Computational Intelligence: Methods and Applications

Lecture 38

Sample questions and remarks to all lectures

Włodzisław Duch
SCE, NTU, Singapore
Google: Duch

L 1

Organization: not much too ask here.

Perhaps I should stress one more time that we will cover a lot of material, some of this in a rather sketchy form, but since this is an advance course I do not want to teach you a lot of details that you will forget after the exam!

The intention of this course is rather to show you many computational intelligence tools, explain how and why they work, present some background knowledge, but leave the theoretical details for self-study those that are interested.

Since your backgrounds and needs are quite diverse you may find different tools useful for your work – these lectures may give you a starting point.

All things that you study at the University are just a starting point for real life ...

L 2

Problems requiring CI

1. Define CI as a branch of science.
2. Give examples of some non-algorithmizable problems
3. How does CI differs from AI, how are they related to cognitive sciences?
4. What pattern recognition problems require CI methods?
5. What pattern recognition problems do not require CI methods?
6. What data mining problems require CI methods?
7. What information selection problems require CI methods?
8. How would you approach the problem of semantic search in the Internet?
9. How to go from observations to problem formulation?

L 2a

CI: problems & inspirations

1. What type of Information Retrieval problems may be solved using CI techniques?
2. Why CI is useful in decision support?
3. What type of control/planning problems may benefit from CI?
4. List some applications of detection of regularities (unsupervised learning).
5. What inspirations are used to solve non-algorithmic problem?
6. Provide some examples of neurobiological (biological, psychological) inspirations used in CI.
7. Provide some examples of biomedical (logical, machine learning, pattern recognition etc.) inspirations used in CI.
8. What can you do if you get some data but do not know what to look for? Where to look for inspirations?

L 3

Histograms and probabilities:

1. What makes a good feature?
2. How to represent objects in feature spaces?
3. How are histograms constructed? How are bins selected and frequencies calculated?
4. Write contingency table, explain quantities involved, including sums of rows and columns.
5. Explain relation between conditional and joint probabilities.
6. Show Bayes formula follows from definition of conditional probability?
7. Do simple examples to see how to calculate posterior probability from other quantities that are easier to calculate.

L 4

Simple visualization.

1. Explain how are 2-D histograms constructed and why it is difficult to use 3D histograms.
2. How to change 3D histograms to make them more legible?
3. What type of information can you find in scatterplots?
4. How will approximately linearly dependent features look like in a scatterplot?
5. Draw a few starplots and explain their construction.
6. How to improve legibility and decrease the number of starplot pictures?
7. Draw lines, cubes, spheres etc. in parallel coordinates representation; what other simple geometrical figures have interesting representation in parallel coordinates.

L 5

EDA & linear transformations.

1. Explain the idea of visualization using Chernoff faces.
2. Can you think of other useful visualizations of this type?
3. Draw the unit circle in Euclidean and Manhattan metric.
4. Draw points that are at the same distance from two fixed points using Euclidean and Manhattan metric; is there anything strange for Manhattan metric? Can you think about other distances function that will have strange behavior?
5. Why is it advantageous to make distances invariant ?
6. Explain the idea of Mahalanobis metric and provide the formula for this metric.
7. Write the formulas for standardization of data – what is gained by standardization? Will it always improve results?
8. Show that standardized data have zero mean and unit variance.

L 6

... and Principal Component Analysis.

1. Write the formula for covariance matrix and explain the meaning of diagonal and off-diagonal terms.
2. Prove that Mahalanobis metric is invariant to the linear transformations.
3. What is the most important property of principal components?
4. How are principal components calculated?
5. What is the interpretation of the first principal component?
6. Calculate variance of principal components.
7. What could PCA be used for?
8. What are the disadvantages of PCA?
9. Draw some examples of data distributions in which PCA may not be useful.

L 7

.. and Discriminant Component Analysis.

