Big Data – Assignment 1

Yotam Lifschytz – 209579077, Pan Eyal – 208722058

Question 1:

The code files are attached in the zip file.

Question 2:

1. Chart, line chart

   Description automatically generated

Question 3:

1. is the set of vectors that describes students height and age where:

height is (meters) and age is (years) (the set of all possible examples).

is the group containing the labels ‘drama’ and ‘comedy’ (the set of all possible labels).

|  |  |  |
| --- | --- | --- |
| Height () | Age () |  |
| 160 | 20 | Drama |
| 160 | 40 | comedy |
| 180 | 25 | Drama |
| 180 | 35 | comedy |

2. We have seen in class that:

Where:

Therefore:

|  |  |  |
| --- | --- | --- |
| Height | genre |  |
| 160 | drama | 33% |
| 160 | comedy | 30% |
| 180 | drama | 17% |
| 180 | comedy | 20% |

2. We will show the bayes optimal predictor:

|  |  |
| --- | --- |
| Height () |  |
| 160 | Drama |
| 180 | comedy |

We have seen in class that:

Where:

Therefore:

We can see from the result that optimal bayes error has grown, this is since we narrowed the representation of the examples, thus omitting relevant information – this increases the relative proportion of the probability of all non-maximal-probability labels per each example, thus enlarging the Bayes optimal error – meaning increases for each , and we know that:

So total error increases.

1. We saw in class that the expected error of the Memorize rule is:

and therefor in our case:0.175006

We are allowed to use it because has a deterministic label conditioned on the example.

Question 4:

1. Our objective is to show that for , the output from 1-nearest-neighbor algorithm on with distance ,satisfies the conditions of an ERM algorithm for hypothesis class . meaning:

First, we will show that and then, that and thus we will conclude:

On any given sample size:

To show that we will find an equivalent function from that will receive the same output for any given .

For any we will choose: and define function as follows:

Claim:

Proof:

Given ,

For some :

or or

Case 1: :

And:

Therefor .

Case 2: :

And:

Therefor .

Case 3: :

And:

Therefor .

Case 4: :

And:

Therefor .

Claim:

Proof:

The error defined as:

From algorithm, because , for any :

Thus:

1. For Sample size

And for 3-nearest-neighbor algoritem :

Since 3-nearest-neighbor algoritem do not behave like an ERM algorithm for the hypothesis class .

Question 5:

1. We’ve seen in class that if there are only two labels, , we can set

Thus, the Bayes-optimal predictor error of D as a function of is: