IE30301-Data Mining Assignment 4 (70 Points)

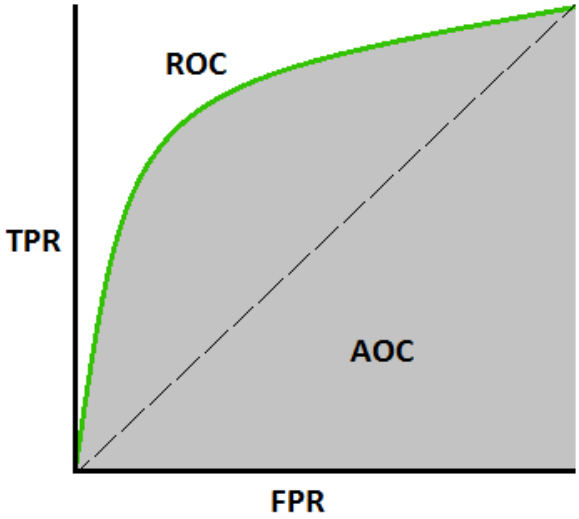
Eldor Fozilov

April 13, 2022

Exercise 1

Summarize the following concepts in 3 ~ 4 sentences each. (write it in your own words). If not, there are 2 points deduction per problem. [12 pts, 3 pts for each.]

1. **Linear Discriminant Analysis** is a machine learning method designed to solve classification problems by performing dimensionality reduction, considering the label information. It aims at optimally projecting data points into a subspace so that those data points belonging to different classes end up being separated from each other after the projection. The dimension of that subspace depends on the number of classes: if there are **C** classes, then the data (excluding the target variable) will be reduced to **C-1** dimensional subspace. After the optimal subspace is determined, new records will be projected to that subspace and they will be assigned to a particular class which represents the closest projected sample records in the train data set to that new record.
2. **Cross Validation** is a way of evaluating the performance of machine learning models. It offers a particular method to estimate model performance on unseen data not used while training. Firstly, data is split into **k** subsets (k is defined by the user) of equal size, and then each subset in turn is used for testing and the remainder for training. To determine the overall error rate, the error estimates are averaged across all iterations. Cross validation is very useful when the size of a given data set is small, in which case splitting the data set into training, validation, and testing set might negatively affect accuracy of a model.
3. **AUC for ROC curves** is an evaluation metric for checking performance of a classification model. The graphic example of a ROC curve is provided below. As we can see, ROC curve plots TRP (true positive rate) vs. FPR (false positive rate) at different classification probability thresholds. Decreasing the probability threshold classifies more items as positive, which increases both TPR and FPR. What **AUC** measures is the area underneath the ROC curve, which represent an aggregate measure of performance across all possible classification thresholds (from 0 to 1). So, that is why we want AUC to be as high as possible. The range of values AUC can take is from 0 to 1, and a model whose predictions are all wrong has an AUC measure of 0; one whose predictions are all correct has an AUC measure of 1. The AUC benchmark for the performance of a model is 0.5, which represents the performance of a strategy in which classification predictions are made randomly.



4. **Decision Tree** is a supervised-machine learning method used for both classification and regression. What makes this method different from other machine learning methods is that it is a rule-based model: the constructed decision tree provides a set of rules, which determine how sample records are classified or what the regression prediction ends up being for those sample records. To determine the set of rules, we first decide how to split the records into branches on each level in our decision tree. For doing this, **multi-way split** or **binary split** methods are used when dealing with nominal variables; **discretization** and **binary decision** methods are used when dealing with continuous variable). Then, for classification, optimization criteria such as **GainRATIO**$_{\textbf{split}}$ and for regression, optimization criterion focused on **MSE** or **variance** of the target variable are employed to select a feature that results in the most optimal split.

Exercise 2

Consider a multinomial logistic regression model with the dependent variable that has three or more nominal type categories. If we define v_{ij} as the value of category j from the r_i independent trial $\left(\text{instead of the usual binary logistic regression formula } v_{ij} = \begin{cases} 1, & \text{for } y_i = j \\ 0, & \text{for } y_i \neq j \end{cases}\right)$ then the v_{ij} follows a multinomial distribution with probabilities $(P_1, \ldots, P_j)$.
Construct the likelihood function for this case. (It is not that complicated. You just need to use the probability mass function of the multinomial distribution) [10 pts]

