**Spring 2021**

**SISU MBA-FT1 BI Take Home Final Exam**

**Professor Han Zhang**

**100 Points**

**Submission Deadline: Sunday, July 25, 2021, 11:59pm.** **For each 15-minute interval your exam is late we will deduct 10 points from your exam grade**

**Please Note:**This exam has three questions. Please answer all three questions. Note that you have to answer all parts of all questions.

Please work alone. The exam is open book and open notes. You can refer to any material you need to for completing this exam (but not your classmates or anyone else!). If you are referring to any material, please provide the appropriate citation. DO NOT DISCUSS THIS EXAM WITH ANY OF YOUR CLASSMATES OR ANYONE ELSE.

* Please name your exam as follows: lastname.firstname-FT1-BI-Final
* Please use Microsoft Word for this exam (please don’t use PDF).
* Please submit your exam via the school system
* Please write your name on each page of your exam
* Also, please use Times New Roman with a font size of 11 points or higher. Please use one inch margins. Single Spacing is allowed.
* There is no page limit for the take-home example. However, please try your best to provide concise answers.
* If you have any questions, please e-mail us at [hanzhang.gt@gmail.com](mailto:hanzhang.gt@gmail.com) or on WeChat.
* OPEN BOOK, OPEN NOTES. You are expected to turn in your own work. No assistance of any sort may be sought from any other individual. Please provide references if you cite other material.

BEST OF LUCK!

**Question (1) (33 points)**

Now consider a real-world dataset, vote.arff, which gives the votes of 435 U.S. congressmen on 16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. This is a purely nominal dataset with some missing values (corresponding to abstentions). (You automatically downloaded this dataset when you downloaded the WEKA software.) Please use WEKA ***J48*** and use the “training set” to build a decision tree to predict **party affiliation** based on voting patterns. (Note: Apart from treating missing value as an attribute value on its own, in the case of the J48 classifier any split on an attribute with missing value will be done with weights proportional to frequencies of the observed non-missing values.)

A. (6 points). Please discuss the Confusion Matrix in detail. What does this Confusion Matrix mean (please explain the numbers in the Confusion Matrix)?

B. (6 points). How would this instance be classified using the decision tree?

physician-fee-freeze = y, synfuels-corporation-cutback = y, mx-missile = n,

adoption-of-the-budget-resolution = y, anti-satellite-test-ban = n.

C. (5 points). Please copy and paste the decision tree in your answer.

D. (8 points). Assume in your decision tree, you got the following leave:

physician-fee-freeze = n: democrat (253.41/3.75).

What does “democrat (253.41/3.75)” mean? Please explain it in detail.

E. (8 points). Why did you get decimal numbers rather than integers in your decision tree? Please explain.

**Question (2) (35 points)**

Now consider a real-world dataset, vote.arff, which gives the votes of 435 U.S. congressmen on 16 key issues gathered in the mid-1980s, and also includes their party affiliation as a binary attribute. This is a purely nominal dataset with some missing values (corresponding to abstentions). (You automatically downloaded this dataset when you downloaded the WEKA software.) Please use WEKA ***Apriori*** association-rule mining to seek interesting associations.

A. (6 points). What is the cutoff of confidence used in selecting the top 10 rules (based on the default setting)?

B. (6 points). Based on the default output, what is the support for this item set?

adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y

C. (8 points). What is the rule support and rule confidence for the following rule (please use the default setting except *numRules*):

adoption-of-the-budget-resolution=y aid-to-nicaraguan-contras=y ==> physician-fee-freeze=n

D. (8 points). It is interesting to see that none of the rules in the default output involve Class = republican. Why do you think that is?

E. (7 points). One person seeks to explain the following rule as antecedent and consequent.

Rule:

adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y 198 ==> Class=democrat 198

In his/her opinion: “adoption-of-the-budget-resolution=y physician-fee-freeze=n aid-to-nicaraguan-contras=y” is the cause, and “Class=democrat” is the effect. Therefore, he/she thinks that the above rule reveals causation between the antecedent and consequent. Is that correct? Why?

**Question (3) (32 points)**

Now consider a real-world dataset, diabetes.arff, which contains measurements for 768 female subjects from the Pima Indian population, all aged 21 years and above (you automatically downloaded this dataset when you downloaded the WEKA software). The Pima Indian population are based near Phoenix, Arizona (USA). They have been heavily studied since 1965 on account of high rates of diabetes. The attributes are as follows, and I list them here since they weren’t explicitly stated in the version of the data that came with Weka and I only found them after a bit of digging online:

preg - the number of times the subject had been pregnant

plas - the concentration of blood plasma glucose (two hours after drinking a glucose solution)

pres - diastolic blood pressure in mmHg

skin - triceps skin fold thickness in mm

insu - serum insulin (two hours after drinking glucose solution)

mass - body mass index ((weight/height)\*\*2)

pedi - ‘diabetes pedigree function’ (a measurement I didn’t quite understand but it relates to the extent to which an individual has some kind of hereditary or genetic risk of diabetes higher than the norm)

age - in years

class – categorical (or nominal) variable: tested positive for diabetes; tested negative for diabetes

Note: K-means cluster analysis is designed for continuous (numeric) variables. Some data mining tools limit cluster analysis only to coninuous (numeric) variables. WEKA’s K-means cluster analysis can process nominal (categorical) variables. In WEKA, for nominal attributes, distance is set to 1 if values are different (or if one or both are missing), 0 if they are equal. However, please keep it in mind that the K-means cluster analysis based on nominal data are rather bad since k-means is all about means, but what is the mean of “bread”, “milk” and “banana”? The diabetes database contains all numeric attributes except the *class*.

Load *diabetes.arff* in WEKA. Use *Cluster* panel, choose *SimpleKMeans*, and select *two* clusters for your cluster analysis. There are nine attributes in the *diabetes.arff* table. Before you run cluster analysis, please choose to ignore the attribute of *Class*.

Please use “*Use training set*” in the Cluster mode to answer the following four questions: A, B, C and D.

* 1. (3 points) What is the number of iterations?
  2. (6 points) What does the number of iterations mean? Please explain it.
  3. (6 points) Please report the percentage of each cluster out of 768 instances.
  4. (4 points) What is the number of the *Within cluster sum of squared errors* in your cluster analysis (please round to two decimal places)?

You answered the previous four questions by using “*Use training set*” in the Cluster mode. Now please choose “*Classes to clusters evaluation*” and answer the following three questions: E, F and G.

* 1. (5 points) What is the percentage for the incorrectly clustered instances?
  2. (4 points) How many instances, that were tested negative in reality, are mistakely clustered into “Tested\_Posittive” cluster in the cluster analysis?
  3. (4 points) How many instances, that were tested potive in reality, are mistakely clustered into “Tested\_Negative” cluster in the cluster analysis?