1. Why increasing separation of cluster centers after projection is not the best idea?
2. How to orthogonalize all data to several directions?
3. Define within-class scatter and between-class scatter matrices.
4. Define the Fisher criterion in terms of these matrices.
5. Find the vector that maximizes Fisher criterion.
6. Explain what type of projections are obtained using Fisher coordinates.
7. How to generate second and higher discriminant components?
8. How to use such criterion for more than 2 classes? Write the formula and think if there is a simpler way to do it.
9. What is a lattice projection?

L 8

CI: Projection Pursuit & Independent Component Analysis.

1. What is Projection Pursuit?
2. Give some examples of interesting projection pursuit indices.
3. Write the formula for kurtosis (for zero mean distributions) and explain what does it measure and why it is useful.
4. How to make two variables uncorrelated?
5. When two variables are statistically independent? How does independence and lack or correlation differ?
6. Write the projection pursuit condition for the first principal component and explain how to get higher components.
7. If several unknown signal sources are linearly mixed using unknown coefficients and only the results are given, how can the sources and the mixing coefficients be calculated?

L 9

CI: Self-Organized Mappings

1. Give some examples of topographically organized maps in the brain – why topographic organization is useful?
2. Explain the idea behind SOM, major steps in the algorithm.
3. Sketch the SOM algorithm, explain the variables and functions used, and explain how they affect the algorithm.
4. What happens when 2-D data is represented by processors arranged on a line? How will the values of their parameter vectors be arranged in the 2-D space?
5. When do the maps get distorted, and why it may be dangerous to learn too many things too quickly?

L 10

Self-Organized Mappings and GCS

1. What are the properties of SOM algorithm?
2. Estimate the complexity of the SOM algorithm.
3. Define quantization error, what is its meaning?
4. Why fixed SOM maps may represent some data in a wrong way?
5. How can SOM be improved?
6. Describe the growing cell structure (GCS) algorithm, explain the concepts involved and possible variants of this algorithm.
7. Explain the concepts of Voronoi set and Delaunay triangulation.
8. List some SOM applications.

L 11

Multi-Dimensional Scaling & SOM

1. What are disadvantages of the SOM algorithm for visualization?
2. Write some topographical distortion measures.
3. Present the MDS algorithm; how many parameters are optimized?
4. Explain the difference between metric and non-metric MDS algorithms.
5. How will any 3 vertices of a simplex in *d*-dimensions be represented by MDS in two dimensions?
6. Compare SOM with MDS; when will they be difficult to use?
7. Explain how semantic map may be constructed from sentences.
8. Why SOM and MDS semantic maps preserve natural hierarchical classification?
9. How do semantic maps capture the meaning of words?

L 12

Bayesian decisions

1. What does learning from data means?
2. Why are we not satisfied with conditional probability but want to have the posterior probability to make a decision?
3. Write all normalization conditions that involve a priori, conditional and posterior probabilities.
4. Show that Bayesian decisions minimize average errors.
5. Formulate Bayesian decisions using likelihood ratio.
6. Explain what the confusion matrix is, and what the risk matrix is.
7. How is the total risk of using a decision procedure calculated?
8. For a two class problem and a binary variable *x* with joint probability P(C,*x*) write the confusion matrix of MAP classifier.
9. If the test is 95% accurate for both types of errors (C1⇔C2) but prior probability is only 0.1% what is the posterior probability?

L 13

Bayesian risks & naive approach

1. Explain the structure of confusion matrix is a general case (including rejections and default classes).
2. Present Bayesian decision procedure.
3. What is a discriminating function? What transformations can be applied to it?
4. List discriminating functions for 4 main types of classifiers. Explain which one is the best and when they may be used.
5. What type of discrimination function is obtained assuming class-conditional distributions defined by single Gaussian functions?
6. Under what conditions linear discriminating functions are obtained?
7. Under what conditions Bayesian classification is reduced to the nearest prototype rule?
8. When is Naive Bayesian approach expected to work well?

