Thatcher Rickertsen

tor0002

09/05/18

Data Mining HW 01

1.
    a. What is Data Mining?
        - Data mining is simply the extraction of data. More formally, it is the process of gathering interesting parts of large sets of data in order to generate new sets of information from it.
    b. Explain knowledge discovery steps in a Framework like Web Mining.
        - The knowledge discovery (KDD) process involves several steps to transfer data into information:
            1. Selection
                - Where data is chosen.
            2. Preprocessing
                - Where the necessary adjustments to data are performed in order to make it processable.
            3. Transformation/Exploration
                - Where necessary data is identified and picked out to be further evaluated.
            4. Data Mining
                - Where patterns or other interesting data is sorted out of the preprocessed data.
            5. Interpretation
                - Where the data is made into an understandable format so that people can make decisions about it.
    c. Write the Data Preprocessing, Data Mining, and Data Postprocessing inclusive process in a knowledge discovery process from a Machine Learning and Statistics approach.
        - This can be more discretely summarized compared to the KDD process:
            1. Data Preprocessing
                - Contains all of the steps that involve preparing data for the data mining step, such as getting rid of useless data and normalization.
            2. Data Mining
                - Where patterns or other interesting data is sorted out of the preprocessed data.
            3. Data Postprocessing
                - Involves the evaluation of patterns discovered within the previous step and the processes by which you interpret that data and transform it into relevant information.
2. Write 5 data mining functions and give an example application for each one of them.
    a. Medical data mining – Helps with increasing the accuracy of diagnosis prediction.

   b. Credit data mining – Determining whether or not a group of people is likely to default on a loan, requiring companies to charge them more interest.
   c. TV watching data mining – Determining whether or not an audience would be well suited to a new show of a certain genre or whether viewership would fall.
   d. Demographic data mining – How likely certain people of a given demographic are likely to do a certain thing, such as how many 40-50 year old’s regularly buy new technology.
   e. Use statistics data mining – This could be many things, such as finding out what features users use most in an application and choosing to expand that feature rather than one that is used less.

3. Explain the difference between noise and outlier:
   a. Noise – This is data that is not helpful at all, at least to what is currently being studied. It could also be referred to as clutter. This tends to cloud data sets, making it harder to mine the data that is actually important to a given study.
   b. Outlier – This would be a data point that is completely out of the ordinary for a given data set. An example of this would be an 80 year old who is highly proficient in technology despite many of their peers being technologically incompetent.