# Study hints for the Exam

- The exam will consist of 14 questions.
- Two different types of questions: multiple choice questions and short answer questions.
- Eight of the questions are multiple choice questions where only one of five possible answers is correct. Each multiple choice question is worth 1 mark.
- Answer the multiple choice questions on the back of page 1 on the answer booklet that will be provided to you.
- The remaining 6 questions are short answer questions. Start answering each short answer question on a new page!
- Each short answer question has three to six sub questions.
- Marks for each exam question (and sub-question) are specified.
- Answers to all question can be abstracted from the lecture notes, and laboratory exercises.
- Exam duration is 3 hours. The time constraints **will** be a factor in the exam.
- You are permitted to take one approved calculator into the exam. It is recommended that you bring a calculator with you as you will need to perform some calculations..
- Everything covered during the lectures and labs can be covered by the exam.
- A sample exam question of type "multiple choice" may look as follows:
  - Which of the following is not a main problem with the Self-Organizing Map clustering algorithms

    (a) Computational time complexity.<br>
    (b) Setting of training parameters.<br>
    (c) Dealing with unknown attributes.<br>
    (d) Dealing with very small training sets<br>
    (e) Obtain clusters that correspond to the class membership of the data.

  The correct answer for this question would be (a) because the computational time complexity of the SOM is linear and hence, this is not a main concern. Option (b) cannot be correct because the SOM requires a range of training parameters (such as the learning rate, size of the SOM, choice of neighborhood function, number of training iterations, etc.) that need to be set before training can commence. It is not a-priori clear which set of training parameters would lead to the best classification result. In fact, you experienced this problem during one of the labs. Option (c) cannot be right because the Euclidean distance measure used in the training algorithm does not allow for missing attribute values. Option (d) cannot be correct because the SOM cannot cluster data if there are not enough data to form any clusters. Option (e) must be wrong because the SOM is trained in absence of knowledge of a class membership (an unsupervised training algorithm)
- A sample exam question of type "short answer" may look as follows:

  (A) Draw the architecture of a multi-layer perceptron network which features 2 nodes in the input layer, 2 nodes in a single hidden layer, and 1 node in the output layer. Assign values from within the range [-1;1] to the weights which connect the neurons such that the output is 1 if the input is (1,0) and -1 if the input is (1,1). Assume that the transfer function associated with every neuron is f(x)=x (a linear transfer function).

  There are many possible answers to this question. For example, a simple answer could look as follows:

More sample exam questions are given in Appendix A at the end of this document.

**Best way to prepare for the exam:**

- Keep social networking, surfing, TV at a minimum.
- Create a study friendly environment (away from bed and TV)
- Reward yourself every two hours of good studies.
- Ability to solve assignment tasks yourself.
- Discuss the results of a marked assignment with your tutor and seek clarification.
- Read the lecture slides, and the sections in the book surrounding the material presented in the slides.
- Study in groups, discuss with group members.

**Proposed strategy to answer exam questions.**

- Read the questions carefully!
- Use Brainstorming.
- Start to answer a questions on a new sheet of paper. Answer may span across several sheets.
- Questions can be answered in any order.
- Start by answering the easiest question first, the hardest question last.

**Best of luck for your exam.**

## APPENDIX A:

Sample exam questions (a collection of "old" exam questions).
Note: Multiple choice questions consisted of four options to choose from. Due to changes in school policy, the multiple choice questions in your exam will have five options to chose from.
Note: Sample answers will not be provided. Reason: At least one of the questions listed here will appear in the final exam.

**Question 1: (1 mark)**
**From which of the following area does Data Mining draw its ideas from?**
(a) Pattern Recognition
(b) Database systems
(c) Statistics
(d) Machine Learning
(e) All of the above

**Question 2: (1 mark)**
**What is the main purpose of Knowledge Discovery in Data Mining?**
(a) To develop algorithms that make a machine learn from data.
(b) To remove noise and outliers from data.
(c) To simulate the ability of the human brain to learn and understand.
(d) To process large amounts of data.
(e) To find hidden relations amongst a given set of data.

**Question 3: (1 mark)**
**What is not normally a data mining task?**
(a) Modelling and Clustering.
(b) Modelling and Classification.
(c) Sequential Pattern Discovery and Contextual Pattern Discovery.
(d) Deviation detection and regression.
(e) Learning and understanding.

**Question 4: (1 mark)**
**What is the main problem with K-means clustering algorithm?**
(a) Scalability problem.
(b) Complexity problem.
(c) Large cluster problems.
(d) Initialization problem.
(e) Data Mining problem.

**Question 5: Data (3+2 marks)**
(A) Explain with the help of diagrams the following terms noise, outliers, missing values, duplicate data. Explain what can be done about each of these problems.
(B) Explain the term “Dimension Reduction”, and give an example on how this can be achieved.

**Question 6: Classification (3+2 marks)**
(A) Explain with the help of a diagram the terms “Curse of Dimensionality”, “overfitting”, and “generalization”.
(B) Draw the architecture of a multi-layer perceptron network which features 2 nodes in the input layer, 1 node in a single hidden layer, and 1 node in the output layer. Assign values from within the range [-1;1] to the weights which connect the neurons such that the output is 1 if the input is (0,1) and -1 if the input is (1,1). Assume that the transfer function associated with every neuron is f(x)=x (a linear transfer function).

