Summary: machine learning and this course

Summary

In the next few slides we try to get another view into the course contents and their relation to machine learning in general by breaking them up along slightly different lines than the order of the lectures and textbook chapters:

- General ideas and overall principles
- Technical content: mathematics, algorithms and other tools
- Practical issues and applications

General ideas

- We considered the main concepts of machine learning process, such as
  - task
  - model
  - machine learning algorithm as a method of creating a model using data
- This leaves out many work phases that are not usually considered part of machine learnign as such but have crucial importance for the overall result, such as
  - defining the problem
  - getting the data
  - interpreting and utilising results
- In practice these work phases often tend take place as an iterative loop

General ideas (2)

- Regarding the overall basic machine learning process, as well as specific tools and techniques, it should be remembered that similar problems are studied also under other topics (data mining, statistics, pattern recognition, signal processing, . . .)
- There may be different points of view of what is interesting or important, what assumptions are realistic etc. even if the mathematics is similar
- Also many application areas (biology, medicine, ecenomics, . . .) have their own well established tools and practices for analysing data

General ideas (3)

- Also the discussion model complexity and under vs. overfitting is very general
  - we discussed it mainly in context of supervised learning, but similar issues arise also in unsupervised learning (e.g. clustering, density estimation)
  - The basic idea is related to modelling in general:
  - Is your model flexible enough to capture the essentials of the phenomenon you are trying to model?
  - Is your model simple enough that it can be reliably constructed given available resources (such as data and computation time)
  - There is a lot of theoretical work on these issues as well, but they are beyond the scope of our course

Technical content: algorithms

- There are a lot of machine learning algorithms, and the ones included in the course were selected based on several criteria
  - Is it actually a good algorithm?
  - What can we learn from it? What general principles does it help to understand?
  - Can it be explained without getting into too complicated mathematics?
  - Can we do something concrete with it (in the course context, e.g. homework)?

Technical content: algorithms (2)

- There are some state-of-the-art algorithms and more general methods that are included in the textbook but we didn't have time to cover:
  - support vector machine (SVM)
  - Gaussian mixtures and Expectation Maximisation (EM)
  - ensemble learning (boosting, bagging)

  (these would also need a bit more mathematical background than most of the material that we did include)

- Even for the algorithms we did include, the presentation was usually brief. Before actually using the algorithms in practice, you should read up more on their details, variants, limitations etc.

Technical content: model quality etc.

- We spent a lot of time on Bayes optimality
  - despite its apparent simplicity, experience has shown that the idea may be difficult to graps
  - understanding at least the basic setting is required to follow the discussion on ROC curves and probabilistic models
  - Bayes optimality is likely to appear on more advance courses, perhaps also outside machine learning (artificial intelligence, decision theory)
- ROC curve and related notions are very important for discussing model performance in many practical situations
  - for example information retrieval: class distribution unbalanced, false positive vs. false negative have different cost
  - extensive use of coverage plots to explain decision trees, rule sets etc. is not really standard part of basic courses (and perhaps included in the textbook because of its author's particular expertise in the topic)

Technical contents: mathematics

- The lectures did try to explain some mathematical background that students with BSc in computer science often lack
- Main places where we did this were linear regression and multivariate Gaussians
- The choice of course textbook and topics were also done with consideration of typical maths background of beginning MSc students in computer science
- However, if you are going to do your MSc in machine learning, you will need heavier maths than this
- It is recommended that you take a well-rounded selection of courses on maths and statistics, instead of just patching the gaps with cookbook recipes as you go along

Practical issues: experimental methodology

- We did some simple experimentation in homework, but textbook has a whole section on experimentation
- There two points of view into experimental machine learning
  - if you are a practitioner, you want to know how well you are actually doing on a given problem
  - if you are a machine learning researcher, you want to know if your algorithm is better than the others
- In any case, proper methodolody is very important
- Hopefully term such as *confidence interval* and *p-value*, and the basics of *significance testing*, are familiar from statistics courses, but if they are not, you definitely should read up on them before embarking on serious experimentation

Practical issues: tools and techniques

- Cross-validation and using hold-out test sets are basic tools in almost any practical machine learning
- Finding the proper features for describing the data is
  - usually very important
  - often specific to the application domain (so ask the experts)
- Instead of, or in addition to, hand-crafting domain specific features, one can use general well-known feature transformations (polynomials etc.)
  - many algorithms (such as SVM) allow this naturally and efficiently via *kernels*
- We discussed features only very briefly, but again the textbook has a whole chapter on that