Machine Learning
Assignment 1

Ida Bagus Dwi Satria Kusuma
1301140297

September 9, 2017

1. What is machine learning?

   (a) Machine learning usually refers to the changes in system that perform tasks associated with artificial intelligence. [1]

   (b) Machine learning is the subfield of computer science that, according to Arthur Samuel, gives computer the ability to learn without being explicitly programmed. [2]

2. Give an example of machine learning implementations!

   An example of implementations of machine learning is the automated classification made by Fayyad et al. The classification was designed whit two purposes: first, to classify objects in DPOSS (Digitized Second Palomar Observatory Sky Survey) to the faintest limits of the data; second, to fully generalize to future classification efforst, including classification of galaxies by morphology and improving the existing DPOSS star/galaxy classifiers once a larger volume of data are in hand [3]

3. There are 3 models of machine learning that can be distinguished according to their main intuition. Explain the 3 models, and give an example of algorithms/methods that belongs to each model!

   According to their main intuition, machine learning can be distinguished into 3 models, which is geometric models, probabilistic models, and logical models. The details about these models will be explained below. [4]

(a) **Geometric Models**

Geometric Models are models that constructed directly in instance space, using geometric concept such as lines, planes and distances. Geometric models also exploit geometric notions (such as planes, translations and rotations, and distance) and are simple, powerful and allow many variations with little effort. An example of geometric models' method is Manhattan Distance.

(b) **Probabilistic Models**

Probabilistic Models are models that will learn the probability distirbution over the data set that given and used. The statistical approach is to assume that there is some underlying random process that generates the values for variables that measured (target variables), according to a well-defined but unknown probability distribution. An example of probabilistic models' method is naive-Bayes.

(c) **Logical Models**

Logical Models are models that can be easily translated into rules that are understandable by humans. This models have have rules (called feature tree) that are easily organised in tree structured. An example of logical models's method or algorithm is the decision tree.

4. What are supervised learning, unsupervised learning, semi-supervised learning and reinforcement learning?

(a) **Supervised learning**

Supervised learning is a type of learning where we have input variables and an output variable, and use one or more algorithms to learn the mapping function from the input to the output. The goal is to apporoximate the mapping function so well that when we have new input, we can predict the output for that input. The process of an algorithm learning from the training dataset can be thought as a teacher supervising the learning process. We know the correct answers, the algorithm iteratively makes prediction on the training data and is corrected by the teacher. Learning stops when the algorithm achieves an acceptable level of performance.

Supervised learning problem can be further grouped into regression and classification problems.[5]

(b) **Unsupervised learning**

Distinct with supervised learning, unsupervised learning is where we only have input data and no corresponding output variables. The goal of unsupervised learning is to model the underlying structure or distribution in the data in order to learn more about the data. Unlike supervised learning, there is no correct answers and there is no teacher. Algorithms are left to their own devises to discover and present the interesting structure in the data. Unsupervised learning problems can be further grouped into clustering and association problems. [5]

(c) **Semi-supervised learning**

Semi-supervised learning is where we have large amount of input data and only some of the data is labeled. These problems sit in between both supervised and unspervised learning. [5]

(d) **Reinforcement learning**

Reinforcement learning occurs when we present the algorithm with examples that lack labels, but we can accompany an example with positive or negative feedback according to the solution the algorithm proposes. Reinforcement learning is connected to applications for which the algorithm must make decisions and the decisions bear consequences. It is just like learning by trial and error. [6]

5. What are dataset, a data object and an attribute?

   A dataset is a collection of data objects. A data object is described a set of atrributes. An Attribute describes one aspect of data object.

6. What is data preprocessing?

   Data preprocessing is a data mining technique that involes transforming raw data into an understandable format. Data preprocessing is a method to resolving issues of real-world data, where real-world data often incomplete, inconsistent, or lacking in certain behaviors or trends, and is likely to contain many errors. [7]

7. Many machine learning algorithms use measures of similarity or dissimilarity between data objects. Explain differences between similarity and dissimilarity measures!

   (a) Numerical measure of the degree to which two objects in 'similarity' are alike, but in 'dissimilarity' are different.

   (b) 'Similarity' is higher for objects that are alike, but 'dissimilarity' is higher for objects that are different.

   (c) 'Simmilarity' is typically between 0(no similarity) and 1 (completely similar), but 'dissimilarity' is typically between 0 (no difference) and infinity/positive number other than 0 (completely different). [8]

8. There are many similarity measures, explain 3 of them!

   (a) **Jaccard Index**, is a satistic used for comparing the similarity and diversity of sample sets:

   $$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|} \tag{1}$$

   (b) **SMC(Simple Matching Coefficient)** is a statistic used for comparing the similarity and diversity of sample sets. SMC defined as :

   $$SMC = \frac{\text{number of matching attributes}}{\text{number of attributes}} = \frac{M_{00} + M_{11}}{M_{00} + M_{01} + M_{10} + M_{11}} \tag{2}$$

   (c) **Overlap Coefficient** is a similarity measure related to the Jaccard index that measures the overlap between two sets, and is defined as the size of the intersection divided by the smaller of the size of the two sets:

   $$\text{overlap}(X,Y) = \frac{|X \cap Y|}{\min(|X|, |Y|)} \tag{3}$$

9. There are many dissimilarity measures, explain 3 of them!

   (a) **Jaccard distance**, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained

by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

$$d_j(A, B) = 1 - J(A, B) = \frac{|A \cup B| - |A \cap B|}{|A \cup B|} \tag{4}$$

(b) **SMC(Simple Matching Coefficient)** which measures dissimilarity between sample sets, is given by 1 - SMC.

(c) **Lee distance** is distance between two string $x_1x_2...x_n and y_1y_2...y_n$ of equal length n of the $q-ary$ alphabet $0, 1, ..., q-1$ of size $q \geq 2$. Defined as

$$\sum_{i=1}^{n} \min(|x_i - y_i|, q - |x_i - y_i|) \tag{5}$$

10. What is the goal of data visualization?

    The goal of data visualization is to communicate information clearly and efficiently via satistical graphics, plots, and information graphics. Effective visualization helps users analyze and reason about data and evidence.[9]