Department of Computer Science
Hong Kong Baptist University
COMP7560

Project – 22 Marks
Due: 2/December. Submission at 18:00 (6:00pm) in form of a hard-copy of the slides. One submission per group. Submission must be made to the lecturers. No late submissions!

This is a group project. Each project can have a minimum of 3 and a maximum of 5 members. Each student is to sign up for one project no later than 9am on 27/November by entering his/her name to the project list which is on display at the door of office RRS715. Groups are formed on a first come first served basis.
Your task is prepare a presentation for one of the topics listed below. The presentation time is limited to 15 minutes for each group plus 3 minutes for questioning time. Each group is to prepare a set of slides (around 15 to 20 slides will be sufficient). The title slide must list the topic of the presentation, the full name and student numbers of all students in the group. The second slide must contain a table which summarizes the contribution of each group member to the project and presentation. The last slide must draw a conclusion on the topic chosen. All other slides must address the given topic.

The presentations are to be given on Tuesday, 2/December between 6:00pm and 9:30pm. The exact time slot will be assigned to each group. **All students are expected to attend the presentations.** Each student is to chose one of the following ten topics.

Topic 1
Title: "An introduction to Hidden Markov Models for Data Mining"
Difficulty: medium
Description of the task: Hidden Markov Models (HMM) are capable of detecting pattern in a given set of data. This renders HMMs particularly useful for Knowledge Discovery tasks for which "hidden" patterns are to be discovered. You task is to provide a description of the HMM training algorithm and provide a view on the suitability of HMM for data mining applications.
Reading: All the required information can be extracted from the paper "HMM.pdf" which is available on the subject's web site.

Topic 2
Title: "On the relationship between Recursive MLP and MLP"
Difficulty: medium
Description of the task: Recursive MLP, alternatively known as BPTS, is a generalisation of the MLP algorithm. Your task is to explain how a Recursive MLP processes data, and why a Recursive MLP contains the standard MLP as a special case. The expected outcome of this presentation is that the audience will have developed a better understanding to how BPTS works, and how BPTS is related to standard error-backpropagation.
Reading: All the required information can be extracted from the papers "RMLP.pdf" and "NNIntro.pdf" which are available on the subject's web site.

Topic 3

Title: “An overview to Auto-associative memories”

Difficulty: medium-low

Description of the task: The main purpose of auto-associative memories is to store information in a fault tolerant fashion inside a neural network. Your task is to present an overview and a comparison of the two auto-associative memory approaches discussed in the lectures. The expected outcome of this presentation will be to enable the audience to more thoroughly comprehend the two different mechanisms.

Reading: All the required information can be extracted from the paper “ANNTutorial.pdf” and from the lecture notes which are available on the subject's web site, and from http://www.scholarpedia.org/article/Hopfield_network, and http://www.heatonresearch.com/articles/2/page4.html.

Topic 4

Title: “Inside MLP”

Difficulty: medium-low

Description of the task: MLP is said to be a “general approximator”. This means that an MLP can learn any problem to any arbitrary precision as long as the problem can be described by a contentiously differentiable function. Your task is to describe the “inner workings” of MLP. The expected outcome is that the audience will have developed an understanding to how an MLP learns, and how information is encoded by the MLP.

Reading: All the required information can be extracted from “NNIntro.pdf” and from the lecture nodes which are available on the subject's web site.

Topic 5

Title: “An alternate view on association rule mining”

Difficulty: easy-medium

Description of the task: Association rule mining is is a method for discovering interesting relations between variables in large databases. The most popular algorithm to association rule mining is known as a-priori rule as was proposed by Agrawal et.al. In real world applications, association rule mining is concerned with processing many million transactions each of which may contain up to several 10,000 items. Your task is to provide a brief overview to the apriori algorithm with special focus on the scalability to real world applications. The expected outcome is that the audience will have developed an understanding to why the apriori rule is successful on a combinatorial task which would would otherwise prohibitifly expensive to compute.

Reading: All the required information can be extracted from the paper “Agrawal.pdf” which is available on the subject's web site.

Topic 6

Title: “The effects of unbalanced data on supervised machine learning”

Difficulty: medium

Description of the task: Unbalanced datasets can severely influence the ability of a machine learning method to learn anything useful. Your task is to describe the problem, and to provide some insight to why an unbalanced training set negatively affects supervised and unsupervised training algorithms. Give one example which shows that such problems are in fact a desirable property.

Reading: The required information can be deducted from the paper “NNIntro.pdf”, “SOM.pdf”, and the lecture notes which is available on the subject's web site.

Topic 7

Title: “Handling unbalanced training sets”

Difficulty: easy-medium

Description of the task: Unbalanced datasets can severely influence the ability of a machine learning method to learn anything useful. Your task is to describe several approaches which can be taken to minimize the effect of unbalanced datasets. The expected outcome is that the audience will have developed an understanding to the various approaches that can be taken to minimize problems that arise out of unbalanced training data sets.

Reading: None. Discussion with the lecturer is encouraged.

Topic 8

Title: “On the relationship between K-means and LVQ”

Difficulty: Easy

Description of the task: The purpose of both K-means and LVQ is to compute prototypes for a given set of training data. Your task is to briefly describe the two training algorithms, and to offer a comparison between the two algorithms. The expected outcome is that the audience will be able to understand how the two algorithms are related (and how they differ).

Reading: All the required information can be extracted from the paper “LVQ.pdf” and the lecture notes which is available on the subject's web site.

Topic 9

Title: “How SOM achieves topology preserving mapping”

Difficulty: easy-medium

Description of the task: Self-Organizing Maps are very popular for data mining tasks requiring the visualization of high dimensional data. This is because SOMs maintain the topology on the display space. This property is not explicit in the training algorithm, but rather an effect of the training algorithm. Your task is to describe why the SOM is able to achieve the topology preserving property amongst the input data. The expected outcome is that the audience understands why a SOM achieves topology preservation.

Reading: All the required information can be extracted from the paper “SOM.pdf” and the lecture notes which are available on the subject's web site.

Topic 10

Title: “One the relationship between SOM-SD and SOM”

Difficulty: medium

Description of the task: SOM-SD is a generalisation of the SOM algorithm. Your task is to explain how a SOM-SD processes data, why the SOM-SD is limited to processing tree-structured data, and why SOM-SD the standard SOM as a special case. The expected outcome of this presentation is that the audience will have developed a better understanding to how SOM-SD works, and how SOM-SD is related to the standard SOM.

Reading: All the required information can be extracted from the papers “SOM.pdf” and “SOM-SD.pdf” which are available on the subject's web site.