



University | School of  
of Glasgow | Computing Science

THE AWARDS  
2020

UNIVERSITY  
OF THE YEAR

# Categorical and Continuous Variables

Dr. Fani Deligianni,

[fani.deligianni@glasgow.ac.uk](mailto:fani.deligianni@glasgow.ac.uk)

Lecturer (Assistant Professor)

Lead of the Computing Technologies for Healthcare Theme

<https://www.gla.ac.uk/schools/computing/staff/fanideligianni>

WORLD  
CHANGING  
GLASGOW



# Categorical and Continuous Variables

- Categorical Variables:
  - Nominal
  - Dichotomous
  - Ordinal
- Continuous Variables:
  - Interval
  - Ratio



# Ordinal (Integer) Encoding

Blood Types
A+
A-
B+
B-
AB+
AB-
O



Encoding
1
2
3
4
5
6
7



# One-Hot Encoding

Blood Types
A+
A-
B+
B-
AB+
AB-
O



A+	A-	B+	B-	AB+	AB-	O
1	0	0	0	0	0	0
0	1	0	0	0	0	0
0	0	1	0	0	0	0
0	0	0	1	0	0	0
0	0	0	0	1	0	0
0	0	0	0	0	1	0
0	0	0	0	0	0	1



# Hash encoding

- Hashing converts categorical variables to a high dimensional space of integers
- The distance between two vectors of categorical variables is maintained
- The number of dimensions are significantly less than that of one-hot encoding



# Target (Mean) Encoding

- Takes into account the 'Target'/Predicted variable
- **Binary Target:**
  - When the target attribute  $Y$  is binary,  $Y \in \{0,1\}$ , the transformation maps individual values  $X_i$  of a high-cardinality categorical attribute  $X$  to a scalar,  $S_i$ , representing an estimate of the probability of  $Y=1$  given that  $X=X_i$

$$X \rightarrow S_i = P(Y|X = X_i)$$



# Target (Mean) Encoding - Formulation

- Split dataset into training set containing  $n_{TR}$  records
- Only training samples is taken into account

$$S_i = \frac{n_{iy}}{n_i}$$

$$S_i = \lambda n_i \frac{n_{iy}}{n_i} + (1 - \lambda(n_i)) \frac{n_y}{n_{TR}}$$



$$P(Y|X = X_i) \quad P(Y)$$



# Target (Mean) Encoding

Features

Blood Types
A+
B-
A+
O
O
B-
O

Target

Counts
5
3
6
2
4
1
5

Encoding

Counts
5.5
5.5

Mean





# Target (Mean) Encoding

Features

Blood Types
A+
B-
A+
O
O
B-
O

Target

Counts
5
3
6
2
4
1
5

Encoding

Values
5.5
2
5.5
2

Mean



# Target (Mean) Encoding

Features

Blood Types
A+
B-
A+
O
O
B-
O

\_\_\_\_\_

\_\_\_\_\_

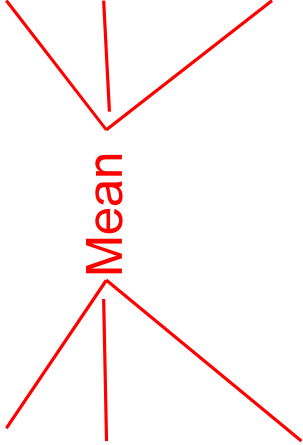
\_\_\_\_\_

Target

Counts
5
3
6
9
4
1
5

Encoding

Values
5.5
2
5.5
6
6
2
6



# Target (Mean) Encoding

- Missing values can be handled by treating them as any other value
- In multiclass classification categorical variable are encoded with  $m-1$  new variables where  $m$  is the number of classes.
- It a more compact representation than one-hot-encoding



# Target (Mean) Encoding - Limitations

- It is very sensitive to the target variable
- Tends to overfit – because it associates every category's value with the same numerical value.
- Few training examples and the average value can get extreme values.
- It does not extract information from intra-category target variable distribution.



# Leave One Out Target Encoding

## Features

Blood Types
A+
B-
A+
O
O
B-
O

## Target

Counts
<del>3</del>
3
6
2
4
1
5

## Encoding

Counts
6

Mean



# Leave One Out Target Encoding

Features

Blood Types
A+
B-
A+
O
O
B-
O

Target

Counts
5
<del>2</del>
6
2
4
1
5

Encoding

Values
6
1

Mean



# Leave One Out Target Encoding

Features

Blood Types
A+
B-
A+
O
O
B-
O

\_\_\_\_\_

\_\_\_\_\_

\_\_\_\_\_

Target

Counts
5
3
6
4
1
5

X

Encoding

Values
6
1
5
4.5

Mean



# Leave One Out Target Encoding

Let  $u$  is the target-encoded value for all samples having category  $C$

$N_c$  is the number of samples having category  $C$

' $j$  belongs to  $C$ '

$$v = \frac{1}{N_c} \sum_{j \in C} y_j$$

$$S_c = \sum_{j \in C} y_j$$

$$v_i = \frac{S_c - y_i}{N_c - 1}$$



# Leave One Out Target Encoding

- Similar to Target (Mean) Encoding
- Mean response over all rows for this category, excluding the row itself
- Avoids direct response leakage
- Reduce the effect of outliers
- Computational efficient



# Target Encoding

- Number of Categories
- Effects of Category Imbalance
- Interaction between variables



# Summary

- One-hot-encoding treats all values of categorical variables equally
- Target (Mean) Encoding encodes categorical variables with conditional mean
- Takes into account the 'Target'/Predicted variable Categorical Variables
  - Binary problems
  - Multi-class problems
  - Prediction Problems
- Leave-One-Out Target encoding reduces overfit
- Interactions often common in EHR might be difficult to be captured with Target Encodings.



# References

- Shreffler et al. Types of Variables and Commonly Used Statistical Designs, <https://www.ncbi.nlm.nih.gov/books/NBK557882/> , 2021.
- Scikit-learn: Encoding Categorical Features, <https://scikit-learn.org/stable/modules/preprocessing.html#encoding-categorical-features>
- Chen et al. ‘Representation Learning for Electronic Health Records: A Survey’, J. Phys Conf, 2020.
- Micci-Barreca. ‘A Pre-processing Scheme for High-Cardinality Categorical Attributes in Classification and Prediction Problems’, ACM SIGKDD Explorations Newsletter 3(1), 2001.