Answer:
Let's assume we have **n** independent trials and **C** number of categories, where **C** ≥ 3. Let's also say that the random variable V_i represent the value of an observed category in the i-th independent trial.
Then,

$$P(V_i = v_{ij}) = \prod_{j=1}^{C} P_j^{a_{ij}}, \quad \text{where } a_{ij} = \begin{cases} 1, & \text{for } V_i = v_{ij} \\ 0, & \text{for } V_i \neq v_{ij} \end{cases}$$

Now, we can write the likelihood function in the following way:

$$L = \prod_{i=1}^{n} \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_C!} \prod_{j=1}^{C} [P(V_i = v_{ij})]^{n_j} = \prod_{i=1}^{n} \frac{n!}{n_1! \cdot n_2! \cdot \ldots \cdot n_C!} \prod_{j=1}^{C} P_j^{n_j},$$

where n_j (j = 1,2,...,C) represents the number of trials that fall in category j out of n trials, and $\sum_{j=1}^{C} n_j = n$

Exercise 3

Compute the Linear Discriminant projection for the following two-dimensional dataset. [10 pts]

Variable_A	Variable_B	Result
1.84	7.57	1
1.37	9.83	1
2.26	7.82	1
2.18	8.71	1
1.58	4.97	0
1.16	6.31	0
2.27	4.32	0

3.1

Calculate the class statistics: scatter matrices S and mean $\boldsymbol{\mu}$ ($S_1, S_2, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2$) [4 pts]

Answer:

We will use the following formula to calculate S_1, related to **class "0"** and S_2, related to **class "1"**:

$$S_i = \sum_{\mathbf{x} \in w_i} (\mathbf{x} - \boldsymbol{\mu}_i)(\mathbf{x} - \boldsymbol{\mu}_i)^T$$

First we will calculate the mean vectors:

$$\boldsymbol{\mu}_1 = \frac{1}{3}\left(\begin{bmatrix}1.58\\4.97\end{bmatrix} + \begin{bmatrix}1.16\\6.31\end{bmatrix} + \begin{bmatrix}2.27\\4.32\end{bmatrix}\right) = \begin{bmatrix}1.67\\5.2\end{bmatrix}$$

$$\boldsymbol{\mu}_2 = \frac{1}{4}\left(\begin{bmatrix}1.84\\7.57\end{bmatrix} + \begin{bmatrix}1.37\\9.83\end{bmatrix} + \begin{bmatrix}2.26\\7.82\end{bmatrix} + \begin{bmatrix}2.18\\8.71\end{bmatrix}\right) = \begin{bmatrix}1.9125\\8.4825\end{bmatrix}$$

Then,

$$S_1 = \begin{bmatrix} 1.58 - 1.67 \\ 4.97 - 5.2 \end{bmatrix} \begin{bmatrix} 1.58 - 1.67 \\ 4.97 - 5.2 \end{bmatrix}^T + \begin{bmatrix} 1.16 - 1.67 \\ 6.31 - 5.2 \end{bmatrix} \begin{bmatrix} 1.16 - 1.67 \\ 6.31 - 5.2 \end{bmatrix}^T + \begin{bmatrix} 2.27 - 1.67 \\ 4.32 - 5.2 \end{bmatrix} \begin{bmatrix} 2.27 - 1.67 \\ 4.32 - 5.2 \end{bmatrix}^T$$

$$= \begin{bmatrix} -0.09 \\ -0.23 \end{bmatrix} \begin{bmatrix} -0.09 \\ -0.23 \end{bmatrix}^T + \begin{bmatrix} -0.51 \\ 1.11 \end{bmatrix} \begin{bmatrix} -0.51 \\ 1.11 \end{bmatrix}^T + \begin{bmatrix} 0.6 \\ -0.88 \end{bmatrix} \begin{bmatrix} 0.6 \\ -0.88 \end{bmatrix}^T = \begin{bmatrix} 0.6282 & -1.0734 \\ -1.0734 & 2.0594 \end{bmatrix}$$

$$S_2 = \begin{bmatrix} 1.84 - 1.9125 \\ 7.57 - 8.4825 \end{bmatrix} \begin{bmatrix} 1.84 - 1.9125 \\ 7.57 - 8.4825 \end{bmatrix}^T + \begin{bmatrix} 1.37 - 1.9125 \\ 9.83 - 8.4825 \end{bmatrix} \begin{bmatrix} 1.37 - 1.9125 \\ 9.83 - 8.4825 \end{bmatrix}^T +$$
$$\begin{bmatrix} 2.26 - 1.9125 \\ 7.82 - 8.4825 \end{bmatrix} \begin{bmatrix} 2.26 - 1.9125 \\ 7.82 - 8.4825 \end{bmatrix}^T + \begin{bmatrix} 2.18 - 1.9125 \\ 8.71 - 8.4825 \end{bmatrix} \begin{bmatrix} 2.18 - 1.9125 \\ 8.71 - 8.4825 \end{bmatrix}^T$$

$$= \begin{bmatrix} 0.4919 & -0.8342 \\ -0.8342 & 3.4391 \end{bmatrix}$$

3.2

Calculate the within- and between-class scatter (S_B, S_W) [3 pts]

