible applications in various areas of life.

Time series analysis and forecasting

- Compare and combine various time series methods: neural networks, Gaussian Processes, State-Space models.
- Focus on accuracy, computational speed and probabilistic forecasting.
- Address the problems of missing observations and unevenly sampled time series.
- Currently we use Astronomical and Electricity Consumption data.



Aalto University

Department of Computer Science

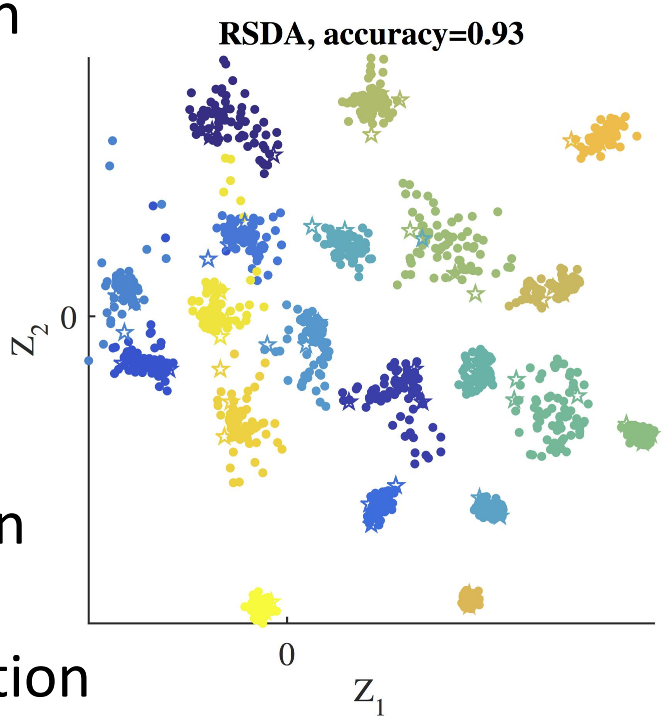
Dimensionality reduction - Variable selection

- The data are often too high-dimensional for methods used.
- The computation time can explode.
- This curse of dimensionality can be handled by data compression or variable selection.
- In variable selection, one selects the most important variables for the task at hand.
- The other variables are discarded.
- We have studied different methods for variable selection.
- And tested them with many real-world data sets.

Dimensionality reduction - 2D Linear projections

Supervised distance preserving projections, SDPP.

- Local pairwise-match of squared distances in projection and response or label space
- Optimisation via QSDP/SDLP or CG
- Kernel-SDPP for nonlinear problems

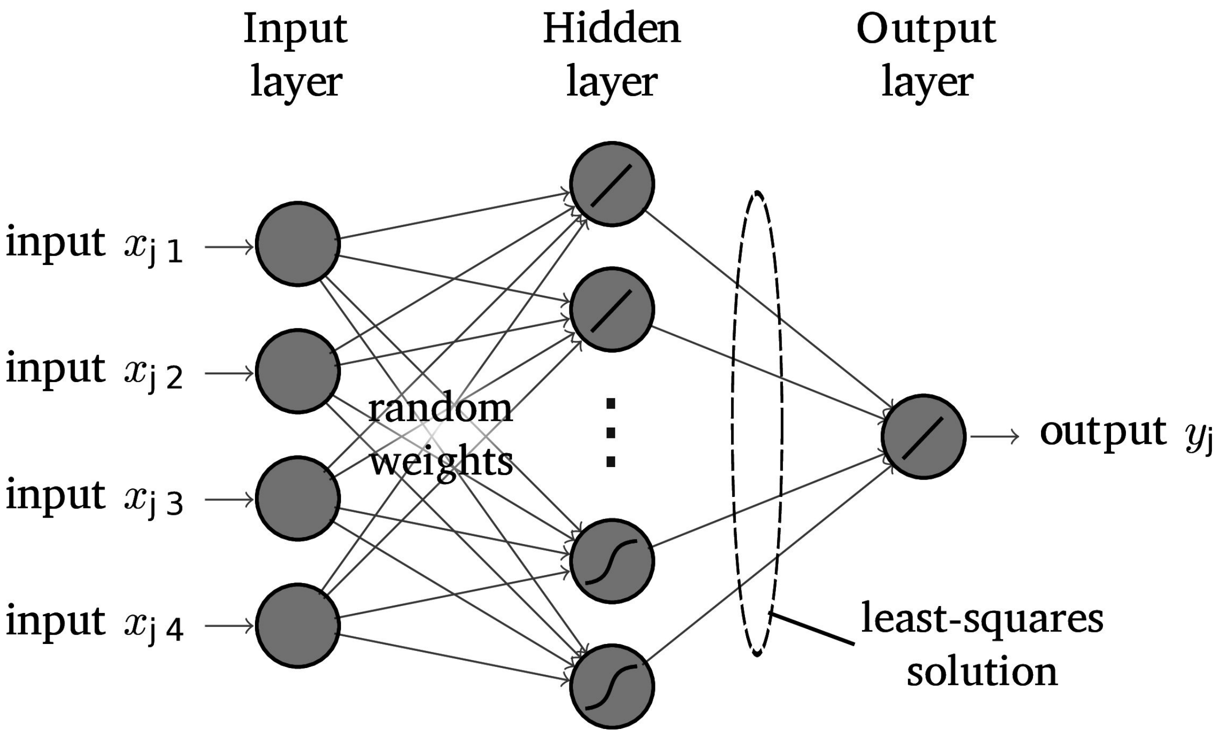


Stochastic discriminant analysis, SDA.

- Pairwise-match of Student's t probabilities in projection and label space (KL divergence)
- Gradient-based optimisation and regularisation

Neural Networks - Extreme Learning Machines

- Efficient and effective neural networks based on random nonlinear feature extraction, scalable to large data sets due to fast training.



Neural Networks - Extreme Learning Machines

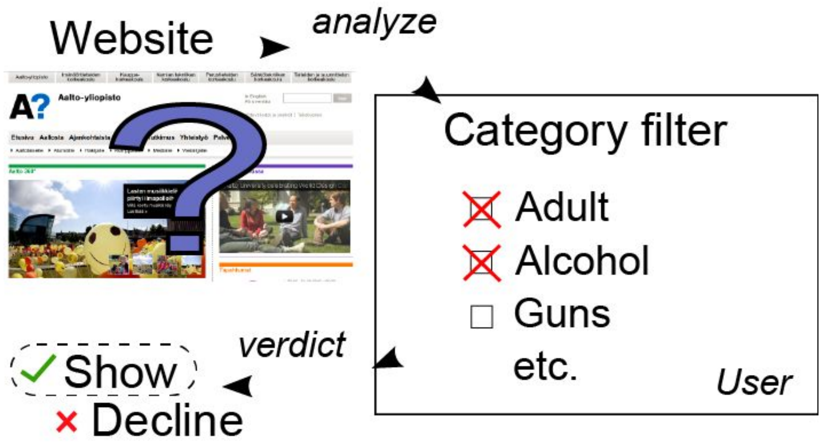
- Many improvements have been explored in our group:
  - hidden layer pruning
  - proper and fast regularization
  - improved accuracy through ensembles
  - GPU-acceleration and parallelization
  - sparse binary/ternary features
  - feature selection
  - compressive training algorithms

Environmental and industrial applications

- Dr. Francesco Corona leads this part.
- Multivariate predictive control of wastewater treatment plants (EU project, DIAMOND).
- Monitoring nitrate concentration in wastewater treatment plants (Viikinmäki, Helsinki).
- Property prediction of fuels in oil refineries (Sarroch, Italy).
- Equipment aging related noise measurement with TVO Olkiluoto nuclear power plant (Miki Sirola and a student making his Diploma thesis).

Classification of web sites

- Web sites have been tried to classify thus far only based on the text they have.



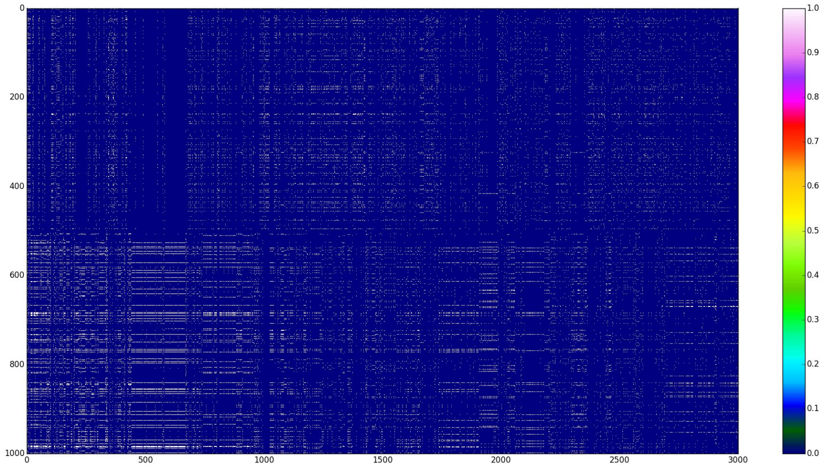
- We are using images on web sites for their classification.
- Trying to separate benign (harmless) web sites from undesirable ones.
- The classes of these undesirable web sites are for example crime, porn, racism, war, etc.

CloSe Project: Android Malware Detection

- The research is done in collaboration with F-Secure corporation
- They provided us a huge dataset of 120K malicious and benign files
- Main research goal is how to efficiently reduce the dimensionality of high-dimensional sparse binary data set for minimizing the desired cost function
- First publication on this topic: “Efficient detection of zero-day Android Malware using Normalized Bernoulli Naive Bayes”
- Graduate student Luiza Sayfullina works in this project
- Her instructor is Dr. Emil Eirola, and advisor Prof. Alex Jung

Dealing with high-dimensional sparse data

- Major issues are how to deal with sparsity, how to make a concise representation of the data, and what properties of the dataset will affect the choice of the dimensionality reduction.
- Below you can see sparse bag of words model sample from our malware dataset. In practice, random projections work well for sparse data compression.



Teaching responsibilities

- Prof. Juha Karhunen lectures the course T-61.5130 Machine Learning and Neural Networks (autumn).
- The course was renovated in autumn 2015.
- The assistant Dr. Mark van Heeswijk of this course comes from our AML research group.
- Juha Karhunen is the supervising professor for several Master’s (Diploma) thesis made outside our department every year.
- Alexander Grigorevskiy is a teaching assistant of T-61.3050 Machine Learning: Basic Principles. Previously he has been teaching assistant of T-61.3040 Statistical Signal Modeling.

Homepage

For more info, see our homepage
http://research.ics.aalto.fi/eiml/