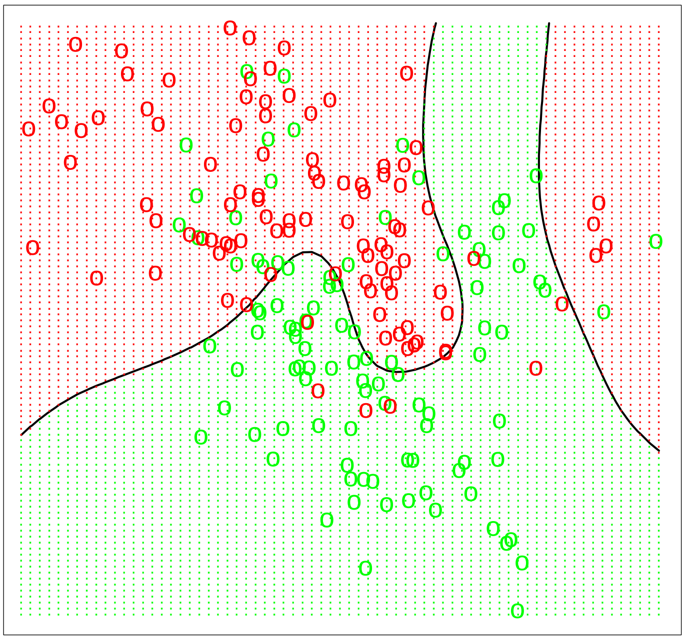
Course description
Learning systems, 7.5hp, Spring-12
Halmstad University

Learning Systems



antanas.verikas@hh.se

Contents

1	Welcome to the Learning systems course!	3
1.1	General content of the course	3
2	Course purpose and learning objectives	4
3	Lectures	4
3.1	Lecture: Introduction	4
3.2	Lecture: Introduction into classification	5
3.3	Lecture: Introduction into regression	5
3.4	Lecture: Project related issues	5
3.5	Lecture: Learning in linear systems	5
3.6	Lecture: Learning in nonlinear systems	6
3.7	Lecture: Generalization in learning systems	6
3.8	Lecture: Support vector machines (SVM)	6
3.9	Lecture: Other machine learning techniques	6
3.10	Lecture: Self-organization	7
4	Seminars - read, comprehend and present scientific articles	7
5	Projects	7
5.1	Classification project: Thyroid disease	9
5.2	Regression project: Predicting power load	10
6	Evaluation	13
7	Course evaluation	13
References		13

1 Welcome to the Learning systems course!

The purpose of this document to give you a rather comprehensive description of the "Learning systems" course. The course consists of lectures, two projects, and seminars. The course is of 7.5 hp split into 3.0 hp for lectures, 3.0 hp for projects, and 1.5 hp for seminars.

This year the course is given by Antanas Verikas (antanas.verikas@hh.se).

Preliminary schedule

Week 12				
M	19 Mar	10:15-12:00	Lecture	D215
T	20 Mar	08:15-10:00	Lecture	D215
Week 13				
M	26 Mar	10:15-12:00	Lecture	D215
T	27 Mar	08:15-10:00	Lecture	D215
Week 14				
M	2 Apr	10:15-12:00	Lecture	D215
T	3 Apr	08:15-10:00	Lecture	D215
Week 16				
M	16 Apr	10:15-12:00	Lecture	D215
T	17 Apr	08:15-12:00	Lecture	D215
Week 17				
M	23 Apr	10:15-12:00	Lecture	D215
T	24 Apr	08:15-10:00	Lecture	D215
Week 18				
M	30 Apr	10:15-12:00	Seminar	D215
Week 19				
M	07 May	10:15-12:00	Seminar	D215
Week 20				
M	14 May	10:15-12:00	Project	D215

1.1 General content of the course

Below given is a list of main topics of the course.

(1) Introduction to the learning systems area.
(2) Learning in linear systems.
(3) Learning in nonlinear systems.
(4) Multilayer perceptron (MLP).
(5) Support vector machines (SVM) and relevance vector machines (RVM).
(6) Generalization issues in learning systems.

(7) Fuzzy rules-based systems.
(8) Random forests.
(9) Self-organization in learning systems.