**Question 7: Clustering (2+2+2 marks)**
(A) What is clustering? What is the role of clustering in Data Mining and Knowledge Discovery?
(B) Explain the difference between K-means and LVQ.
(C) Explain the term “topology preserving properties” of a self-organizing map.

**Question 8 (5 marks)**

(a) What is the difference between "classification" and "numeric prediction" problems? (2 marks)

(b) Suppose that a machine prints a status message every minute. Each status message consists of one of the letters A,B,C,D,E or F depending on the machine's internal diagnostics. You suspect that there is a pattern that could predict the occurrence of the next status message being D based on the sequence the previous four (4) messages.

You have available a large sequential stream of historical status messages produced by the machine over several days. How would you arrange or transform this data stream, creating appropriate attribute names where required, such that a decision tree learning task could be used to investigate this hypothesis. Further demonstrate this with the first 35 minutes of this status stream included below.

B,D,E,A,D,E,E,B,C,B,A,E,E,B,D,B,D,B,F,A,A,B,B,D,D,B,E,E,A,D,E,C,E,E,F, …

(3 marks)

**Question 9 (8 marks)**

Consider the selection of training examples shown in the following table. This data corresponds to nine decisions (*Action*) that were made in different situations. The decisions recorded here reflects the experts judgment of whether to ***release*** or ***hold*** the deployment of a specialised tracking balloon that will be able to monitor and record certain atmospheric weather conditions.

| **Observation** | **Overcast** | **StrongWinds** | **UV_ratio** | ***Action*** |
|---|---|---|---|---|
| 1 | yes | yes | 1.0 | *release* |
| 2 | yes | yes | 6.0 | *release* |
| 3 | yes | no | 5.0 | *hold* |
| 4 | no | no | 4.0 | *release* |
| 5 | no | yes | 7.0 | *hold* |
| 6 | no | yes | 3.0 | *hold* |
| 7 | no | no | 8.0 | *hold* |
| 8 | yes | no | 7.0 | *release* |
| 9 | no | yes | 5.0 | *hold* |

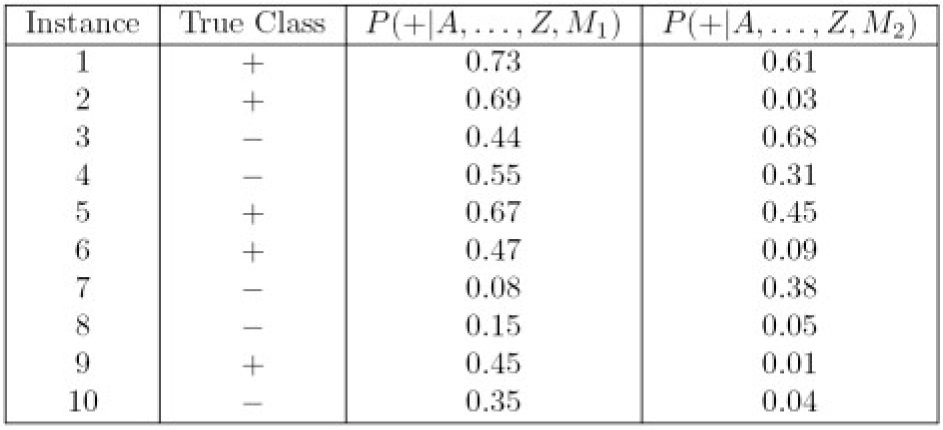
(a) What is the entropy from this data with respect to the decision for releasing balloons? (2 marks)

(b) What are the information gains for the conditions of *"Overcast"* and *"StrongWinds"* relative to this selection of training examples? (3 marks)

(c) For the continuous attribute, *"UV_ratio"* compute the information gain for every possible split, and thus determine which attribute from *Overcast, StrongWinds and UV_ratio* which provides the best split according to the information gain? (3 marks)

**Question 10 (7 marks)**

You are asked to evaluate the performance of two classification models, M1 and M2. The test set used contains 26 binary attributes, labeled as A through Z. The following table shows the models performance as posterior probabilities obtained by applying the models to the test set. (Only the posterior probabilities for the positive class are shown). As this is a two-class problem, P(−) = 1−P(+) and P(−|A,...,Z) = 1−P(+|A,...,Z). Assume that you are only interested in detecting instances from the positive class.



| Instance | True Class | $P(+\|A,\ldots,Z,M_1)$ | $P(+\|A,\ldots,Z,M_2)$ |
|---|---|---|---|
| 1 | + | 0.73 | 0.61 |
| 2 | + | 0.69 | 0.03 |
| 3 | − | 0.44 | 0.68 |
| 4 | − | 0.55 | 0.31 |
| 5 | + | 0.67 | 0.45 |
| 6 | + | 0.47 | 0.09 |
| 7 | − | 0.08 | 0.38 |
| 8 | − | 0.15 | 0.05 |
| 9 | + | 0.45 | 0.01 |
| 10 | − | 0.35 | 0.04 |

Plot the ROC curve for both M1 and M2. (You should plot them on the same graph.) Which model do you think is better? Explain your reasons.

END OF DOCUMENT