– Part 1  
**Question 1.1**

Besides the number of instances, what is another main difference between train\_full.txt and train\_sub.txt?

The main difference is that train\_sub.txt is rather unbalanced dataset in comparison to train\_full.txt. From figure 1, it can be observed that all the labels are well represented within the range from 599 to 671 observations. This makes the dataset relatively balanced. On the other hand, when examining figure 2, we can see that the representation of each label is not uniform, ranging from 21 up to 187 observations. This makes this dataset relatively unbalanced.

Figure 1: Train\_full.txt:

Text

Description automatically generated

Figure 2: Train\_sub.txt

Text

Description automatically generated

**Question 1.2**

What kind of attributes are provided in the dataset (Binary? Categorical/Discrete? Integers? Real numbers?) What are the ranges for each attribute in train\_full.txt?

All the attributes are integers. The dependant variable (y vector) contains categorical/discrete strings. The ranges for each attributed, labelled from left to right using integers, e.g. 0, 1, 2 etc., can be found in table 1.

Table 1: ranges of each attribute for train\_full.txt

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Attribute index | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |  |  |  |
| Min value | 0 | 0 | 1 | 0 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 4 | 0 | 1 | 0 | 1 |
| Max value | 10 | 15 | 11 | 12 | 14 | 14 | 14 | 10 | 12 | 13 | 11 | 15 | 12 | 15 | 11 | 14 |

**Question 1.3**

Train\_noisy.txt is actually a corrupted version of train full.txt, where we have replaced the ground truth labels with the output of a simple automatic classifier. What proportion of labels in train\_noisy.txt is different than from those in train full.txt? (Note that the observations in both datasets are the same, although the ordering is different). Has the class distribution been affected? Specify which classes have a substantially larger or smaller number of examples in train\_noisy.txt compared to train\_full.txt.

Figure 3:

**Text, table

Description automatically generated with medium confidence**

Figure 4:

Chart, bar chart

Description automatically generated

# – Part 2

**Task 2.1**

**Task 2.2**

# – Part 3

**Question 3.1**

**Question 3.2**

**Question 3.3**

– Part 4  
**Task 4.1**

**Question 4.1**