2 Course purpose and learning objectives

The purpose of the course is to let you develop basic skills in the field of machine learning. The objective of the course is to provide an overview of machine learning systems for classification, regression, and self-organization, to study basic learning algorithms in detail. Following successful completion of the course the student should be able to:

- Knowledge and understanding
  · describe basic linear machine learning algorithms;
  · describe basic nonlinear machine learning algorithms;
  · describe main application areas of machine learning algorithms.
- Skills and abilities
  · apply machine learning methods on real world problems;
  · present scientific results in the learning systems area.
- Judgement and approach
  · assess when and which machine learning methods are applicable;
  · analyze and explain scientific results from the machine learning area.

3 Lectures

3.1 Lecture: Introduction

Antanas Verikas
Format of the course (lectures, seminars, projects, examination), goals, content, schedule (see the table below), literature [1,2], short introduction into learning systems:

- Links with other research areas
- Definition of learning and self-organizing systems
- Examples of learning systems
- Typical tasks solved with learning and self-organizing systems

3.2 Lecture: Introduction into classification

Antanas Verikas
Formal description of classification task, examples of classification tasks, notations used, illustration of notations using an example from health care, Bayes classification rule, expected conditional risk, approaches to classification, parametric and non-parametric approaches, examples of decision boundaries of different complexity.

3.3 Lecture: Introduction into regression

Antanas Verikas
Regression task examples, data and assumptions, idealized regression and real regression, examples of model families, error measures, model bias and model variance, examples of regression results obtained using models of different complexity and different number of data points, relation between a prediction error and model complexity.

3.4 Lecture: Project related issues

Antanas Verikas
Tasks of the projects, important aspects to consider when solving classification and regression tasks, data exploration using: scatter plots, correlation coefficients, histograms, Fisher index, data transformations: standardization, PCA, whitening, Box-Cox transformations, outlier detection, comparing regression models, comparing classification models.

3.5 Lecture: Learning in linear systems

Antanas Verikas
Definition of a linear system, assumptions and goals, data representation, error function and error surface, optimal solution by pseudo-inverse, ridge regression, gradient descent and on-line gradient descent, ADALINE, simple perceptron, training perceptron by gradient descent, linear Gaussian classifier, Euclidean distance classifier, logistic regression.

Demonstrations using MATLAB: comparison of linear decision boundaries obtained for two different data sets using pseudo-inverse, simple perceptron, and Euclidean distance classifier. Discussion of the results.

3.6 Lecture: Learning in nonlinear systems

Antanas Verikas
Assumptions, polynomial model family, generalized linear model, K nearest neighbours regression, kernel regression, quadratic Gaussian classifier, K-NN classifier, multilayer perceptron (MLP), error backpropagation (BP) training algorithm, BP with momentum, adaptive learning rate, resilient propagation (RPROP), second order learning algorithms, examples of classification and nonlinear regression using MLP.

Demonstrations using MATLAB: comparison of decision boundaries obtained using pseudo-inverse, simple perceptron, Euclidean distance classifier, and MLP. MLP runs using different data sets and different number of hidden nodes. Discussion of the results.

3.7 Lecture: Generalization in learning systems

Antanas Verikas
Idealized and real regression, well-posed and ill-posed learning problems, estimating the generalization error, cross-validation, model complexity and model bias and model variance trade-off, Vapnik-Chervonenkis (VC) dimension, ways of controlling over-fitting: early stopping, regularization, committees, model pruning & growing, variable selection.

3.8 Lecture: Support vector machines (SVM)

Antanas Verikas
Margin of linear classifiers, maximum margin, mathematical description of linear SVM, solving the optimization problem, primal and dual problems in SVM, linearly nonseparable data, soft margin classification, "non-linear SVM", feature spaces, kernel trick, kernel functions, mathematical description of nonlinear SVM, properties of SVM.

3.9 Lecture: Other machine learning techniques

Antanas Verikas
Drawbacks of SVM, relevance vector machines (RVM), decision trees, random forests, fuzzy sets, techniques based on fuzzy rules.

3.10 Lecture: Self-organization

Antanas Verikas
Definitions, auto-encoder based on one hidden layer MLP, auto-associative neural network, K-means clustering, objective function and update equations, learning vector quantization (LVQ), self-organizing maps (SOM), multidimensional scaling (MDS), t-Stochastic neighbor embedding.

