

The Chinese University of Hong Kong
Department of Systems Engineering and Engineering
Management

**SEEM5770/ECLT5840 Open Systems/Electronic
Commerce**

2013-14 Assignment 3: Data Mining

The assignment is due at 11:00pm on November 29. This assignment counts for 20% of the unit's assessment. You must submit to the assignment box **B10** (ERB 5th floor). You need to put down your student id and your name.

The questions in this assignment ask for explanations. Answer all the questions. The explanations should be concise descriptions of your understanding. You should clearly show all the formulas you use, and all the steps followed to get your answers.¹

1 Classification

Table 1 provided by J. Cendrowska in UCI Machine Learning Repository² gives a simple dataset for testing the fitting for contact lenses. There are four attributes: ID, Age, SP (Spectacle Prescription), A (Astigmatic), and TPR (Tear Production Rate), plus the Class label. There are 22 records. For the 1st Age attribute, there are three values: 1 (young), 2 (pre-presbyopic), and 3 (presbyopic). For the 2nd SP attribute, there are two values: 1 (myope) and 2 (hypermetrope). For the 3rd A attribute, there are two values: 1 (no) and 2 (yes). For the 4th TPR attribute, there are two values: 1 (reduce) and 2 (normal). For the Class label, there are three classes: Class-1 (The patient should be fitted with hard contact lenses), Class-2 (The patient should be fitted with soft contact lenses), and Class-3 (3: The patient should not be fitted with contact lenses). In addition to the training dataset (Table 1), there are two testing cases given in Table 2.

¹Departmental Guideline for Plagiarism (Department of Systems Engineering and Engineering Management): If a student is found plagiarizing, his/her case will be reported to the Department Examination Panel. If the case is proven after deliberation, the student will automatically fail the course in which he/she committed plagiarism. The definition of plagiarism includes copying of the whole or parts of written assignments, programming exercises, reports, quiz papers, mid-term examinations and final examinations. The penalty will apply to both the one who copies the work and the one whose work is being copied, unless the latter can prove his/her work has been copied unwittingly. Furthermore, inclusion of others' works or results without citation in assignments and reports is also regarded as plagiarism with similar penalty to the offender. A student caught plagiarizing during tests or examinations will be reported to the Faculty office and appropriate disciplinary authorities for further action, in addition to failing the course.

²<http://archive.ics.uci.edu/ml>

Age	SP	A	TPR	Class
1	1	1	2	2
1	1	2	1	3
1	1	2	2	1
1	2	1	1	3
1	2	1	2	2
1	2	2	1	3
1	2	2	2	1
2	1	1	1	3
2	1	1	2	2
2	1	2	1	3
2	1	2	2	1
2	2	1	1	3
2	2	1	2	2
2	2	2	1	3
2	2	2	2	3
3	1	1	2	3
3	1	2	1	3
3	1	2	2	1
3	2	1	1	3
3	2	1	2	2
3	2	2	1	3
3	2	2	2	3

Table 1: Contact Lenses: Training Dataset

Age	SP	A	TPR
1	1	1	1
3	1	2	2

Table 2: Contact Lenses: Two Testing Cases

- Q-1:** (Decision Tree) Find the best attribute at the root node for building a decision tree based (1) Gini Index and (2) Entropy, using the multi-way split³ as the attribute test conditions. Detailed steps on your calculations must be given.
- Q-2:** (Naive Bayesian Classification) With the training dataset in Table 1, predict each of the two testing cases given in Table 2. Give details on your calculations step-by-step.
- Q-3:** (kNN: k -Nearest Neighbor Classification) Consider the following distance function $d(x, y) = \sqrt{\sum_{i=1}^4 (x_i - y_i)^2}$, where i means the i -th attribute in the two tables ($1 \leq i \leq 4$). With the training dataset in Table 1, predict each of the two testing cases given in Table 2. Give details on your calculations step-by-step.

³Refer to the lecture note on data classification part-1, slide 25.

TID	Items Bought
1	{Milk, Beer, Diapers}
2	{Bread, Butter, Milk}
3	{Milk, Diapers, Cookies}
4	{Bread, Butter, Cookies}
5	{Beer, Cookies, Diapers}
6	{Milk, Diapers, Bread, Butter}
7	{Bread, Butter, Diapers}
8	{Beer, Diapers}
9	{Milk, Diapers, Bread, Butter}
10	{Beer, Cookies}

Table 3: Market Basket Transactions

2 Association Rules

Consider the dataset given in Table 3, answer the following questions.

1. Show the supports for all 1-itemsets and 2-itemsets, and give a 2-itemset that has the largest support. (*To understand that any frequent i -itemsets can be done based on $(i-1)$ -itemsets for a given min support threshold.*)
2. Find a pair of 2-itemsets, X and Y such that the rules $\{X\} \rightarrow \{Y\}$ and $\{Y\} \rightarrow \{X\}$ have the same confidence. Hint: X and Y are a 2-itemset.
3. Explain the procedure of merging a frequent 3-itemsets using two frequent 2-itemsets, using an example taken from Table 3. And show that the procedure does not miss any candidates. Hint: You can use any thresholds.