**MIS419**

**Business Data Mining**

**Assignment 1**

**Problem 1 (40 pts).** ***In your own words***, provide definitions for the following terminologies and provide at least one example (in no more than 3 lines):

1. Unsupervised learning

Doesn’t have target value just describe the data

Example: Customer classes

1. Supervised learning

Splitting the data into historical data and it has target value

Example: Determine who is more likely to buy a particular product

1. Business data mining

Companies that use data mining to develop their business

Example: In supermarkets, if two items has high correlation between them, then they know where to place them

1. Prediction in data mining

Prediction represented in testing data to predict the future

Example: 1R method

1. Model accuracy

A percentage that shows how accurate the model is

Example: when you compare between the predicted and the actual result.

1. One-rule method

Method that help you predict something

Example: Based on historical data and observations

1. Normal distribution

When the data are clustered around the mean

Example: when the student grades are around (C, C+, B) this mean that the grades are normally distributed

1. Overfitting and underfitting

Overfitting: When the gap between the error of the test data and training data is huge

Example: Complex model

Underfitting: When the gap between the error of the test data and training data is small

Example: Linear model

**Problem 2 (40 pts).** The following is a hypothetical example representing information on past customers of a specific online store.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Name** | **Address available?** | **Age** | **Gender** | **Income** | **Purchased** |
| Alex D. | Yes | 1 | Male | 25 | 0 |
| Stacy M. | Yes | 2 | Female | 38 | 0 |
| John C. | Yes | 3 | Male | 66 | 1 |
| Pete M. | Yes | 1 | Male | 25 | 0 |
| Cheng H. | Yes | 1 | Male | 35 | 1 |
| Levine M. | Yes | 1 | Female | 29 | 1 |
| Frank L. | Yes | 2 | Male | 49 | 0 |
| Tim M. | Yes | 2 | Male | 40 | 1 |
| Rachel K. | No | 3 | Female | 60 | 0 |
| Marry B. | No | 2 | Female | 53 | 0 |
| Sara D. | No | 1 | Female | 42 | 1 |
| Bryan A. | No | 2 | Male | 38 | 1 |
| Brittany W. | No | 3 | Female | 48 | 0 |
| Omar C. | No | 3 | Male | 50 | 1 |
| Martin L. | No | 1 | Male | 22 | 1 |

Answer the following questions:

1. [5 pts] How many records and how many attributes are there?

There are 15 records and 6 attributes

1. [10 pts] Specify the attribute type for each of the attributes above.

Name = Nominal Address Available? = Binary Age = Ratio Gender = Binary Income = Ratio Purchased = Binary

1. [30 pts] Suppose we want to predict whether or not a customer will purchase our product (the target variable is “Purchased”). Use the 1-rule method to generate rules for the attributes Age and Gender.

|  |  |  |  |
| --- | --- | --- | --- |
| Attribute | Rules | Error | Total error |
| Age | Age1= 4 Purchased / 2 Didn’t | 2/6 | 6/15= 0.4 |
| Age2= 2 Purchased / 3 Didn’t | 2/5 |
| Age3= 2 Purchased / 2 Didn’t | 2/4 |
| Gender | Male= 6 Purchased / 3 Didn’t | 3/9 | 5/15= 0.33 |
|  | Female= 2 Purchased / 4 Didn’t | 2/6 |

1. [5 pts] Using the results from c), which attribute better predicts purchase? And why?

Gender is better because it has less error

**Problem 3 (20 pts).** Specify whether the following is a **useful** or an **obvious** pattern, and state why?

* The sales increase when the store introduced deep discounts.

Obvious, Because the lower the price the more demand it will get

* A customer’s recurring transaction with above than average value takes place every beginning of the month.

Obvious, Customers withdraw their salary in the beginning of the month

* Productivity of some employees do not change when salary was increased.

Useful, Salary does not have a relation with productivity

* Low number of customers for a telecommunication company at rural places.

Obvious, people that lives their dos not use telecommunication as much

* A bank customer receives an amount of 50,000 to his account. The customer has never received an amount greater than 10,000 for the past 5 years.

Useful, it’s useful for the bank to keep an eye on this customer

* Street traffic are heavy every first Saturday of the month between 10pm and 11pm

Useful, this information is more useful for marketing.

* Customers whose age is between 28 and 34 buy more frequently during weekdays from a fast food restaurant that claims to be healthy

Useful, it’s useful for the fast food restaurant so they target people in this age.

* Customers wearing their watch on right hand buy more from the store.

Useful, it’s useful for the store so they take care of people that are wearing watches on their right hand