4 Seminars - read, comprehend and present scientific articles

You are expected to present scientific articles in the area of machine learning. Each student has to give a seminar (15 mins). Material is provided from the course responsible. The table below contains different articles you may choose from. Choose an article from the topic you are interested in.

When reading and presenting a paper, you may focus on finding the: objectives, methods, results, conclusions and scientific contribution. Other machine learning techniques than the ones discussed at the lectures may be included in the articles. To understand these techniques better it might be necessary to follow up reference articles.

One of the key aspects when presenting an article is to keep within the time you have been assigned to talk. The seminar leader will inform you about the time constraints.

To accumulate material for discussion you are expected to formulate at least one question for each of at least five articles. The questions are emailed to the seminar leader before the seminar starts. When you are presenting, your student colleagues will do the same for you.

5 Projects

You will do projects in groups of two students. Each group shall do two projects, one regression and one classification project. The project will be performed in MATLAB. Results are to be presented in a report and orally. Below given is a list of steps usually taken when accomplishing a typical classification and regression project. Each group will be given a preliminary list of steps to be taken in their specific projects to achieve the projects' objectives.

After four weeks of the course start a project seminar will be arranged. During the seminar students will be asked to give a short presentation of project

Articles for the seminar.

Topic	ID	Title
Neural networks	4	Color control of printers by neural networks
	6	Multivariate outlier detection and remediation in geochemical databases
	9	PID neural networks for time-delay systems
	11	Multivariate monitoring of fermentation processes with non-linear modelling methods
	14	Bootstrap for neural model selection
	16	WeAidUa decision support system for myocardial perfusion images using artificial neural networks
	18	Robust control of nonlinear stochastic systems by modelling conditional distributions of control signals
	19	Application of four-layer neural network on information extraction
	20	A comparison of supervised and unsupervised neural networks in predicting bankruptcy of Korean firms
	23	Optimal Ensemble Averaging of Neural Networks
	24	Soft sensors for on-line biomass measurements
Unsupervised learning	1	Wood inspection with non-supervised clustering
	3	The Evolving TreeA Novel Self-Organizing Network for Data Analysis
	17	Unsupervised and auto-adaptive neural architecture for on-line monitoring. Application to a hydraulic process
Hybrid methods	1	Medical image classification using genetic algorithm based fuzzy-logic approach
	7	A dynamic overproduce-and-choose strategy for the selection of classifier ensembles
	8	Ensemble learning via negative correlation
	10	Combining Neural Networks and Context-Driven Search for On-Line, Printed Handwriting Recognition in the Newton
	12	Classifier combination based on confidence transformation
	13	Divide-and-conquer approach for brain machine interfaces: nonlinear mixture of competitive linear models
	21	A Hybrid Neural Network Model for Noisy Data Regression
	25	Modeling and control of non-linear systems using soft computing techniques
SVM	5	An Analytical Method for Multiclass Molecular Cancer Classification
	15	Non-Technical Loss Analysis for Detection of Electricity Theft using Support Vector Machines
	22	Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters

results achieved so far.

5.1 Classification project: Thyroid disease

5.1.1 Task

Tell if a particular set of measurements (test results) comes from a person who is healthy, or suffers from being hypothyroid or hyperthyroid (i.e. 3 output categories).

5.1.2 Data

There are 7200 observations representing patients. You are given 5000 of these, and 2200 are withheld for out-of-sample testing. There are 21 variables (there is no information on what these represent).

You are given the file *thyroidTrain.mat*, which contains the matrices *trainThyroidInput* (5000×21), *trainThyroidOutput* (5000×3), and *testThyroidInput* (2200×21). The first matrix, *trainThyroidInput*, contains the input patterns for the training data. The second matrix, *trainThyroidOutput*, contains the outputs coded in a "1-out-of-3" fashion. That is, the outputs are coded as (1,0,0), (0,1,0), or (0,0,1). The third matrix, *testThyroidInput*, contains the inputs for the test data. You are supposed to use the latter file to produce outputs that are handed in to me.