Answer:

$$S_W = S_1 + S_2 = \begin{bmatrix} 0.6282 & -1.0734 \\ -1.0734 & 2.0594 \end{bmatrix} + \begin{bmatrix} 0.4919 & -0.8342 \\ -0.8342 & 3.4391 \end{bmatrix} = \begin{bmatrix} 1.1201 & -1.9076 \\ -1.9076 & 5.4985 \end{bmatrix}$$

$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T = \begin{bmatrix} 1.67 - 1.9125 \\ 5.2 - 8.4825 \end{bmatrix} \begin{bmatrix} 1.67 - 1.9125 \\ 5.2 - 8.4825 \end{bmatrix}^T = \begin{bmatrix} 0.0588 & 0.796 \\ 0.796 & 10.7748 \end{bmatrix}$$

3.3

Based on the results 3.1 and 3.2, calculate the optimal $\mathbf{w}^\star$. [3 pts]

Answer:

$$\mathbf{w}^\star = S_W^{-1}(\mu_1 - \mu_2)$$

S_W^{-1} exists, and it is (approximately) equal to $\begin{bmatrix} 2.182 & 0.757 \\ 0.757 & 0.4445 \end{bmatrix}$

So,

$$\mathbf{w}^\star = \begin{bmatrix} 2.182 & 0.757 \\ 0.757 & 0.4445 \end{bmatrix} \begin{bmatrix} 1.67 - 1.9125 \\ 5.2 - 8.4825 \end{bmatrix} = \begin{bmatrix} 2.182 & 0.757 \\ 0.757 & 0.4445 \end{bmatrix} \begin{bmatrix} -0.2425 \\ -3.2825 \end{bmatrix} = \begin{bmatrix} -3.014 \\ -1.643 \end{bmatrix}$$

Exercise 4

The following tables are confusion matrices of the test dataset from the two methods. (Logistic Regression and Decision Tree). [12 pts]

Table 1: Logistic Regression

Predicted / Actual	Disorder	No Disorder
Disorder	8	18
No Disorder	45	929

Table 2: Decision Tree

Predicted / Actual	Disorder	No Disorder
Disorder	12	14
No Disorder	60	914

4.1

Explain how we can interpret the accuracy, sensitivity, and specificity, respectively.
Calculate accuracy rate, sensitivity, and specificity for each method. (Positive class = 'Disorder')
[3 pts]

Answer:

For **logistic regression**, we have the following:

$$\text{Accuracy rate} = \frac{8 + 929}{8 + 18 + 45 + 929} = 0.937$$

$$\text{Sensitivity} = \frac{8}{8 + 18} = 0.308$$

$$\text{Specificity} = \frac{929}{45 + 929} = 0.954$$

For **decision tree**, we have the following:

$$\text{Accuracy rate} = \frac{12 + 914}{12 + 14 + 60 + 914} = 0.926$$

$$\text{Sensitivity} = \frac{12}{12 + 14} = 0.462$$

$$\text{Specificity} = \frac{914}{60 + 914} = 0.938$$

4.2

Compare the accuracy obtained in (4.1) with that of the naïve rule.
(naïve rule: classify all records as belonging to the most prevalent class) [3 pts]

Answer:

Table 3: Naive method

Predicted / Actual	Disorder	No Disorder
Disorder	0	26
No Disorder	0	974

When **naive rule** is used, **accuracy rate** is $\frac{974}{26 + 974} = 0.974$, which is higher than accuracy rates found using **decision tree** and **logistic regression** in (4.1).