L 14

Bias-variance tradeoff

1. How are linear models trained on data and why not to use Bayesian probability estimations directly?
2. What is meant by overfitting the data and why does it occur?
3. What are the sources of errors in learning from data?
4. Formulate the goal of learning from data. What should we try to approximate and how does it relate to the observed values?
5. Write the decomposition of the MSE to noise and the remaining terms, explain the meaning of each term.
6. What is meant by bias? Provide formula and explain it.
7. Write the bias-variance decomposition of the MSE error, explain all terms in it (you may skip the detail of the derivation).
8. Provide some examples of the bias-variance tradeoff.

L 15

Model selection

1. Explain why is crossvalidation used, how does it work and what information does it give us?
2. What is stratified crossvalidation?
3. How is CV called when the number of folds is equal to the number of data samples? When will you use it this way? How does it differ from typical k-fold CV?
4. What is the curse of dimensionality and when does it arise?
5. How does the curse of dimensionality relates to the bias?
6. What is the simplicity/accuracy tradeoff? How to generate different models along these lines?
7. What is confidence vs. rejection rate tradeoff? How to generate different models along these lines?

L 16

Model evaluation & ROC

1. Write the confusion matrix for a two-class plus rejection case, explain how to obtain it, and what is the meaning of its entries.
2. List all quantities that may be derived from the confusion matrix. Explain what do sensitivity, specificity and precision measure. How to increase confidence in the model?
3. How to introduce costs? Write the appropriate error function.
4. Explain how are lift charts used.
5. How is the ROC curve constructed? Show examples of good and bad ROC curves.
6. Draw sample ROC curves for two classifiers that give only yes/no answers (ex. logical rules), and compare them.
7. Prove using MAP rule for binary feature that the ROC curve is defined by one point at $(S_+, 1-S_-)$, and that AUC = $(S_+ + S_-)/2$ is identical along $S_+ + (1-S_-)$=const line.

L 17

WEKA/Yale intro & knowledge from simplest trees

1. What do you think of WEKA? Have you played with this?
2. What do you think of Yale? Have you played with this?
3. What do you think of GhostMiner? Have you played with this?
4. What categories of models are in Yale/Weka?
5. What are lazy methods, functions and meta models?
6. What type of knowledge representation is used in Weka/Yale? What are the limitations of such representation?
7. Describe ZeroR method; why is it included in Weka/Yale?
8. Sketch the 1R tree algorithm.
9. How are continuous values handled in 1R?

L 18

Decision trees in WEKA and GM

1. Describe 4 essential elements of decision tree algorithm.
2. How to calculate the amount of the information gained by splitting a node?
3. Why most compact trees are preferred?
4. How to find compact trees? What type of search techniques decision tree use?
5. How do decision borders of the univariate trees look like?
6. How do decision borders look for multivariate trees?
7. Explain why decision trees are not stable.
8. Show that trees may overfit the data, achieving classification rates that are below that of the majority classifier.

....

Recess week: time to think it over ...

1. How am I doing so far ?
2. What have I learned for the half semester?
3. What do I want to achieve during the coming year?

And for the most ambitious students ...

What is the meaning of life, universe and everything?
And what has 42 to do with it ?

L 19

Pruning of decision trees

1. Sketch general Top-down Iterative DT algorithm.
2. How to avoid overfitting of data by trees?
3. What is meant by pruning?
4. What type of pruning are used?
5. Sketch the ID3 algorithm.
6. What tests and split criteria does C4.5 algorithm use?
7. Describe the CHAID decision tree, what criterion is used here and why does it work?
8. What split and stop criteria are used by CART trees?

L 20

SSV trees

1. Define the SSV tree split criterion.
2. What type of tests are used in SSV?
3. How rules of different complexity may be created?
4. Explain the tradeoff between complexity and accuracy of rules.
5. How can rules of optimal complexity be found?
6. How can simplest rules be found?
7. Describe advantages of decision trees.
8. Describe disadvantages of decision trees.

L 21

Linear discrimination, linear machines

1. How to create trees for regression? What type of approximation will they provide?
2. Can the concept “the majority is for it” be learned from data using univariate decision trees?
3. Draw graphical representation of how discrimination function is calculated.
4. Calculate the distance from discrimination plane.
5. What problems appear in K-class situation when hyperplanes are used for discrimination?
6. How to construct linear machine? What decision borders will it create?
7. How linear discrimination may be used to create non-linear decision borders?

