Jerzy Stefanowski
Institute of Computing Science,
Poznań University of Technology
Course Title: Data Mining and Analysis
June 4th, 2010

REMARKS FOR PREPARING TO THE EXAM (FIRST ATTEMPT "PRZEDTERMIN" - MONDAY JUNE 14)

This note should give you some useful hints on topics which are the basis for preparing the exam on the Data Mining and Data Analysis course for Software Engineering specialization at Computer Science. However, you should treat them as a kind of warnings what seems to be the most important in Magdalena Deckert and my teaching. These remarks are not definitely the precise questions, etc. You should expect that real questions may be formulated in a slightly different way (although basing on these topics). Remember - besides these topics you should read material / notes from my lectures and labs. It is good to look in some additional materials (books, web presentations, etc.) to clarify what you have not understand or should be somehow studied more carefully.

> You should expect that the exam could be a list of questions (to be answered by 1-2 sentences or by a necessary drawing), test questions (multiple choice from the given list a,..,d) and simple tasks - problems for calculation sth.
>
> We will prepare a special sheet of paper with these questions and problems. You can use your own empty pages if you need more free space. Moreover, you can use your own calculator (but a classical one, not mobile phone, notebook, etc.).

Below you have the list of main topics:

1. Definition of KDD process. Name steps of this process. Shortly explain the aim of the given step.
2. Difference between predictive and descriptive goals of data mining.
3. Types of attributes in data table - explain difference between nominal, ordinal, interval and ratio-scale.
4. Simple statistical analysis of dependency between nominal attributes - you have to know how to calculate chi-2 test (given a table of observed values) and V-Crammer coefficient.
5. Differences between V-Crammer coefficient and other strength measures (e.g. Φ-Yule).
6. Correlation analysis - interpretation of ρ Pearson coefficient. Relation to determination R2 coefficient in a simple regression (y and x variables).
7. Interpret parameters in a linear regression model.
8. Methods of statistical diagnostics of a linear regression model. We will give you numbers from a software report (like Excel or Statistica) and ask to name its element and interpret the value -- to decide whether the regression is statistically valid or not.
9. How to define determination R2 coefficient basing on residual analysis (SST, SSR and SSE). When we are using corrected coefficient.
10. Write hypothesis and explain them in local and global tests for linear regression.
11. Other evaluation measures for predictions (where in the historical data confidence intervals are greater) and what are mean square error?

12. What Conditions Must Hold in order for us to Legitimately Apply Regression Techniques? Just name them in points.
13. How to interpret figures from the residual analysis (understand differences between the good or bad residual plots - e.g. standardized residuals should be located in a kind of „belt" shape – approx. equally distributed around the expected 0.).
14. How can we identify outliers using analysis of standardized residuals?
15. Simple task - transform given non-linear function into linear one (used in non-linear regression).
16. General rules for selecting variables in la inear multi-dimensional regression.
17. Differences between regression and classification trees.
18. Pre-processing data: List names of methods for dealing with missing attribute values.
19. Discretization methods: simple calculation tasks for equal width or equal frequency methods.
20. Entropy-based discretization - you should know the method and introduce proper values of probabilities into the formula - the final calculations are not necessary (as we have logarithms).
21. List components of typical algorithms for attribute selections in pre-processing.
22. Differences between filter and wrapper approaches - draw figures and name the elements.
23. Evaluation criteria for comparing classifiers - list and name them.
24. Task: calculation of basic measures (total accuracy, sensitivity, specificity, ..) basing on a confusion matrix.
25. Explain when to use hold-out, cross-validation and other statistical techniques for calculating classifier evaluation measures.
26. Decision tree - how to use ID3 algorithm.
27. Simple calculating task for choosing attribute to nodes of the tree - simple examples without calculating precise entropy.
28. Properties of the entropy measure - in particular for a binary data / when it is equal to min or max values.
29. Differences between info-gain and gain-ration / and when to use them?
30. How could we handle numerical attributes in decision tree induction.
31. Define overfitting of the data - in particular for trees.
32. Typical criteria for forward pruning.
33. The cost-complexity approach to tree post-pruning.
34. Task: simple calculation of a single rule induction from nominal data (PRISM from a lecture).
35. List typical evaluation functions for rules (simple criteria, vs, weighted formulas).
36. Differenced between classification strategies with ordered and unordered rule sets.
37. Multiple classifiers - explain the concept of bagging (see laboratory notes).
38. Why should we normalize / standardize attributes before calculated distance measures?
39. Differences between main distance measures - make a drawing showing their differences.
40. List in points k-means or AHC algorithm.

41. Medoids vs. centroids in partitioning algorithms (as k-means).
42. Interpret differences between linkage techniques in AHC (distances between clusters).
43. Criteria for evaluating good clustering.
44. Density based approaches to clustering - main ideas in DBSCAN.
45. What is the aim of SOM / multi-dimensional projections?
46. How to calculate frequent itemset from a set of transactions - the first steps of Apriori algorithm.
47. Pruning property of Apriori algorithm - a simple example.
48. Calculate basic rule evaluation measures for association rules - simple examples for calculating support or confidence.
49. Multiple-level Association Rules - what is the idea behind them?
50. Why could we use lift against confidence measure in evaluation of association rules?
51. Simple calculation task - to check whether a sequence supports sequential patterns.

ADDITIONAL SOURCES FOR LEARNING ...

See pdf files of slides from my lectures -- see e-moodle or my Web-page.

Go through Magdalena Deckert materials from labs (in particular calculations

Daniel Larose: Discovering Knowledge in Data: An Introduction to Data Mining. Wiley, 2005.
Odkrywanie wiedzy z danych, Larose D. (polskie tłumaczenie), PWN, Warszawa, 2006.

Any good course book on Statistics (including testing hypothesis and regression) - In Polish I could recommend for instance:

A. Aczel: Statystyka w zarządzaniu, PWN 2000.

or Koronacki J., Ćwik J., Statystyczne systemy uczące się, WNT 2005 - it is better to read II edition.

You can also consider reading:

Statystyka dla studentów kierunków technicznych i przyrodniczych, Koronacki Jacek, Mielniczuk Jan, WNT, 2001

As to induction of decision trees, rules, Bayes classifier or neural networks you could read:

Krawiec K, Stefanowski J., Uczenie maszynowe i sieci neuronowe, Wyd. PP, 2004.

In Internet you can find nice sets of slides accompanying the well known course books:

Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, Jian Pei, Morgan Kaufmann, 2005.

Tan, Steinbach, Kumar: Introduction to Data Mining.

WEKA teaching materials (see web page of WEKA project at Waikato University)

Gregory Piatetsky – Shapiro: KDD and Data Mining Course -- see also material in KDDNuggets service.