10/18/2021 Name:

DATA SCIENCE

MIDTERM TEST

1. To solve a classification problem, we build three models on the data. The accuracy on the training and validation sets of the three models is summarized in the table below. (12%)
   a. Which model would you choose and why?
   b. What phenomena occur regarding the models you did not choose?
   c. Let us assume that all models are kNN models with different *k* values. How would you modify the *k* values of the models you did not choose?

ACCURACY	Training	Validation
Model 1	0.85	0.84
Model 2	0.89	0.79
Model 3	0.79	0.77

a. Model 1, it has the highest accuracy on the validation set
b. Model 2: overfitting, Model 3: underfitting
c. Model 2: increase k value, Model 3: decrease k value

2. Are the following statements TRUE or FALSE? Support your answer with justification and correct the false statements. You only get a point if you justify your answer correctly. (18%)

   a. For interval variables, calculating the variance is a meaningful operation.

TRUE: interval variables are quantitative variables (without a clear concept of 0), so calculating the variance is meaningful

   b. For two binary vectors, the Jaccard index is always less than or equal to the SMC (simple matching coefficient).

TRUE: $\frac{M_{11}}{M_{11}+M_{01}+M_{01}} \le \frac{M_{11}+M_{00}}{M_{11}+M_{01}+M_{01}+M_{00}}$ **, since** $\frac{M_{11}+M_{00}}{M_{11}+M_{01}+M_{01}+M_{00}} - \frac{M_{11}}{M_{11}+M_{01}+M_{01}} \ge 0$

   c. In the kNN algorithm, the variance of the model will be high for low k values.

TRUE: low k values indicate that the model is sensitive for the training data itself, therefore it has a high variance

   d. The *maximum a posteriori* and *maximum likelihood* estimates agree if we assume that the attributes and independent from each other.

FALSE: The two estimates agree if the the prior P(C=c_i) probabilites of the labels are equal for all possible values of hthe label

   e. Laplace estimation helps to eliminate the bias due to the naive Bayes assumption.

FALSE: It does not help to eliminate the bias due to the naive Bayes assumption, it helps tackle the problem of zero probability in the Naïve Bayes machine learning algorithm

   f. In a diagnostic test, a high recall value is more important than a high precision value.

TRUE: High recall indicates that we have a low number of false negatives, meaning that we will not tell somenone they are heatly when they are not

3. We are given two 6-dimensional feature vectors (16%):

$$a = (3, 5, 0, 4, 6)$$
$$b = (6, 9, 1, 9, 13)$$

   a. Calculate the **Minkowski distance** (L_r distance) of the two vectors with exponents $r = 1, 2, \infty$.
   b. Name the **special cases** of Minkowski distance with exponents $r = 1, 2, \infty$.
   c. Calculate the **cosine** similarity and dissimilarity of the two vectors.
   d. What is the **main difference** between Minkowski distance and cosine (dis)similarity? Name a situation when you think using cosine (dis)similarity is more reasonable!

a. $\sum_{i=1}^{5}|a_i - b_i| = |3-6| + |5-9| + |0-1| + |4-9| + |6-13| = 20$

$$\sqrt{\sum_{i=1}^{5}(a_i - b_i)^2} = \sqrt{(3-6)^2 + (5-9)^2 + (0-1)^2 + (4-9)^2 + (6-13)^2} = 10$$

$$\max_{i\in\{1,\,\dots,\,5\}}|p_i - q_i| = |6-13| = 7$$

b. Manhattan, Euclidean, Chebysev

c. $$\cos(a, b) = \frac{a\cdot b}{\|a\|\cdot\|b\|} = \frac{3\cdot 6+5\cdot 9+0\cdot 1+4\cdot 9+6\cdot 13}{\sqrt{3^2+5^2+0^2+4^2+6^2}\cdot\sqrt{6^2+9^2+1^2+9^2+13^2}} \approx 0.995$$

dissimilarity: 1-cos(a, b)=0.005

d. While Minkowski distance takes into consideration the magnitude of the vectors, cosine (dis)similarity only looks at the direction of the vectors. For document term matrices it might be more resonable to use the cosine (dis)similarity.