L 22

Linear discrimination - variants

1. Formulate linear discrimination problem and present it in a form of linear equations.
2. Write LDA solution using pseudoinverse matrix; what properties does this matrix have and how to calculate it?
3. How may LDA equations be solved using the perceptron algorithm? Write the criterion and the update formula.
4. Write the Fisher criterion, explain how it is used in classification and how is it related to LDA.
5. Write the QDA, or Quadratic Discriminant Analysis function, and explain how to estimate its parameters.
6. Describe the idea behind RDA, or Regularized Discrimination Analysis.

L 23

Logistic discrimination and support vectors

1. What is the basic assumption of Logistic DA?
2. Write the likelihood for the two-class case, and explain how does the classification rule obtained from it differs from the standard linear discriminant.
3. Why maximum likelihood formulation is used in this case instead of the linear equation solution?
4. How to define margin of classification and measure it for linear discriminants?
5. Formulate LDA problem with margin maximization for separable data.
6. Write the Lagrange form of the constrained minimization problem, write the discriminant function using Lagrange multipliers, and show that Lagrangian depends only on the inner products of support vectors.

L 24

SVM in non-linear case

1. Formulate the linear discrimination problem for non-separable data.
2. Explain the meaning of the C coefficient in SVM and how does it influence the results.
3. How to optimize C? What other parameters should be set by the user in SVM program?
4. What are support vectors and which vectors will be usually selected as SVs?
5. Describe mechanical analogy for SVM.
6. What idea is the sequential minimal optimization (SMO) algorithm implemented in WEKA based on?

L 25

Kernel methods.

1. Use quadratic kernel in 2-D and write down the equivalent functions defining features in the extended feature space.
2. Give examples of some useful kernels.
3. Write covariance matrix after transformation of vectors to a new basis, assuming zero means after transformation.
4. Write discriminant function using kernel form.
5. Derive the kernel PCA equation.
6. For 2D with several Gaussian clusters, how will the contours of the constant kernel PCA components look like?
7. How many PCA components may be extracted from the d-dimensional data? How many kernel PCA components?
8. Write the criterion that is optimized for the Fisher kernel method.

L 26

Density estimation and EM algorithm.

1. If you could find good PDF, what could you do with it?
2. How may PDFs be used to discuss psychological phenomena?
3. How to assign values to features if the value is missing in some vectors? When will it work well?
4. Describe the idea behind likelihood maximization and write the formula for log likelihood; why logarithmic version is used?
5. Given frequency of observation of 3 types of birds, and some parametric formula for probability, write the log-likelihood.
6. Describe general steps of the expectation-maximization algorithm – when is such algorithm useful?

L 27

Expectation Maximization algorithm

1. What is “a generative model” and what is it useful for?
2. Explain the idea of likelihood maximization.
3. What alternatives for parameter optimization exist?
4. Describe properties of the EM algorithm.
5. Sketch the EM algorithm including missing feature values in some data vectors.
6. List some applications of the EM algorithm for likelihood maximization.

L 28

Non-parametric density modeling

1. Explain why histograms provide a useful approximation for the density.
2. Explain difficulties in non-parametric density estimation.
3. Describe the Parzen windows algorithm in 1-dimensional case.
4. Describe the Parzen windows algorithm in d-dimensional case.
5. What are the factors that determine the degree of smoothing in the Parzen windows algorithm?

L 29

Approximation theory, RBF and SFN networks

1. Decompose basis function into the activation and output functions and give a few examples of useful functions.
2. Define radial basis functions and show a few examples of RBFs.
3. What type of contours of constant values may be created using mixed activation functions?
4. Define separable functions and relate them to the Naive Bayes method.
5. Why polynomial functions are not useful in high D cases? Why Gaussian or sigmoidal functions with weighted activations are used?
6. Draw and explain the structure of RBF network, write the simplest solution to determine the linear parameters.
7. Explain the structure of SBF networks, explain how they may be used to find classical logical rules and fuzzy rules.

