ACT4311 Homework 1

Deadline:2020-03-29 23:59:59

(Accounting for 5% of overall grade)

Q1. [6 points] Our task is to recognize species of animals, such as cat or dog, based on their pictures. Then the pictures of animals with the species of the animals in the picture are ___ data for our task?

A) Unstructured and unlabeled data
B) Unstructured and labeled data
C) Structured and unlabeled data
D) Structured and labeled data

Q2. [6 points] If feature F1 represents the most favorite movie name of a student. Then which of the following statement is true?

A) Feature F1 is an example of nominal variable.
B) Feature F1 is an example of ordinal variable.
C) Feature F1 is an example of numerical variable.
D) All above

Q3. [6 points] We want know what are the products that are usually bought together in a super market, then this task should be:

A) Supervised learning
B) Unsupervised learning
C) Semi-supervised learning
D) All above

Q4. [6 points] If we need to accomplish following 4 tasks as follows:

1. What is the expected GDP in the coming year?
2. What is the specie of an animal shown in a given picture?
3. Which top X customers should I target with a special offer (X can be any positive integer number)?
4. Are there any interesting natural groupings of my customers?
Please select most accurate categories for above tasks:

A) 1-Regression, 2-Ranking/scoring, 3-Classification, 4-Clustering
B) 1- Ranking/scoring, 2- Regression, 3- Clustering, 4- Classification
C) 1-Classification, 2-Ranking/scoring, 3- Clustering, 4- Regression
D) 1-Regression, 2- Classification, 3- Ranking/scoring, 4-Clustering

Q5. [6 points] Entropy is:

A) a measure of impurity
B) a measure of information gain
C) a measure of correlation between numeric variables
D) None of above

Q6. [6 points] We now have a collection of 500 words with positive sentiment, and then we need to check 1 million new words and determine whether these new words are positive words, which kind of Python collection you should use to store and maintain these positive words?

A) List
B) Tuple
C) Dict
D) Set

Q7. [6 points] Which of the following data types CANNOT be used as the key of dictionary in Python?

A) Integer
B) String
C) List
D) Tuple

Q8. [6 points] Below are the 12 actual values of target variable in the training data.

[0,0,0,1,1,1,1,1,0,0,1,1]

What is the entropy of the target variable?

A) 5/12*log(5/12) - 7/12*log(7/12)

B) 5/12*log(5/12) + 7/12*log(7/12)

C) -(5/12*log(5/12) + 7/12*log(7/12))

D) -5/12*log(5/12) + 7/12*log(7/12)

Q9. [6 points] Which one of the following statements is NOT true for decision tree:

A) Decision tree is easy to interpret
B) Decision tree can be represented by a set of rules
C) Decision tree is a linear classifier
D) Decision tree can handle categorical variables naturally

Q10. [6 points] You are constructing an information-gain-based decision tree to predict a dependent variable "loan payment," based on bank records. "Loan payment" has two possible values: "yes" and "no". Your dataset also includes the following 6 variables:

"Balance", "Residence" (privately owned or rent), "age", "gender", "employment", "marital status"

After beginning to build the tree, you reached the tree structure presented below. Which variables would you consider as candidates to be used for additional splitting of the node marked by the circle if no specific restriction is required?

A) "Balance", "Residence", "Age", "Gender"
B) "Residence", "Gender", "Employment", "Marital status"
C) "Gender", "Employment", "Marital status"
D) All

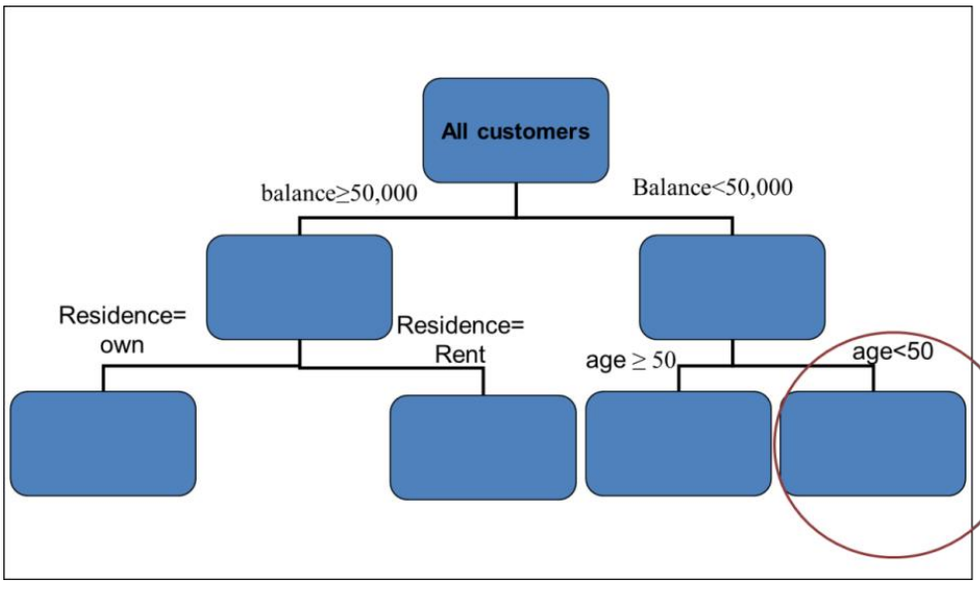


Figure for Q10

Q11. **[15points] The figure as follows shows a split on a training data of 100 samples into two groups, In each node, the number of observations with class + is given in the left part, and the number of observations with class - in the right part.**

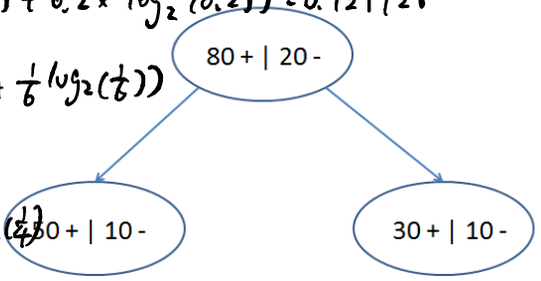
entropy (not split) $= -(0.8 \times \log_2(0.8) + 0.2 \times \log_2(0.2)) = 0.721928$

entropy (split 1) $= -(\frac{5}{6} \times \log_2(\frac{5}{6}) + \frac{1}{6}\log_2(\frac{1}{6}))$

$= 0.65$

entropy (split 2) $= -(\frac{3}{4}\log_2(\frac{3}{4}) + \frac{1}{4}\log_2(\frac{1}{4}))$

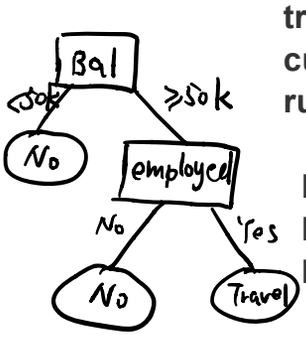
$= 0.811278$



Please compute the information gain if we split the 100 samples into 2 subgroups as above (at least 3 digits after the decimal point)?
Note: please calculate logarithm with base 2, like $\log_2 X$.

information gain $= 0.721928 - (0.6 \times 0.65 + 0.4 \times 0.811278) = -0.010\,5832 \approx -0.011$

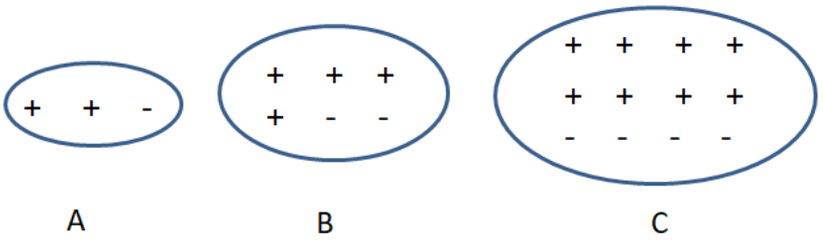
Q12. **[10 points] We are now trying to predict whether a user will buy our travel product based on their bank balance and is whether employed currently. Based on the feedback from sales, we now have following rules about the prediction:**



IF (Balance<50K) THEN Class= Not Travel
IF (Balance>=50K) AND (Employed = Yes) THEN Class= Travel
IF (Balance>=50K) AND (Employed = No)THEN Class= Not Travel

Please draw the decision tree that can describe above rules.

Q13. **[15 points] Given three different leaf nodes in a decision tree as follows, now we need to estimate the probably of a new sample to be class + when falls into these leaf nodes respectively.**



In particular, please estimate the probability by frequency and by Laplace correction respectively, and fill your estimation of probability being class + (in the table provided in the answer sheet).

	Leaf A	Leaf B	Leaf C
Frquency	2/3 = 0.67	4/6 = 0.67	8/12 = 0.67
Laplace	3/5 = 0.6	5/8 = 0.625	9/14 = 0.643