5.1.3 Steps and subgoals

(1) Get acquainted with the data. Plot it and try to get a feel for the possible relationships between input and output.
(2) Try to transform the variables and see if this changes the information content (this can be measured using e.g. a "Fisher Index").
(3) Construct a k-nearest neighbor (k-NN) classifier for the problem, using all the variables, and estimate the generalization error (use e.g. $k = 5$).
(4) Prune the $k - NN$ classifier by successively removing the variable that results in the least degradation of the generalization error, until the degradation is significant. Note the classification error (generalization).
(5) Construct an artificial neural network (ANN) model using the remaining inputs. Optimize the number of hidden units (one hidden layer) with respect to the generalization error.
(6) Try to prune the ANN model by successively removing the variable that results in the least degradation of the generalization performance, until the degradation is significant. Optimize the number of hidden units for the final model. Note the classification error.

(7) Train a few networks with your optimal number of parameters. Combine these into a committee.
(8) Hand in the test results for your best k-NN classifier and your best ANN committee together with your estimate of the generalization classification error. **These results must be handed in no later than 48 hours before your oral presentation by emailing the report to the course responsible.**
(9) Write a report.

5.1.4 Report and presentation of results

You will present the results from your project in two ways: (1) A written report where the main conclusions are presented together with figures and tables supporting your conclusions. (2) An oral presentation, of about 15 minutes, to your course colleagues.

The report should be about 10 pages, including figures and tables, and should contain the elementary report constituents:

- Introduction (brief presentation of problem, 1 page)
- Methodology (brief listing of methods, 1 page)
- Data (presentation of your data set with important observations, 1-2 pages)
- Results (4-5 pages)
- Discussion (your results and comparison to other researchers' results, 1 page)

When you are finished with your report, and it has been accepted, then you should produce a PDF file with it, and pack it together with your data set and other important parts of your project (like MATLAB M-files). The idea being that someone else could unpack it and repeat the main steps in your analysis without rewriting everything.

5.2 Regression project: Predicting power load

5.2.1 Task

This is a nonlinear regression task: To predict the power load for Puget Sound Power & Light Co. 24 hours in advance, at 8 in the morning, when the current day is a working day and tomorrow is a working day (to make the problem a little bit easier, because things look a little different when tomorrow is a holiday or when the current day is a holiday).

To solve the problem, you get observations for the period January 1985 – October 1990 (in all seasons).

To test your system, the course responsible have withheld data from the winter months (November – March) of 1990/1991 and 1991/1992. You will be given the input data corresponding to those months, without information about the correct output, and asked to provide predictions for them.

The 15 input variables are:

(1) The current power load (MW)
(2) Average power load over the last 24 hours (MW)
(3) Average power load over the last week (MW)
(4) Peak power load during the last 24 hours (MW)
(5) Peak power load during the last week (MW)
(6) Forecasted temperature 24 hours ahead (Fahrenheit)
(7) The current temperature (Fahrenheit)
(8) Average temperature last 24 hours (Fahrenheit)
(9) Average temperature last week (Fahrenheit)
(10) Variance of the temperature over the last 24 hours (Fahrenheit^2)
(11) Variance of the temperature over the last week (Fahrenheit^2)
(12) Average forecasted temperature for the next 24 hours (Fahrenheit)
(13) Variance of forecasted temperature for the next 24 hours (Fahrenheit^2)
(14) $\cos(2\pi \cdot \text{daynum})$ where daynum = (the number of the day in the year)/365
(15) $\sin(2\pi \cdot \text{daynum})$

These inputs are the result of quite a lot of variable selection so you can assume that these variables should all be used.

The output that you shall predict is the load 24 hours ahead.

5.2.2 Data

You are given the file *PowerTrainData.mat*, which contains the following matrices: *powerTrainInput* (15×844), *powerTrainOutput* (1×844), *powerTrainDate* (1×844), and *powerTestInput* (15×115). The *powerTrainDate* is the date for the training observations (in MATLAB datenum format).

5.2.3 Steps and subgoals

(1) Get acquainted with the data. Plot the data and try to get a feel for the possible relationships between input and output.
(2) Construct a standard linear model and estimate the generalization performance.
(3) Construct a linear ridge regression model. Use cross-validation to estimate the ridge regression parameter λ. Estimate the generalization performance.