L 30

Neurofuzzy FSM and covering algorithms

1. How would you initialize a separable function network?
2. Write the Spearman rank order correlation coefficient and explain when it should be used.
3. Describe the idea of covering algorithms using PRISM algorithm as an example.
4. How can PRISM be applied to data with continuous features?
5. What determines the complexity of solutions for covering algorithms?

L 31

Combinatorial reasoning, partial observations

1. Think about electric circuit problem: why is it difficult to write a program that would answer any question related to changes in such system? What other similar problems that are difficult to solve come to your mind?
2. How to solve electric circuit problem using expert system approach?
3. List all true and false facts related to changes of the 3 variables bound by A=f(B,C) additive, multiplicative or inverse additive relation; draw the feature space representation of the true facts.
4. Write a function that describes Ohm's law in a qualitative way.
5. Describe how to use density estimation as a heuristic in searches for possible changes of parameters or subsets of discrete parameters.
6. How to generalize it to a larger subsets of correlated variables?
7. Which variable should be considered first during the search?

L 32

Nearest neighbor methods

1. Prove that MAP decision is equivalent to the nearest neighbor majority rule.
2. What kind of Voronoi tesselations does kNN method create?
3. How do they depend on the *k* value?
4. Why is it easy to calculate the leave-one-out estimate using kNN and difficult using most other methods?
5. What is the complexity of kNN and how to decrease it?
6. Describe properties of kNN; why is it stable while LDA or decision trees are not?
7. Draw the unit circle for different Minkovsky exponents.
8. Describe feature weighting and instance (neighbor) weighting.
9. List some advantages and disadvantages of kNN.

L 33

Information theory

1. What is the idea behind the decision table algorithm?
2. Explain why information is measured using the – p log p Shannon formula, and why $\log_2$ are used.
3. Which Gaussian distribution contains more information, the one with a large variance or the one with a small variance?
4. Define joint information, conditional information, find and prove how they are related.
5. Define mutual information and prove that it is equal to the Kullback-Leibler divergence.
6. Decompose joint information into a sum of mutual and conditional information, or into information in each variable and mutual information between them.

L 34

IT applications and feature selection

1. List some application of information theory.
2. Prove that Mutual Information = Information Gain when binary feature is used to split the data in decision tree.
3. How to use mutual information for data visualization?
4. What are some methods to focus on relevant information in a large database?
5. What is the difference between ranking and selection?
6. What are filters and what are wrappers? Write general algorithm for use of filters and wrappers.
7. Sketch an algorithm for feedforward feature selection using the Naive Bayes algorithm.

L 35

Feature selection and discretization

1. What is the complexity of filters and what is it for wrappers?
2. Given one binary feature and a priori probabilities for two classes, write joint probability matrix using only two unique parameters.
3. How to compare the relevance of two binary features?
4. Write the formula for Laplace corrections to conditional probability estimation. Why are they sometimes used?
5. How to calculate relevance indices for continuous features?
6. List 4 types of discretization algorithms.
7. How may decision trees be used for discretization?
8. Write the formula for mutual information for discretized features.
9. What other indices may be used to estimate relevance of features?

L 36

Meta-learning: committees, sampling and bootstrap

1. What are advantages and disadvantages of committees?
2. Describe the sources of models for committees, what makes them different and how should they differ?
3. Explain basic committee voting procedures.
4. What is bootstrap? Where does the magic number 0.632 comes from?
5. What is bagging?
6. What is boosting?
7. Present the AdaBoost M1 algorithm and explain why this particular weighting has been used? What are the alternatives?
8. Describe the idea of stacking algorithm; how is it related to committees and to other classification models?

L 37

Review and summary

What should one do to make progress in science,
and become rich and famous?

Should I add more reading material after each lecture?

More step-by-step exercises using software packages?

Perhaps lab/tutorial would be a good idea?