4.3

Which method do you prefer for further implementation in terms of accuracy, sensitivity, and specificity? Explain your reasons. (You should note that the class is imbalanced.) [3 pts]

Answer:

In the problem that we are dealing with, which is classifying records to either "Disorder" or "No Disorder" class, it is much more important for us to identify records having a disorder correctly rather than identifying records that have no disorder because it is evident that the cost of making a mistake, which can be a person's life, in misclassifying a **disorder record** as a **no-disorder record** is much higher than the opposite case. So, we want the sensitivity measure to be as high as possible, but we agree to tolerate lower specificity value and overall accuracy rate. We will choose **Decision Tree** method for implementation since it has the highest sensitivity value compared to sensitivity value obtained using logistic regression method and naïve rule, which has a sensitivity value of zero.

4.4

If the accuracy rates of those data mining methods were no better than the naïve rule, what would you do to improve accuracy? (Write your own opinion.) [3 pts]

Answer:

We saw that naïve rule gave us higher accuracy rate compared to logistic regression and decision tree methods. This probably happened because one class, specifically "No Disorder" class, greatly dominated in our data set (974 no-disorder records versus 26 disorder records), and thus logistic regression and decision tree methods struggled to separate classes correctly, showing lower overall accuracy rate.

Thus, in order to increase the accuracy rates of logistic regression and decision tree methods, we can oversample rare class (disorder records) and increase the number of disorder records in our dataset. As a result, logistic regression and decision tree methods will have more information about the rare class and can better understand the differences between it and the dominant class. As a result, we might end up with improved accuracy rates of those methods.

To improve accuracy, we can also try to use a higher cutoff value than 0.5 to classify records in a more conservative way.

Exercise 5

The following are training samples of 12 objects. Each object is represented as variable X and divided into two classes. (Positive : Class 1, Negative : Class 0) [11 pts]

Object	1	2	3	4	5	6	7	8	9	10	11	12
X	24	30	35	37	42	49	54	56	60	68	72	73
Class	0	0	0	0	1	0	1	1	0	1	1	1

5.1

For the data above, compute sensitivity and specificity according to the change of classification criterion(C) value.
- You should fill in the table below
- Use a classification criterion that if X < C, then classify it as class 0. [6 pts]

Classification criterion	Sensitivity	1-Specificity
X < 24	1	1
X < 30	1	0.83
X < 35	1	0.67
X < 37	1	0.5
X < 42	1	0.33
X < 49	0.83	0.33
X < 54	0.83	0.17
X < 56	0.67	0.17
X < 60	0.5	0.17
X < 68	0.5	0
X < 72	0.33	0
X < 73	0.17	0

Answer:

In [1]:
```
X = [24,30,35,37,42,49,54,56,60,68,72,73]
C = [0,0,0,0,1,0,1,1,0,1,1,1]
```

In [2]:
```
table = {}
n = len(C)
for c in X:
    predictions = []
    for x in X:
        if x < c:
            predictions.append(0)
        else:
            predictions.append(1)

    TP = TN = FP = FN = 0
    for k in range(0,n):
        if predictions[k] == 0 and C[k] == 0:
            TN += 1
        elif predictions[k] == 0 and C[k] == 1:
            FN += 1
        elif predictions[k] == 1 and C[k] == 0:
            FP += 1
        else:
            TP += 1
    table["X < " + str(c)] = [TP / (TP + FN), FP / (TN + FP) ]
```

In [3]:
```
table
```

Out[3]:
```
{'X < 24': [1.0, 1.0],
 'X < 30': [1.0, 0.8333333333333334],
 'X < 35': [1.0, 0.6666666666666666],
 'X < 37': [1.0, 0.5],
 'X < 42': [1.0, 0.3333333333333333],
 'X < 49': [0.8333333333333334, 0.3333333333333333],
 'X < 54': [0.8333333333333334, 0.16666666666666666],
 'X < 56': [0.6666666666666666, 0.16666666666666666],
 'X < 60': [0.5, 0.16666666666666666],
 'X < 68': [0.5, 0.0],
 'X < 72': [0.3333333333333333, 0.0],
 'X < 73': [0.16666666666666666, 0.0]}
```