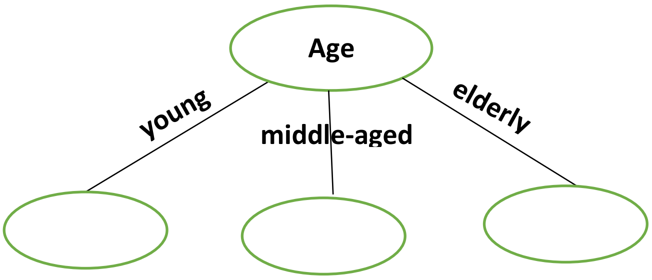
4. A company plans to launch a promotion and asked some people if they were interested. In addition, they recorded three features about each person. We use a **decision tree** to predict who would be interested in the promotion. We are given a dataset with 25 labeled records. We split the data into training and test sets. We train the model on the first 15 instances, and we test it on 10 instances. (32%)
   a. We build a decision tree with depth two on the training data. First, we choose age as the first splitting attribute (see below). Proceed, find **the best splitting attribute** in each child node to build a decision tree with depth two. Use **misclassification error** as the inhomogeneity measure.
   b. Using the decision tree built in part a., test its performance **using the test data**. Determine the **confusion matrix** and calculate the **Accuracy**, **Precision** and **Recall** metrics!
   c. Determine the **confidence scores** (ratio of positive observations) of the leaves based on the training data. Sort the confidence scores of the test instances in ascending order.
   d. Construct the **ROC curve** of the model (using the test instances). If more instances have the same confidence scores ROC curve may change diagonally!
   e. Calculate the **AUC score** of the model!
   f. What is the **probabilistic interpretation** of the AUC score?

Training set

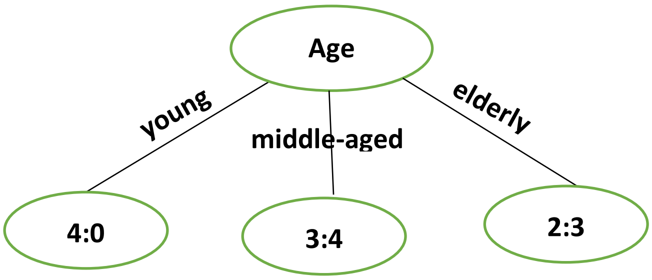
#	*Age*	*Owns a car?*	*Owns a house?*	*Interested in promotion?*
1	young	Yes	Yes	**Yes (+)**
2	young	Yes	No	**Yes (+)**
3	young	No	No	**Yes (+)**
4	young	No	No	**Yes (+)**
5	middle-aged	Yes	Yes	**No (-)**
6	middle-aged	Yes	No	**Yes (+)**
7	middle-aged	Yes	No	**Yes (+)**
8	middle-aged	Yes	No	**No (-)**
9	middle-aged	No	Yes	**No (-)**
10	middle-aged	No	No	**No (-)**
11	middle-aged	No	No	**Yes (+)**
12	elderly	Yes	Yes	**No (-)**
13	elderly	Yes	No	**No (-)**
14	elderly	No	Yes	**Yes (+)**
14	elderly	No	No	**Yes (+)**
15	elderly	No	No	**No (-)**

Test set

#	*Age*	*Owns a car?*	*Owns a house?*	*Interested in promotion?*
16	young	Yes	Yes	**Yes (+)**
17	young	Yes	No	**Yes (+)**
18	young	No	Yes	**No (-)**
19	middle-aged	Yes	Yes	**No (-)**
20	middle-aged	Yes	No	**Yes (+)**
21	middle-aged	No	No	**No (-)**
22	elderly	Yes	Yes	**No (-)**
23	elderly	Yes	No	**No (-)**
24	elderly	Yes	Yes	**Yes (+)**
25	elderly	No	Yes	**Yes (+)**



a.

