Introduction to Data Science and Artificial Intelligence – computer class

In class, we have introduced the data science process, the process through which raw data are transformed into information, and then into knowledge. We have seen that among the main components of this process there are: exploring the data, processing the data, feature extraction and building predictive models.

In this computer class, we will address some of those components and at the end solve a simple classification task: distinguish between lemons and oranges.

Data

The fruits dataset we will use was created by Dr. Iain Murray from University of Edinburgh. He bought a few dozen oranges, lemons, mandarins and apples of different varieties, and recorded their measurements in a table. The original data can be downloaded from here:

https://github.com/susanli2016/Machine-Learning-with-Python/blob/master/fruit_data_with_colors.txt

The original dataset comprises four different types of fruit. In this computer class, we will only use oranges and lemons. The corresponding dataset can be found in the Excel file IDAComputerClass.xlsx. The data look as follow (only few rows showed):

fruit label	fruit_name	fruit_subtype	mass	width	height	color_score
1	orange	turkey_navel	142	7.6	7.8	0.75
1	orange	selected_seconds	204	7.5	9.2	0.77
1	orange	turkey_navel	190	7.5	8.1	0.74
2	lemon	spanish_belsan	216	7.3	10.2	0.71
2	lemon	spanish_belsan	200	7.3	10.5	0.72

There are 16 lemons and 19 oranges. For each fruit, four main numeric features were recorded: mass, width, height, and color score.

Part I – Exploring the data

Before doing anything with our data, it is very good practice to have a look at it. This is fundamental in order to understand what type of data we are working with. Looking at their statistical properties, their distribution, identifying missing values and potential outliers, can help to get a good feeling of the data, and think about what is the best way to proceed.

1. Visualization: box plots for each numeric feature will give us a clearer idea of the distribution of the input features, and presence of potential outliers (points well outside the main body of the data).

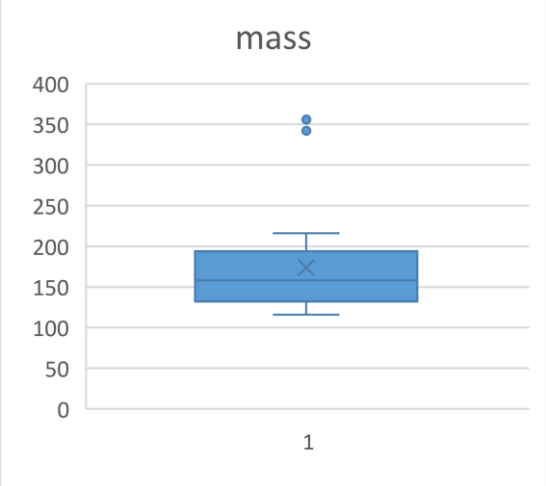
Go to the Excel sheet Part I. Generate a box plot for each of the four features as follows:

- Select the values of the feature of interest (for instance mass)

	A	B	C	D	E	F	G	H
1	fruit_label	fruit label	fruit_name	fruit_subtype	mass	width	height	color_score
2	3	1	orange	spanish_jumbo	362	9.6	9.2	0.74
3	3	1	orange	spanish_jumbo	356	9.2	9.2	0.75
4	3	1	orange	spanish_jumbo	342	9	9.4	0.75
5	3	1	orange	selected_seconds	210	7.8	8	0.82
6	3	1	orange	turkey_navel	180	7.6	8.2	0.79
7	3	1	orange	turkey_navel	142	7.6	7.8	0.75
8	3	1	orange	selected_seconds	204	7.5	9.2	0.77
9	3	1	orange	turkey_navel	190	7.5	8.1	0.74
10	4	2	lemon	spanish_belsan	216	7.3	10.2	0.71
11	4	2	lemon	spanish_belsan	200	7.3	10.5	0.72
12	4	2	lemon	spanish_belsan	196	7.3	9.7	0.72
13	4	2	lemon	spanish_belsan	174	7.3	10.1	0.72

- Insert -> Recommended Charts -> All Charts -> Box & Whiskers (third from the bottom) -> Ok

- You should obtain something like this:



Questions:

1.a. What features seem to show a symmetric distribution? What do not? Why this is important?

1.b. Are there outliers in the data? If yes, do you think they should be removed from the data before further analysis? Why? Please, always try to motivate your answers!

2. Statistical summary. Summary statistics provides a simple but efficient quantitative description of the data. If we simply presented the raw data it would be hard to visualize what the data was showing, especially if there was a lot of it (think of “Big Data”). By computing measures of location (as the average of the data) or spread (as their standard deviation), we can summarize data in a more meaningful way,

which allows simpler interpretation of the data, such that a first level of information can be extracted from it (for example, patterns might emerge from the data).

Fill in the table in the Excel sheet Part I (shown below). You will need to use the following built-in Excel functions: COUNT, AVERAGE, STDEV.S, MEDIAN, QUARTILE.INC. IQR stands for Interquartile Range.

	mass	width	height	color_score
count				
mean				
std				
median				
iqr				

You can call those functions by doing:

-> Formulas -> Insert Function, and look for the function you need.

Please, take some time to understand what property of the data each measure computes. You can also google their names to help you out.

Questions:

2.a. Which ones are measures of location, and which ones of spread?

2.b. Use the summary statistics of each feature to confirm your findings at point 1, and extract additional information if possible.

3. Simple feature selection for classification. Imagine we want to perform a first qualitative classification of lemons and oranges. One of the first questions we would need to address is what features are most discriminative for our problem. In other words, what features allow us optimize the separation between the two fruits. Selecting a subset of relevant features is important because it allows work with simpler models, which generalize better to new data, and are less prone to overfitting.

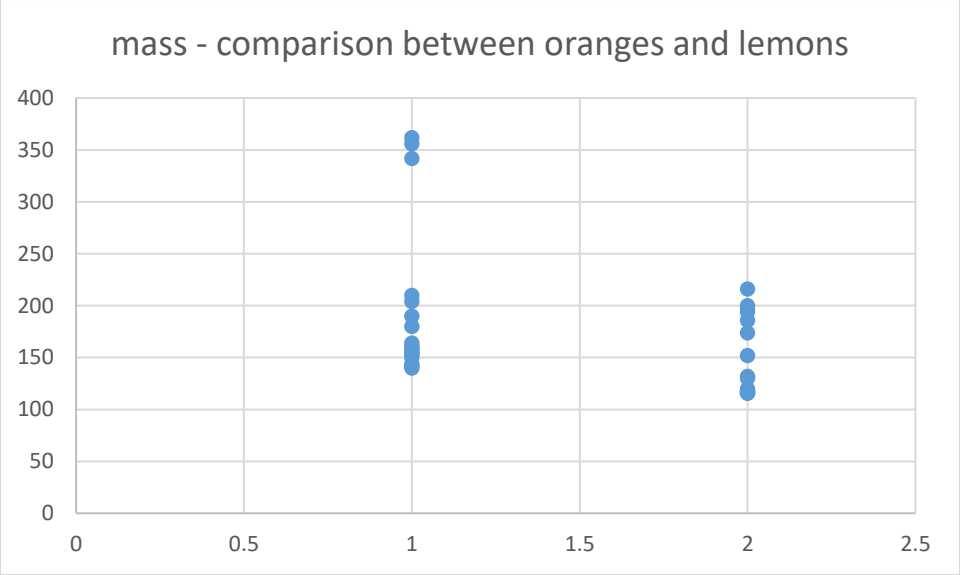
Our task here is to select the most relevant feature among the four features available. A simple approach is to use scatter plots to assess the overlap in the values of a feature for the two fruits. To do so, for each feature select the corresponding values and the values in the column “fruit label” (by holding down Ctrl). For instance (for mass):

	A	B	C	D	E	F	G	H
1	fruit_label	fruit label	fruit_name	fruit_subtype	mass	width	height	color_score
2	3	1	orange	spanish_jumbo	362	9.6	9.2	0.74
3	3	1	orange	spanish_jumbo	356	9.2	9.2	0.75
4	3	1	orange	spanish_jumbo	342	9	9.4	0.75
5	3	1	orange	selected_seconds	210	7.8	8	0.82
6	3	1	orange	turkey_navel	180	7.6	8.2	0.79
7	3	1	orange	turkey_navel	142	7.6	7.8	0.75
8	3	1	orange	selected_seconds	204	7.5	9.2	0.77
9	3	1	orange	turkey_navel	190	7.5	8.1	0.74

Then:

-> Insert -> Charts -> Insert Scatter

You should obtain something like the following:



Questions:

3.a. Based on the information provided by the scatter plots, which one you think is the feature most likely to well separate lemons from oranges? Why?

Part II – Building a decision tree

In this part of the practical we are going to use the same subset of the fruit dataset you have been working with in Part I, and use it to build a decision tree to distinguish between oranges and lemons using the four numeric fields measured.

The algorithm we will use to build the decision tree uses entropy and information gain calculations to decide which split to make next. Each step in our calculations is coloured in the Excel sheet Part II, this document will talk you through the calculations you need to make.

Step 1a – Yellow.

So first we need to calculate the entropy of the whole dataset (the 35 records of oranges and lemons). To do this we use the count of how many samples from the dataset are oranges, how many are lemons and the total number. These values are pre-filled for you in K2:M2.

The entropy is given by, $E= -p^+ \log_2 p^+ – p^- \log_2 p^-$

Where p^+ is the probability of a random fruit being an orange, and p^- is the probability it is a lemon.

- Write a formula to go in the yellow cell N2, which calculates the entropy for this dataset.

The correct value here is 0.994693795 do not go on until you have got this answer!

Step 1b - Yellow

If there are no oranges or lemons in the dataset then the formula should return an entropy of 0. As we will copy and paste this formula to many other places in the sheet, we need it to work in all situations.

- Check that your formula works in this case. If it doesn't then you can use the IFERROR function to catch the error and return the right value.
- REMEMBER to change your inputs back when you are done!

For each variable at each possible split point we are going to calculate the entropy of the dataset on each side of the split, and then make a weighted sum of these and subtract it from the old entropy you just calculated, to get the information gain. For the mass, this has been calculated for you in the first block. Once you have entered the entropy correctly in cell N2, you will be able to see that if we want to split based on mass, then the best point to choose is 136, as below this there are just 9 lemons.

Step 2 – Orange

Now it's your turn. For the width four split points are possible

- Sort the table based on width, from largest to smallest.

I have listed the possible split points for the variable width. The split points are chosen to be half way between the values seen in the samples. To save calculations we only need to consider splits in between the masses of oranges and lemons. So for instance, there is no point in looking at splitting at 9.4, it will always be better to split at 7.4 instead. *Why is this? Convince yourself this is true.*

- Count the number of oranges and number of lemons above and below each split point and fill them in the orange cells.

Step 3 – Green

- Copy and paste the entropy formula from step 1 into the green cells – it should now use the numbers from the orange cells to calculate the entropy of the reduced set if you split it as suggested.

Step 4 – Light blue

- Do the same as above for the samples which are below the split value (NB. You can use your previous counts for those above, and the totals to calculate these automatically)

Step 5 – Dark blue

- The formulas for the weighted entropy and the information gain can be copy and pasted from the mass section above. Check that you understand how they fit with the formula below.

$$Gain(S, A) = E(S) - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} E(S_v)$$

Remember – we have already calculated E(S) and each $\frac{|S_v|}{|S|} E(S_v)$, so this is a simple calculation.

Step 6 – Grey

- Repeat this procedure (copy and paste formulas where you can) to find the information gain for the 10 possible height split points and the 4 possible colour ones.

Step 7 – Choose your first split point

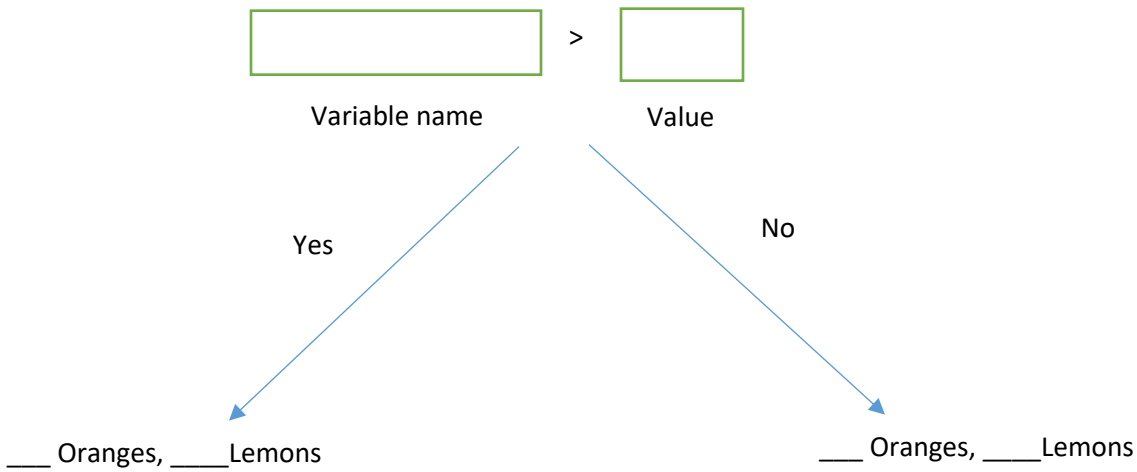
- Using all your calculations, find the split from **any** numeric variable which gives the highest information gain.
- Fill in the diagram on the next page.

Step 8 – Finish the tree

- Copy the samples which are above the chosen split point into a new table. Just through sorting the table find a numerical variable you can use to perfectly classify the samples in this split
- Do the same for the samples on the other branch of the tree.
- Complete the tree on the next page (draw in the new branches below).

Challenge!

Using the same principles, can you perfectly classify the subtypes of apples (in the Excel sheet Challenge) using **only** the color_score? What do you notice about the entropy of this dataset?


>
Variable name
Value
Yes
No
___ Oranges, ____Lemons
___ Oranges, ____Lemons