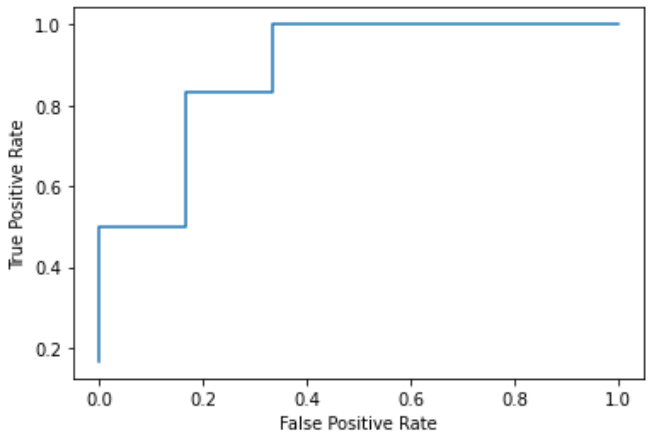
5.2

Generate ROC curve based on the computed sensitivity and specificity. And explain how to interpret the ROC curve. [5 pts]
(Use Python or R to plot the ROC curve, but you should provide a screenshot of the code for generating the plot)

Answer:

In [5]:
```
true_positive_rate = []
false_positive_rate = []
for metrics in table.values():
    true_positive_rate.append(metrics[0])
    false_positive_rate.append(metrics[1])
```

In [6]:
```
import matplotlib.pyplot as plt
plt.plot(false_positive_rate, true_positive_rate)
plt.ylabel('True Positive Rate')
plt.xlabel('False Positive Rate')
plt.show()
```



We know that if the area under the ROC curve is big (higher than the benchmark of 0.5), which is actually the case as we see from the graph above, then the performance of the model can be considered satisfactory.

Exercise 6

The following data set was collected from the survey, consisting of four attributes: Age, Health Concern, Exercise, Health Status, and one target variable: Health Checkup. [15 pts] **(For only exercise 6, Handwriting is allowed. Illegible handwriting will not be graded)**

Age	Health Concern	Exercise	Health Status	Health Checkup
senior	low	frequent	fair	yes
middle-aged	high	seldom	fair	yes
youth	medium	frequent	excellent	yes
middle-aged	medium	seldom	excellent	yes
youth	high	seldom	excellent	no
youth	medium	seldom	fair	no
middle-aged	low	frequent	excellent	yes
middle-aged	high	frequent	fair	yes
senior	medium	seldom	excellent	no
youth	high	seldom	fair	no
senior	low	frequent	excellent	no
senior	medium	seldom	fair	yes
youth	low	frequent	fair	yes
senior	medium	frequent	fair	yes