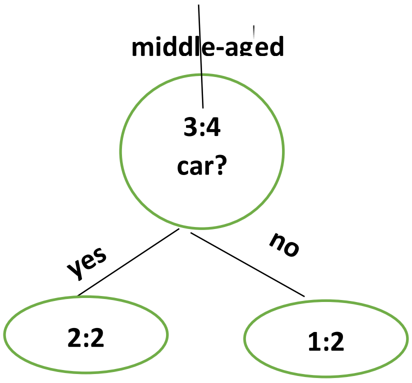


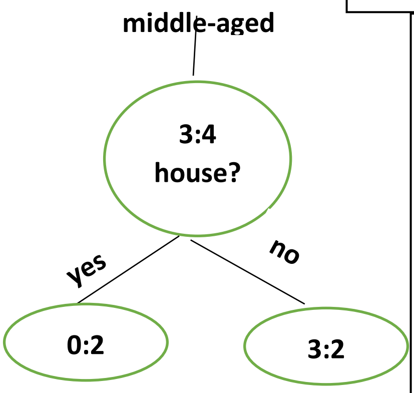
$$I(Age = young) = 1 - \max\left(\frac{4}{4}, \frac{0}{4}\right) = 0$$

$$I(Age = middle.aged) = 1 - \max\left(\frac{3}{7}, \frac{4}{7}\right) = \frac{3}{7}$$

$$I(Age = elderly) = 1 - \max\left(\frac{2}{5}, \frac{3}{5}\right) = \frac{2}{5}$$

Age is totally homogeneous, no need to split, proceed with the other two nodes.





For the middle aged node:

$$I(car = yes) = 1 - \max\left(\frac{2}{4}, \frac{2}{4}\right) = \frac{1}{2}$$

$$I(car = no) = 1 - \max\left(\frac{1}{3}, \frac{2}{3}\right) = 1/3$$

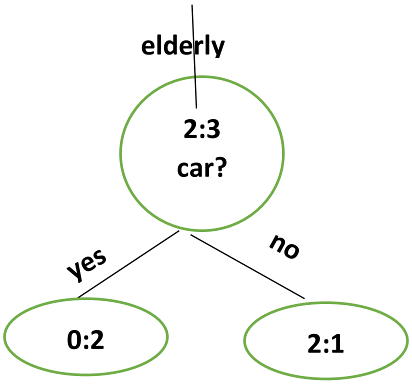
$$\Delta(\text{car}) = \frac{3}{7} - \left(\frac{4}{7} \cdot \frac{1}{2} + \frac{3}{7} \cdot \frac{1}{3}\right) = 0$$

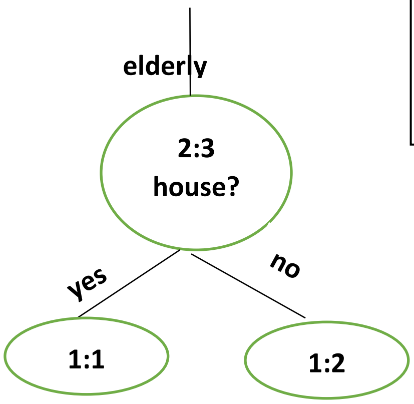
$$I(house = yes) = 1 - \max\left(\frac{0}{2}, \frac{2}{2}\right) = 0$$

$$I(house = no) = 1 - \max\left(\frac{3}{5}, \frac{2}{5}\right) = \frac{2}{5}$$

$$\Delta = \frac{3}{7} - \left(\frac{2}{7} \cdot 0 + \frac{5}{7} \cdot \frac{2}{5}\right) = \frac{1}{7}$$

The gain is larger for house, we will split on that!





