

INFS 755 (Fall 2008)

Assignment 1 (Due on 09/22/2008)

This is an individual assignment. Please ensure that assignment is submitted in class in hard copy before start class. No late submissions allowed. The first part of the assignment are a few questions from Chapter 1 and Chapter 2. The second part gives you a feel for the KDD process and dealing with data. It lets you get acquainted to WEKA.

Part 1 (50 points) (Questions borrowed from the Tan et. al. book)

1. Discuss whether or not each of the following activities is a data-mining task. [2 points each – 10 points]
 - a. Sorting a student database based on the student identification number.
 - b. Predicting the future stock price of a company using historical records.
 - c. Monitoring the heart rate of a patient for abnormalities.
 - d. Monitoring seismic waves for earthquake activities.
 - e. Extracting the frequencies of a sound wave.

2. Classify the following attributes as binary, discrete, or continuous. Also classify them as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some cases may have more than one interpretation, so briefly, indicate your reasoning if you think there may be some ambiguity. [3 points each – 15 points]
 - a. Brightness as measured by a light meter.
 - b. Angles as measured in degrees between 0 degrees and 360 degrees.
 - c. Bronze, Silver, and Gold medals as awarded at the Olympics.
 - d. ISBN numbers for books.
 - e. Military Rank.
 - f. Coat check number.

3. Which of the following quantities is likely to show more temporal autocorrelation: daily rainfall or daily temperature? Why? [5 points]

4. Discuss why a document-term matrix is an example of a data set that has asymmetric discrete or asymmetric continuous features.[5 points]
5. For the following vectors, x and y , calculate the indicated similarity or distance measures.[3 points each = 9 points]
 - a. $x = (1, 1, 1, 1)$, $y = (2, 2, 2, 2)$ - cosine, correlation, and Euclidean.
 - b. $x = (0, 1, 0, 1)$, $y = (1, 0, 1, 0)$ - cosine, correlation, Euclidean, and Jaccard.
 - c. $x = (0, -1, 0, 1)$, $y = (1, 0, -1, 0)$ - cosine, correlation, Euclidean
6. Given a similarity measure with values in the interval $[0, 1]$ describe two ways to transform this similarity value into a dissimilarity value in the interval $[0, \infty]$. [6 points]

Part 2 (50 points) : The KDD Process in Weka

This assignment was borrowed from the TCSS 555A Data Mining Class taught at the University of Washington, Tacoma Branch.

Assignment preparation

This assignment will be using **Weka** data mining tool. **Weka** is an open source Java development environment for data mining from the **University of Waikato in New Zealand**. It can be downloaded freely from <http://www.cs.waikato.ac.nz/ml/weka/>.

Heart disease datasets

The dataset studied is the **heart disease** dataset from **UCI repository**. Two different datasets are provided: **heart-h.arff** (Hungarian data), and **heart-c.arff** (Cleveland data). These datasets describe factors of heart disease. Both these data sets are available to you on the assignment page.

The **data mining project goal** is to better understand the risk factors for heart disease, as represented in the 14th attribute: **num** (<50 means no disease, and values <50-1 to <50-4 represent increasing levels of heart disease).

The **question** on which this machine learning study concentrates is whether it is possible to predict heart disease from the other known data about a patient. The **data mining** task of choice to answer this question will be classification/prediction, and several different algorithms will be used to find which one provides the best predictive power. However this exercise focuses on the various aspects of the KDD process.

1. Data preparation- integration

We want to merge the two datasets into one, in a step called data integration. Revise *arff* notation from the tutorial, which is *Weka* data representation language. Answer the following questions:

- a. Define what data integration means. (in your own words)
- b. Is there an *entity identification* or *schema integration* problem in this dataset? If yes, how to fix it?
- c. Is there a *redundancy* problem in this dataset? If yes, how to fix it?
- d. Are there *data value conflicts* in this dataset? If yes, how to fix it?
- e. Integrate the two datasets into one single dataset, which will be used as a starting point for the next questions, and load it in the *Explorer*. How many instances do you have? How many attributes? (You could do this using Excel or spreadsheet programs. First, save your individual files as “csv” files in weka, Open them in a spreadsheet viewing program. Copy the rows from one file to another. Save the merged file (csv). Open it in weka and save it as “csv”. Take care of the above questions. Think about rectifying potential problems.
- f. Paste a screenshot of the *Explorer* window.

