Exercise 1

1. What is meant by classification?

Classification is assigning labels to unlabeled data. In a given set of data each instance can belong to a class. The process of assigning the correct label/class to the data is called classification.

2. Explain the two phases of classification.

Learning Phase-

In learning phase classifier is built on training data to learn which instance belongs to which class. This is also induction phase.

Querying phase-

In querying phase on test data model is applied and class/labels are predicted. This is also called deduction phase.

3. What is the learning set? Why is this set split into training and test sets?

The set of data/instances on which the classifier models is trained to learned os called learning set. Training data is used to make the model learn which class/bales a typical instance belongs to and then on the test data the accuracy of the model is tested in order to verify if it works.

4. What is a target attribute? What properties must it have in order to be used for

classification?

Target attribute is the class/label the model is trying to predict. It must a discrete value and should be mutually exclusive for the instances for classification.

5. Open Orange1 and create a new workflow. Add a “Datasets” widget to the canvas

and select the “Attrition - Train” dataset. Inspect the data with a “Data Table”

widget. Classify the following attributes as nominal, ordinal, interval or ratio: Age,

BusinessTravel, Department, DistanceFromHome, Education, Gender, JobLevel and

MonthlyIncome

Nominal

Business Travel

Department

Gender

Ordinal

Age,

DistancefromHome,

Education,

Monthly Income

Interval

Ratio

6. A patient goes to see a doctor. Suppose the doctor performs a test with 99% reliability,

i.e., 99% of people who are sick test positive and 99% of the healthy people test negative.

The doctor knows that only 1% of the people in the country are sick. If a patient has

tested positive, what is the probability he is sick?

Very little like, .0009 as per Bayes theorem.

7. How does a Naïve Bayes classifier work? Why is it “naive”?

Naïve Bayes makes a naïve assumption that each feature individually makes impact on the target. If a data set has 2 features then the value of feature 1 doesn’t impact the outcome on target of feature 1. Person shall play golf if day is sunny. Person shall play golf if air is humid. Sunny and humid are independent and makes impact on target independently.

8. How does the Naïve Bayes classifier handle numeric attributes?

It usese it with mean value or standard deviation

9. Explain the zero-frequency problem. How can it be avoided?

Zero frequency problem is when a feature values doesn’t occur in the set. While calculating probability if for a feature the value doesn’t exists then the complete probability turn to be zero.

To solve this smoothing is used, i.e. adding a dummy/virtual example to the dataset to prevent making complete value zero.

10. For the dataset provided in Table 1, compute the conditional probabilities of all attribute

values, given the target attribute Defaulted Borrower (DB=Y and DB=N).

Classify the following instance according to the the Naïve Bayes Classifier (a) without

handling zero conditional probabilities and (b) by using Laplace estimate to calculate

conditional probabilities:

**X** = (Home Owner = N, Marital Status = Married, Annual Income = 90)

1http://orange.biolab.si/

Y= 3/10, n = 4/10

HO (Y)= 1 HO(N)= 4/7

MS(Married)(Y) = 0 n= 4/7

Without smoothing instance is going to be classified as 0. In laplace can add 1 virtual instance.