For the elderly node:

$$I(car = yes) = 1 - \max\left(\frac{0}{2}, \frac{2}{2}\right) = 0$$

$$I(car = no) = 1 - \max\left(\frac{2}{3}, \frac{1}{3}\right) = 1/3$$

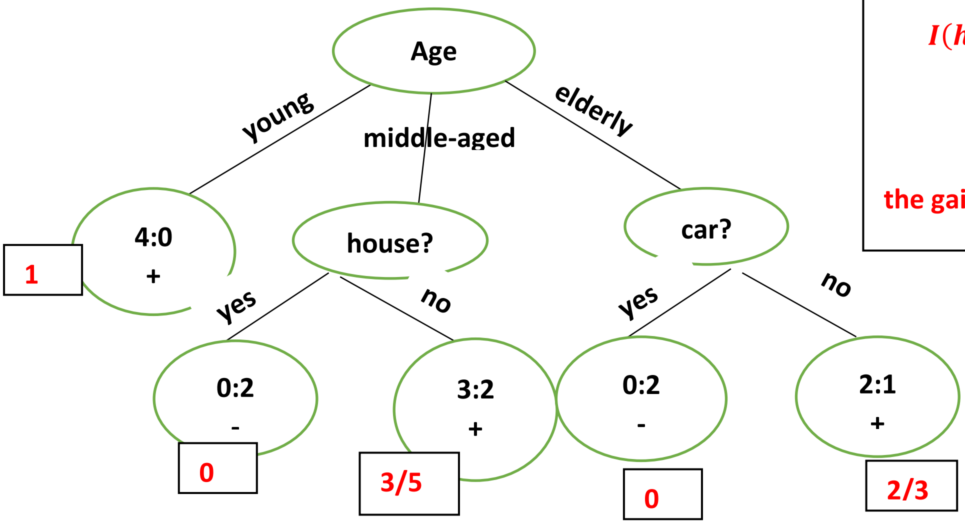
$$\Delta(\text{car}) = \frac{2}{5} - \left(\frac{2}{5} \cdot 0 + \frac{3}{5} \cdot \frac{1}{3}\right) = 1/5$$

$$I(house = yes) = 1 - \max\left(\frac{1}{2}, \frac{1}{2}\right) = \frac{1}{2}$$

$$I(house = no) = 1 - \max\left(\frac{1}{3}, \frac{2}{3}\right) = \frac{1}{3}$$

$$\Delta = \frac{2}{5} - \left(\frac{2}{5} \cdot \frac{1}{2} + \frac{3}{5} \cdot \frac{1}{3}\right) = 0$$

the gain is larger for car, we will split on that!



Test set					***Prediction***	
#	***Age***	***Owns a car?***	***Owns a house?***	***Interested in promotion?***	***Prediction Confidence score***	
16	young	Yes	Yes	**Yes (+)**	**+ 1**	**TP**
17	young	Yes	No	**Yes (+)**	**+ 1**	**TP**
18	young	No	Yes	**No (-)**	**+ 1**	**FP**
19	middle-aged	Yes	Yes	**No (-)**	**- 0**	**TN**
20	middle-aged	Yes	No	**Yes (+)**	**+ 3/5**	**TP**
21	middle-aged	No	No	**No (-)**	**+ 3/5**	**FP**
22	elderly	Yes	Yes	**No (-)**	**- 0**	**TN**
23	elderly	Yes	No	**No (-)**	**- 0**	**TN**
24	elderly	Yes	Yes	**Yes (+)**	**- 0**	**FN**
25	elderly	No	Yes	**Yes (+)**	**+ 2/3**	**TP**

b.

Confusion matrix		**Predicted**	
		+	**-**
Actual	**+**	**4**	**1**
	-	**2**	**3**

c. Confidence scores marked in the table sorted n ascending order together with their actual labels:
(0, -); (0,-); (0,-); (0, +); (3/5, -), (3/5, +), (2/3, +), (1, -); (1, +); (1, +)

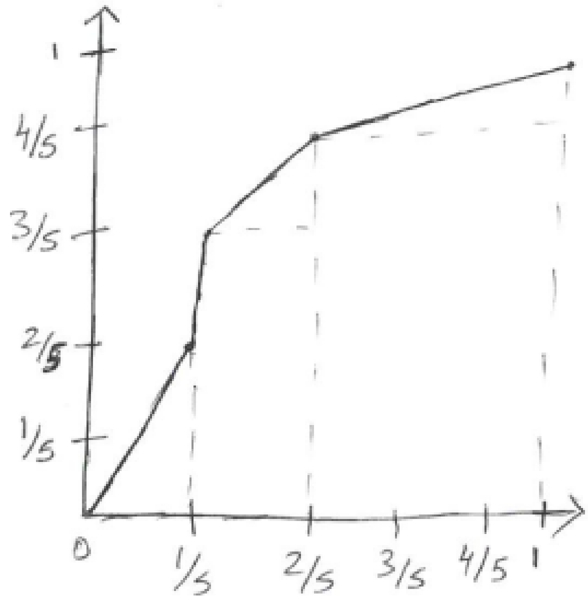
b.

Accuray= $\frac{TP+TN}{TP+TN+FP+FN}=\frac{7}{10}$

Precision= $\frac{TP}{TP+FP}=\frac{4}{6}$

Recall= $\frac{TP}{TP+FN}=\frac{4}{5}$

d.

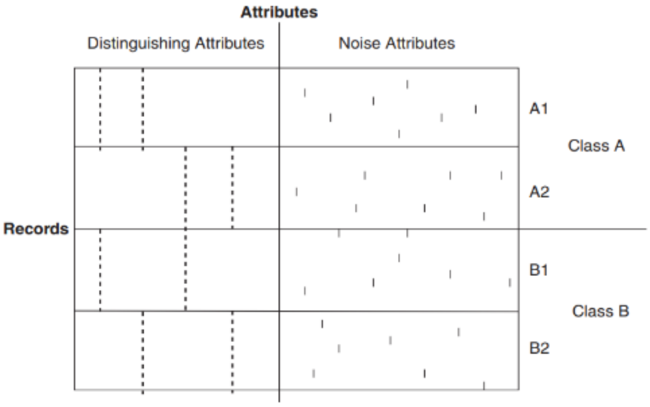


e. AUC score:

$$\left(\frac{1}{2}\cdot\frac{1}{5}\cdot\frac{2}{5}\right)+\left(\frac{1}{2}\cdot\frac{1}{5}\cdot\frac{1}{5}\right)+\left(\frac{1}{2}\cdot\frac{3}{5}\cdot\frac{1}{5}\right)+\frac{1}{5}\cdot\frac{3}{5}+\frac{3}{5}\cdot\frac{4}{5}=\frac{18}{25}$$

f. AUC= the probability that a randomly-chosen positive record is ranked more highly than a randomly chosen negative record

5. Given the following data table with some distinguishing **binary attributes** and some noisy binary attributes that randomly take on the value 1. How would the **k-NN**, **decision tree** and **naive Bayes** classifier perform on the following data? Support your answers with reasoning and sketch calculations. (22%)
   a. Answer the question if the problem is treated as a binary classification problem where the **two class labels** are A and B!
   b. The two classes are further divided into two parts and now the objective is to learn the **four classes** A1, A2, B1 and B2. Evaluate the performance of the three algorithms in this case, too!



a. kNN will not do well due to high number of noise attributes, kNN is sensitive to noise, two records might be close to each other just because they agree in the noise attributes
For the two-class problem, decision tree will not perform well because the homogeneity will not improve after splitting the data using the distinguishing attributes (decision tree is a local greedy algorithm!)
NB will not do well on this data set because the conditional probabilities for each distinguishing attribute given the class are the same for both class A and class B.

b. kNN: same as in a.
If there are four classes, then decision tree will improve considerably, since splitting on the distinguishing attributes will improve the homogeneity
The **performance of NB will improve on the subclasses because the product of conditional probabilities among the distinguishing attributes will be different for each subclass.**
E.g.

$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0|A1) = 1 \cdot 1 \cdot 1 \cdot 1$$
$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0|A2) = \epsilon \cdot \epsilon \cdot \epsilon \cdot \epsilon$$
$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0|B1) = 1 \cdot \epsilon \cdot \epsilon \cdot 1$$
$$P(X_1 = 1, X_2 = 1, X_3 = 0, X_4 = 0|B2) = \epsilon \cdot 1 \cdot 1 \cdot \epsilon$$