2. Descriptive data summarization

Before preprocessing the data, an important step is to get acquainted with the data – also called *data understanding*.

- a. Stay in the *Preprocess* tab for now. Study for example the *age* attribute. What is its *mean*? Its *standard deviation*? Its *min* and *max*?
- b. Provide the *five-number summary* of this attribute. Is this figure provided in *Weka*? This is min, max, median, lower 25% quartile and upper 25% quartile.
- c. Specify which attributes are numeric, which are ordinal, and which are categorical/nominal.
- d. Interpret the graphic showing in the lower right corner of the *Explorer*. How can you name this graphic? What do the red and blue colors mean (pay attention to the pop-up messages that appear when dragging the mouse over the graphic)? What does this graphic represent?
- e. Visualize all the attributes in graphic format. Paste a screenshot.
- f. Comment on what you learn from these graphics.
- g. Switch to the *Visualize* tab. By selecting the maximum jitter, and looking at the *num* column – the last one – can you determine which attributes seem to be the most linked to heart disease? Paste the *boxplot* representing the attribute you find the most predictive of heart disease (Y) as a function of *num* (X).
- h. Does any pair of different attributes seem correlated?

3. Data preparation – selection

The datasets studied have already been processed by selecting a subset of attributes relevant for the data mining project.

- a. From the documentation provided in the dataset, how many attributes were originally in these datasets?
- b. With *Weka*, attribute selection can be achieved either from the specific *Select attributes* tab, or within *Preprocess* tab. List the different options in *Weka* for selecting attributes, with a short explanation about the corresponding method.

4. Data preparation - cleaning

Data cleaning deals with such defaults of real-world data as incompleteness, noise, and inconsistencies. In *Weka*, data cleaning can be accomplished by applying *filters* to the data in the *Preprocess* tab.

- a. **Missing values.** List the methods seen in class for dealing with missing values, and which *Weka filters* implement them – if available. Remove the missing values with the method of your choice, explaining which filter you are using and why you make this choice. If a filter is not available for your method of choice, develop a new one that you add to the available filters as a Java class. (that should be exciting and fun ... send me an email if you plan to do this)
- b. **Noisy data.** List the methods seen in class for dealing with noisy data, and which *Weka filters* implement them – if available.
- c. Save the cleaned dataset into *heart-cleaned.arff*, and paste here a screenshot showing at least the first 10 rows of this dataset – with all the columns.

5. Data preparation - transformation

1. Among the different data transformation techniques, explore those available through the *Weka Filters*. Stay in the *Preprocess* tab for now. Study the following data transformation only:
 - a. **Attribute construction** – for example adding an attribute representing the sum of two other ones. Which *Weka filter* permits to do this?
 - b. **Normalize** an attribute. Which *Weka filter* permits to do this? Can this filter perform Min-max normalization? Z-score normalization? Decimal normalization? Provide detailed information about how to perform these in *Weka*.
 - c. **Normalize** all real attributes in the dataset using the method of your choice – state which one you choose.
 - d. Save the normalized dataset into *heart-normal.arff*, and paste here a screenshot showing at least the first 10 rows of this dataset – with all the columns.

6. Data preparation- reduction

Often, data mining datasets are too large to process directly. Data reduction techniques are used to preprocess the data. Once the data mining project has been successful on these reduced data, the larger dataset can be processed too.

- a. Stay in the *Preprocess* tab for now. Beside attribute selection, a reduction method is to select rows from a dataset. This is called sampling. How to perform sampling with *Weka filters*? Can it perform the two main methods: *Simple Random Sample Without Replacement*, and *Simple Random Sample With Replacement*?