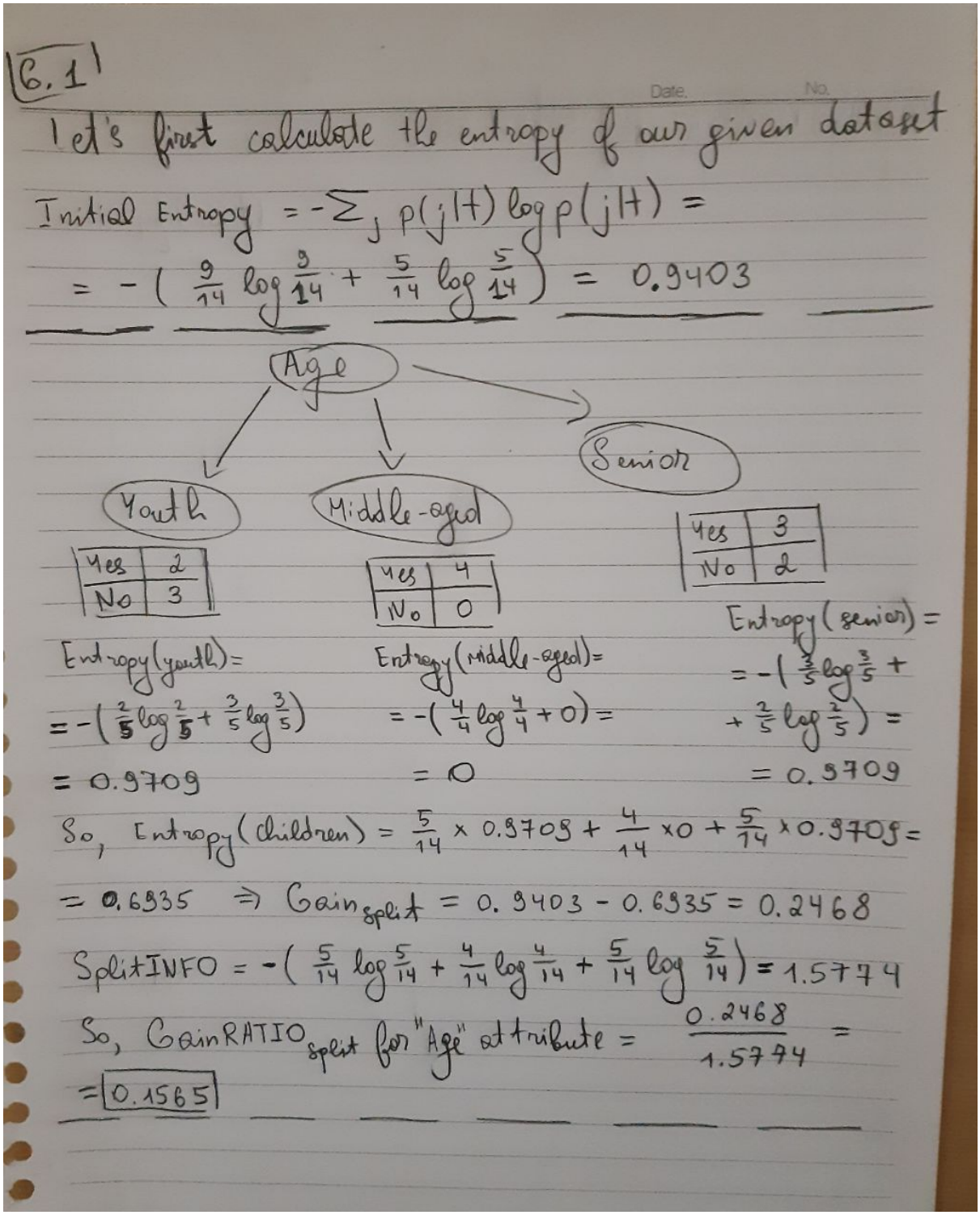
6.1

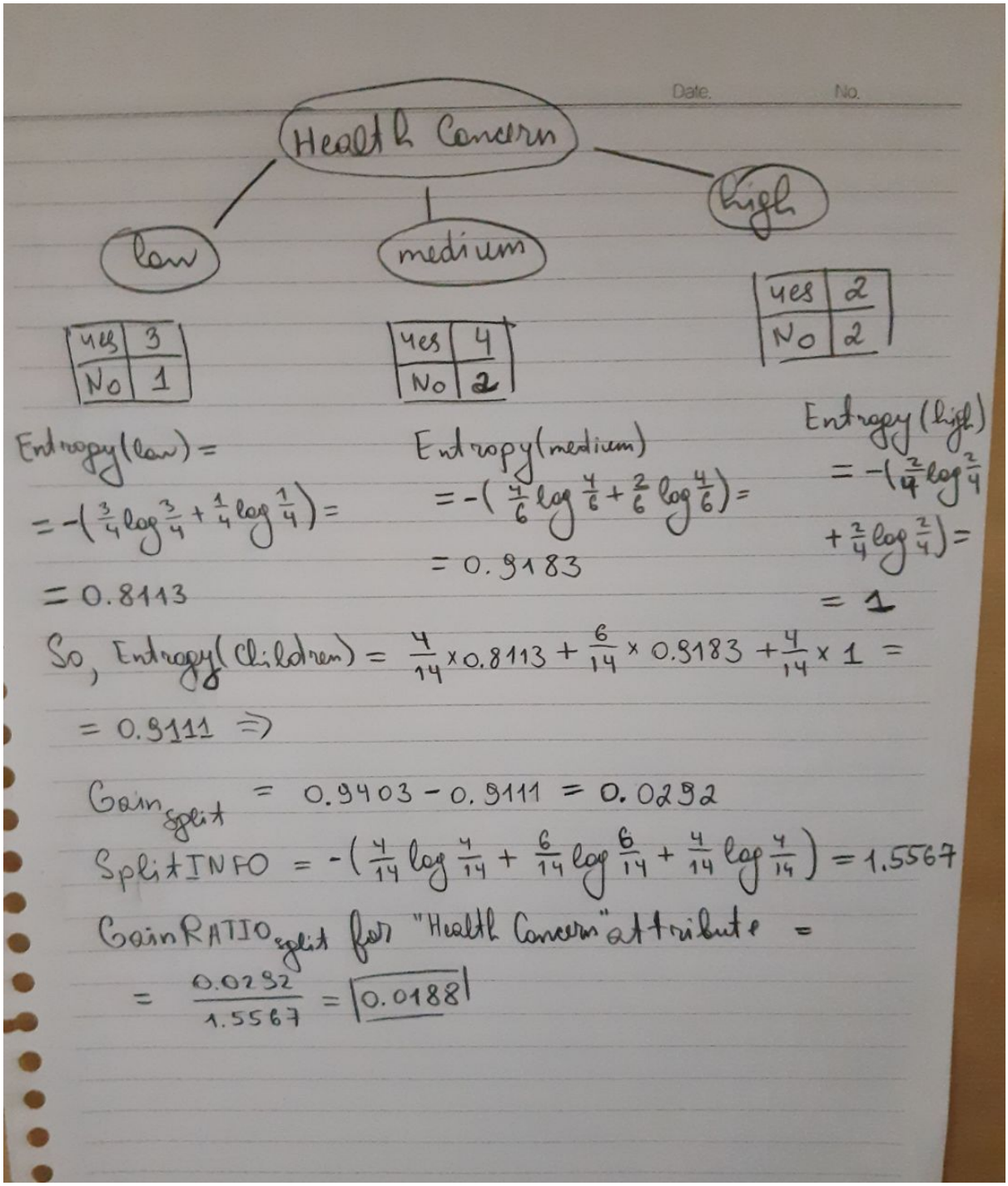
Compute the Gain Ratio for each attribute. Which variable will be a splitting criterion at the root node in terms of Gain Ratio? (Take the multi-split approach for splitting and the binary logarithm (i.e., base 2) for calculating the Gain Ratio. Write down the calculation process) [10 pts]

