

# Stream Mining

## One-Hot Encoding and DGIM

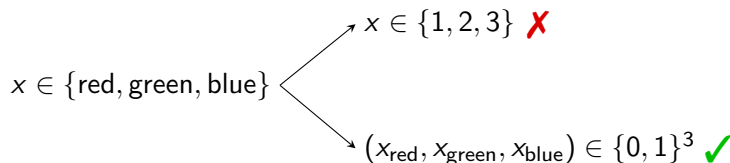
Zeno Adrian Weil

Data Science 1  
Goethe University Frankfurt

7th June 2022

# One-Hot Encoding

- **categorical** features
  - nominal (e.g. colours)
  - ordinal (e.g. satisfaction levels)
- need for numbers in algorithms
- naïve approach: number serially
  - arbitrary orders
  - meaningless differences
- **one-hot encoding**
  - one binary feature for each possible value



# The Datar-Gionis-Indyk-Motwani Algorithm

## Objectives

- **Estimate** the number of **ones** in a bit stream!
- Be **space-efficient**!
- window size  $N$
- $\mathcal{O}(\log_2 N)$  **buckets**
  - **timestamp**
  - **size** = number of ones
    - powers of two
    - one or two of each size
    - sizes never decreasing moving back
  - include all ones; end with ones
- needs only  $\mathcal{O}((\log_2 N)^2)$  **bits**
- **estimation**: sum of sizes of all included buckets + half the size of a partially included bucket (if any)
  - error rate: 50%

...101
10110001 0
 11101
1001 0
 1
1 0

estimation: 16

reality: 14

# The Mushroom Data Set (J.S. Schlimmer, 1987)

- **8124 samples** of 23 mushroom species
  - 4208 edible
  - 3916 poisonous
- **22 attributes** with 128 possible values
- saved as CSV

```
p,x,s,n,t,p,f,c,n,k,e,e,s,s,w,w,p,w,o,p,k,s,u
e,x,s,y,t,a,f,c,b,k,e,c,s,s,w,w,p,w,o,p,n,n,g
e,b,s,w,t,l,f,c,b,n,e,c,s,s,w,w,p,w,o,p,n,n,m
```

...

Are there simple rules to determine edibility? **Yes!**

# The Implementation

- load CSV with Python
- **2D array** for the one-hot encoding of the odour
- Python package **dgim** for the algorithm
- **Streamlit** for the interface
- options
  - odour type
  - window size  $N$
  - error rate

## Topic 4: One-Hot Encoding and DGIM

One-hot encoding denotes the technique of replacing a categorical attribute with  $k$  possible values by a binary  $k$ -ary tuple where the  $i$ -th element is 1 if and only if the attribute was set to the  $i$ -th value. The Datar-Gionis-Indyk-Motwani algorithm is a technique to estimate the number of ones in the last  $N$  bits of a binary string. This program demonstrates the DGIM algorithm on a data set of mushrooms. It estimates the number of edible and poisonous mushrooms for a chosen odour and compares it to the real count.

Please select an odour:

None

Please select a value for  $N$ :

16 205 2048

Please select a maximum absolute value for the error rate of the DGIM algorithm:

1% 50% 100%

☒ Shuffle data

Rerun

### Edible Mushrooms

| Real count | Estimated count | Error   | Number of buckets |
|------------|-----------------|---------|-------------------|
| 205        | 176             | -14.15% | 10                |

### Poisonous Mushrooms

| Real count | Estimated count | Error  | Number of buckets |
|------------|-----------------|--------|-------------------|
| 8          | 6               | -25.0% | 4                 |

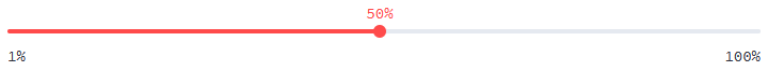
Please select an odour:

None

Please select a value for N:



Please select a maximum absolute value for the error rate of the DGIM algorithm:



Shuffle data

Rerun

### Edible Mushrooms

Real count

205

Estimated count

176

Error

-14.15%

Number of buckets

10

### Poisonous Mushrooms

Real count

8

Estimated count

6

Error

-25.0%

Number of buckets

4

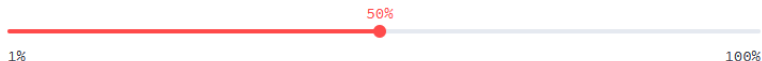
Please select an odour:

None

Please select a value for N:



Please select a maximum absolute value for the error rate of the DGIM algorithm:



Shuffle data

Rerun

### Edible Mushrooms

Real count

214

Estimated count

208

Error

-2.8%

Number of buckets

10

### Poisonous Mushrooms

Real count

11

Estimated count

12

Error

9.09%

Number of buckets

6

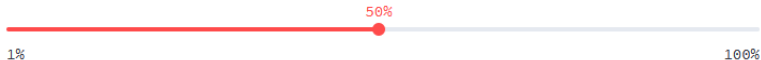
Please select an odour:

None

Please select a value for N:



Please select a maximum absolute value for the error rate of the DGIM algorithm:



Shuffle data

Rerun

### Edible Mushrooms

| Real count | Estimated count | Error  | Number of buckets |
|------------|-----------------|--------|-------------------|
| 1675       | 1872            | 11.76% | 15                |

### Poisonous Mushrooms

| Real count | Estimated count | Error | Number of buckets |
|------------|-----------------|-------|-------------------|
| 67         | 72              | 7.46% | 9                 |



Please select an odour:

None

Please select a value for N:

16

2048

2048

Please select a maximum absolute value for the error rate of the DGIM algorithm:

1%

100%

100%



Shuffle data

Rerun

### Edible Mushrooms

Real count

1651

Estimated count

1872

Error

13.39%

Number of buckets

15

### Poisonous Mushrooms

Real count

59

Estimated count

72

Error

22.03%

Number of buckets

9

Please select an odour:

None

Please select a value for N:



Please select a maximum absolute value for the error rate of the DGIM algorithm:



Shuffle data

Rerun

### Edible Mushrooms

Real count

1678

Estimated count

1672

Error

-0.36%

Number of buckets

410

### Poisonous Mushrooms

Real count

68

Estimated count

68

Error

0.0%

Number of buckets

68

# References

- Project code: <https://github.com/s9770652/DS1-DGIM>
- Mushroom data set:  
<https://archive-beta.ics.uci.edu/ml/datasets/mushroom>
- Streamlit: <https://streamlit.io/>
- Python package *dgim*: <https://pypi.org/project/dgim/>
- Description of one-hot encoding:  
[https://sherbold.github.io/intro-to-data-science/04\\_Data-Analysis-Overview.html#Features](https://sherbold.github.io/intro-to-data-science/04_Data-Analysis-Overview.html#Features)
- Description of the DGIM algorithm:  
<http://infolab.stanford.edu/~ullman/mmds/ch4.pdf>