MATH2319 - Machine Learning
Mid-Semester Test, Semester 1, 2018

First name: ______________________ Surname: __________KEY_____________

Student ID: ____________

Instructions:

1. **Smart phones and smart watches are prohibited. Calculators are allowed.**
2. This test contains 4 pages and 4 problems for a total of 100 points.
3. You may use only one textbook of your choice. Any other notes are prohibited.
4. If you use back of a sheet, you must clearly indicate so.

Problem 1: (30 points) You are given the dataset below with three descriptive features (*Smoker, Obese, Family*) with *Risk* being the target feature.

SMOKER	OBESE	FAMILY	RISK
false	false	yes	low
true	false	yes	high
false	false	no	low
true	true	yes	high
true	true	no	high

A) (5 points) What is the total entropy of this dataset?

H(Risk, D) = -[Pr(Risk=low) * $\log_2$(Pr(Risk=low)) + Pr(Risk=high) * $\log_2$(Pr(Risk=high))]
= -[(2/5) * $\log_2$(2/5) + (3/5) * $\log_2$(3/5)]
= 0.4 * 1.32 + 0.6 * 0.74
= 0.53 + 0.44
= 0.97

B) (20 points) Which one of the two descriptive features would you split at the root node if you are to use the Information gain split criterion: *Family or Smoker*? Show all your work.

Smoker:

Rem(Smoker, D) = Pr(Smoker =true) * H(Risk, D(Smoker =true) +
Pr(Smoker =false) * H(Risk, D(Smoker =true)
= (3/5) * 0 + (2/5) * 0
= 0

IG(Smoker, D) = 0.97 – 0 = 0.97

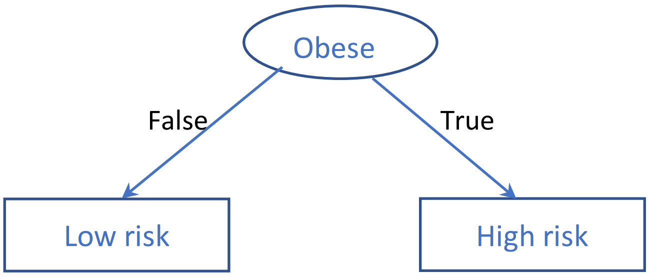
Family:

Rem(Family, D) = Pr(Family = yes) * H(Risk, D(Family = yes) +
Pr(Family =no) * H(Risk, D(Family =no)
= (3/5) * [-(1/3) * $\log_2$(1/3) - (2/3) * $\log_2$(2/3)] +
(2/5) * [-(1/2) * $\log_2$(1/2) - (1/2) * $\log_2$(1/2)]
= 0.6 * 0.92 + 0.4 * 1
= 0.95

IG(Family, D) = 0.97 – 0.95 = 0.02

Highest IG is for Smoker → Choose Smoker

C) (5 points) Suppose you decided to split on the *Obese* variable and you decided to make only one split at the root node. Draw the corresponding decision tree and label the predictions made at each one of the leaf nodes.



Problem 2: (20 points) You would like to build a recommender system for a large online shop that has a stock of over thousands of items. In this domain, the behavior of customers is captured in terms of what items they have bought or not bought. For example, the following table lists the behavior of three customers in this domain for a subset of four items.

ID	**Item 1**	**Item 2**	**Item 3**	**Item 4**
1	True	False	True	False
2	True	False	False	False
3	True	True	True	True

A) (5 points) The company has decided to use a similarity-based model to implement the recommender system. Which of the following three similarity indexes do you think the system should be based on and why? Explain your choice in no more than one sentence.

$$\text{Russell-Rao(X,Y)} = \frac{CP(X,Y)}{P}$$

$$\text{Sokal-Michener(X,Y)} = \frac{CP(X,Y)+CA(X,Y)}{P}$$

$$\text{Jaccard(X,Y)} = \frac{CP(X,Y)}{CP(X,Y)+PA(X,Y)+AP(X,Y)}$$

Jaccard, as co-absences are not very meaningful in the presence of thousands of items and RR similarity is just too simple.

B) (15 points) What items will the system recommend to the following customer?

ID	**Item 1**	**Item 2**	**Item 3**	**Item 4**
4	True	False	True	True

Assume that the recommender system uses the *Russel-Rao* similarity index and is trained on the sample dataset listed above. Also assume that the system generates recommendations by recommending the items that the most similar customer has bought but that the query customer has not bought. Finally, assume that ties are broken by selecting the customer with the highest number of purchases. Show all your work for full credit.

$sim_{RR}(1,4)$: 2/4 = 0.5

$sim_{RR}(2,4)$: 1/4 = 0.25

$sim_{RR}(3,4)$: 2/4 = 0.75*

Item 2 is recommended.

Problem 3: (20 points) Calculate the probability of a model ensemble that uses simple majority voting making an incorrect prediction in the following scenarios. (Hint: the binomial distribution will be useful.)

A) (7 points) The ensemble contains 3 independent models, all of which have an error rate of 40%.

$$\begin{aligned}\text{Pr(incorrect prediction)} &= \binom{3}{2}(0.4)^2(0.6)^1 + \binom{3}{3}(0.4)^3(0.6)^0 \\ &= 3*0.16*0.6 + 1*0.064 \\ &= 0.0288 + 0.064 \\ &= 0.35\end{aligned}$$

Alternatively, compute directly: (A: Accurate prediction, N: Inaccurate prediction)

$$\begin{aligned}\text{Pr(incorrect prediction)} &= \text{Pr(NNA)} + \text{Pr(ANN)} + \text{Pr(NAN)} + \text{Pr(NNN)} \\ &= 3*0.16*0.6 + 0.064 \\ &= \text{same result}\end{aligned}$$

B) (8 points) The ensemble contains 5 independent models, all of which have an error rate of 40%.

$$\begin{aligned}\text{Pr(incorrect prediction)} &= \binom{5}{3}(0.4)^3(0.6)^2 + \binom{5}{4}(0.4)^4(0.6)^1 + \binom{5}{5}(0.4)^5(0.6)^0 \\ &= 10*0.064*0.36 + 5*0.0256*0.6 + 1*0.01 \\ &= 0.02304 + 0.0768 + 0.01024 \\ &= 0.32\end{aligned}$$

C) (5 points) Based on your answers above, which ensemble you would choose? Explain your choice in no more than one sentence.

The ensemble with 5 independent models would be selected. This is because it gives a lower probability for an incorrect prediction.

Problem 4: (30 points) For each one of the scenarios below, circle whether the nearest neighbor (NN) or the decision tree (DT) method would be preferable in general.

1. NN / [✓ DT]: There are many irrelevant descriptive features in the training data.
2. [✓ NN] / [✓ DT]: There is a lot of noise in the training data. **(Both answers accepted)**
3. [✓ NN] / DT: You suspect that there is a concept drift.
4. NN / [✓ DT]: There are a large number of observations in the training data and you would like to make a prediction rather quickly.
5. [✓ NN] / DT: You would like to use a lazy learner.
6. NN / [✓ DT]: Most of the descriptive features are categorical.
7. [✓ NN] / DT: Most of the descriptive features are numerical.
8. NN / [✓ DT]: Your descriptive features are a mix of categorical and numerical variables and you would like to perform only a minimal amount of data preprocessing.
9. [✓ NN] / DT: You would like to easily add new observations to the training data.
10. Circle the correct answer: True / False: I wrote my name and student ID **in a legible manner** on the top of page 1 where indicated.