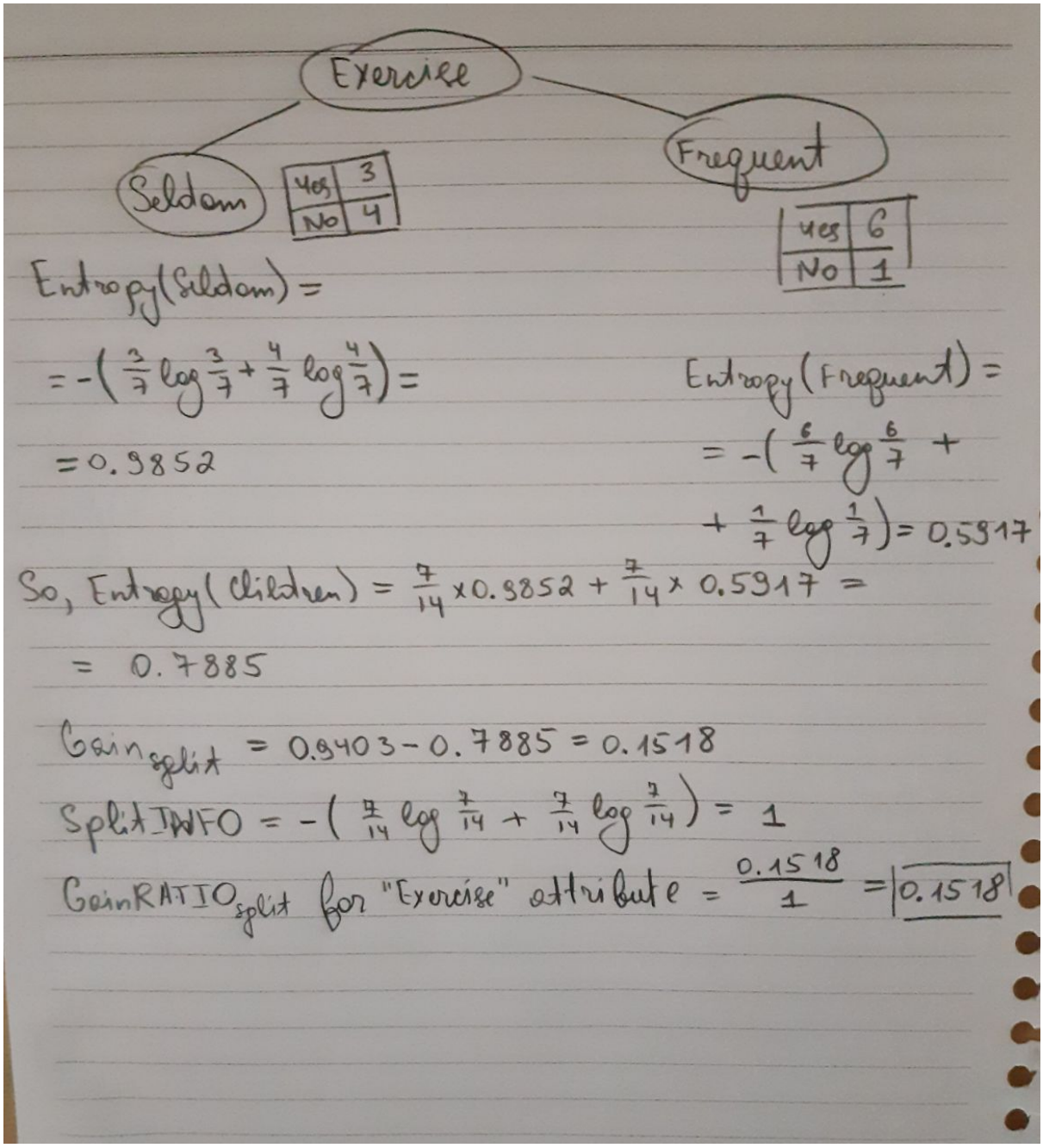
6.2

What is the classification error right after splitting the root node according to the result of 5.1? [5 pts]

Answer for **6.1** and **6.2**:




Date.
No.
Health Concern
low
medium
high
yes | 3
No | 1
yes | 4
No | 2
yes | 2
No | 2
Entropy(low) =
$= -\left(\frac{3}{4}\log\frac{3}{4} + \frac{1}{4}\log\frac{1}{4}\right) =$
$= 0.8113$
Entropy(medium)
$= -\left(\frac{4}{6}\log\frac{4}{6} + \frac{2}{6}\log\frac{4}{6}\right) =$
$= 0.9183$
Entropy(high)
$= -\left(\frac{2}{4}\log\frac{2}{4} + \frac{2}{4}\log\frac{2}{4}\right) =$
$= 1$
So, Entropy(Children) $= \frac{4}{14} \times 0.8113 + \frac{6}{14} \times 0.9183 + \frac{4}{14} \times 1 =$
$= 0.9111 \Rightarrow$
$Gain_{split} = 0.9403 - 0.9111 = 0.0292$
$SplitINFO = -\left(\frac{4}{14}\log\frac{4}{14} + \frac{6}{14}\log\frac{6}{14} + \frac{4}{14}\log\frac{4}{14}\right) = 1.5567$
$GainRATIO_{split}$ for "Health Concern" attribute =
$= \frac{0.0292}{1.5567} = \boxed{0.0188}$


Exercise
Seldom
Frequent
yes 3
No 4
yes 6
No 1
Entropy(Seldom) =
$= -\left(\frac{3}{7}\log\frac{3}{7} + \frac{4}{7}\log\frac{4}{7}\right) =$
$= 0.9852$
Entropy(Frequent) =
$= -\left(\frac{6}{7}\log\frac{6}{7} +$
$+ \frac{1}{7}\log\frac{1}{7}\right) = 0.5917$
So, Entropy(Children) $= \frac{7}{14} \times 0.9852 + \frac{7}{14} \times 0.5917 =$
$= 0.7885$
$Gain_{split} = 0.9403 - 0.7885 = 0.1518$
$SplitINFO = -\left(\frac{7}{14}\log\frac{7}{14} + \frac{7}{14}\log\frac{7}{14}\right) = 1$
$GainRATIO_{split}$ for "Exercise" attribute $= \frac{0.1518}{1} = 0.1518$

Date. No.

Health Status

fair

Yes	6
No	2

excellent

Yes	3
No	3

$Entropy(fair) = -\left(\frac{6}{8}\log\frac{6}{8} + \frac{2}{8}\log\frac{2}{8}\right) = 0.8113$

$Entropy(excellent) = -\left(\frac{3}{6}\log\frac{3}{6} + \frac{3}{6}\log\frac{3}{6}\right) = 1$

$So,\ Entropy(children) = \frac{8}{14} \times 0.8113 + \frac{6}{14} \times 1 = 0.8922$

$Gain_{split} = 0.9403 - 0.8922 = 0.0481$

$SplitINFO = -\left(\frac{8}{14}\log\frac{8}{14} + \frac{6}{14}\log\frac{6}{14}\right) = 0.9852$

$GainRATIO_{split}$ for "Health Status" attribute =

$= \frac{0.0481}{0.9852} = \boxed{0.0488}$

Since the Gain Ratio of the "Age" attribute is the highest, this age variable will be a splitting criterion at the root node.

6.2

Age

Youth

Yes	2
No	3

Middle-aged

Yes	4
No	0

Senior

Yes	3
No	2

Classification error (after splitting) = $\frac{2+2}{14} = \frac{4}{14} = \frac{2}{7}$