(4) Construct a nearest neighbor regression model. Estimate the generalization performance.
(5) Construct a standard multilayer perceptron (MLP) model and estimate the generalization performance.
(6) Construct a committee with 10 different multilayer perceptrons (no weight decay) and estimate the generalization performance.
(7) Produce the predictions for the test inputs and mail to me (for your best linear model, nearest neighbor, single MLP model and the MLP committee), together with your estimates for how well you will do on the test samples. **These results must be handed in no later than 48 hours before your oral presentation by emailing the report to the course responsible.**
8. Write a report.

5.2.4 Report and presentation of results

You will present the results from your project in two ways: (1) A written report where the main conclusions are presented together with figures and tables supporting your conclusions. (2) An oral presentation, of about 15 minutes, to your course colleagues.

The report should be about 10 pages, including figures and tables, and should contain the elementary report constituents:

- Introduction (brief presentation of problem, 1 page)
- Methodology (brief listing of methods, 1 page)
- Data (presentation of your data set with important observations, 1-2 pages)
- Results (4-5 pages)
- Discussion (your results and comparison to other researchers' results, 1 page)

When you are finished with your report, and it has been accepted, then you should produce a PDF file with it, and pack it together with your data set and other important parts of your project (like MATLAB M-files). The idea being that someone else could unpack it and repeat the main steps in your analysis without rewriting everything.

5.2.5 Oral exam

The last part of the course is to do the oral exam. The purpose of this exam is to let you demonstrate the knowledge and skills acquired at the course. You are supposed to answer several questions related to the course material. First, simple questions are given and then continued with questions of a higher degree of difficulty. The white-board is free to use as a tool when answering questions.

6 Evaluation

The following aspects are taken into account when setting final grades.

(1) Student's knowledge shown during oral exam (*40% is the contribution to the final grade*).
(2) Project work: results achieved, the quality of the report and the oral presentation (*40% is the contribution to the final grade*).
(3) Seminars: the presentation quality and activity at seminars given by other students in the course (*20% is the contribution to the final grade*).

7 Course evaluation

After the oral exam there will be an opportunity to give feedback to the teachers involved in the course. This course evaluation is done online and anonymously. The result of the evaluation should offer guidance in the future development and planning of the course. The statistics of the final course evaluation is then made available to all of the students.

References

[1] T. Hastie, R. Tibshirani, J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (Springer Series in Statistics), 2nd Edition, Springer-Verlag, New York, 2009.

[2] R. O. Duda, P. E. Hart, D. G. Stork, Pattern Classification, 2nd Edition, John Wiley & Sons, New York, 2001.

Evaluation matrix

Learning objectives	**Evaluation method**	**Minimum level for grade 3**
Knowledge and understanding		
Describe basic linear as well as nonlinear machine learning algorithms	During the course several lectures and seminars are given to let you acquire brief knowledge of machine learning methods. The acquired knowledge is evaluated both in the written project reports and in the oral exam.	You shall at the oral exam being able to answer most basic to medium-level questions concerning machine learning algorithms. The methods section in the project report shall contain a brief and correct explanation of machine learning methods used.
Knowledge and understanding		
Describe main application areas of machine learning algorithms	The oral exam includes explaining the application areas for different machine learning algorithms. Also, at the seminars you are expected to explain the machine learning methods used in the specific application area.	You are expected to answer correctly on questions concerning which machine learning methods are suitable for a specific application area.
Skills and abilities		
Apply machine learning methods on real world problems	The skills and abilities are evaluated based on written and oral project reports. You are also expected to answer project-related questions during oral exam.	Both regression and classification projects are to be fulfilled and sensible results are to be demonstrated. A very basic discussion of the results should be given in the project report and during oral presentation.
Skills and abilities		
Present scientific results in the learning systems area	Given a research article, you are expected of being able to analyze techniques used for learning system designing in the article, to highlight the main results, advantages and drawbacks of the techniques at the seminars.	You are expected of being able to understand and explain the main ideas presented in the article.
Judgement and approach		
Assess when and which machine learning methods are applicable	During the oral exam several problems will be briefly introduced to you. You are expected of being able to provide reasoning regarding methods applied to solve the tasks.	You are expected of being able to provide a correct basic discussion regarding methods applied.