Multi-field categorical data encodings

Adrien Pavao

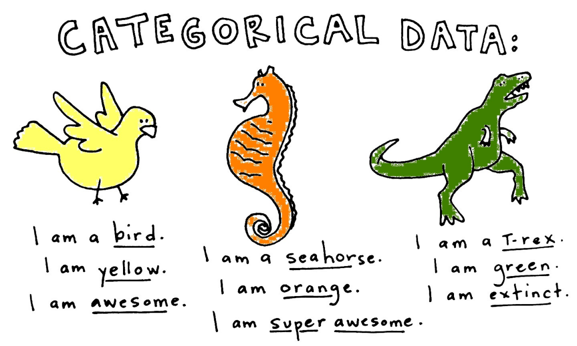
June 22 2018

Laboratoire de Recherche en Informatique

Table of contents

Introduction

Introduction



Definitions

Categorical variable : variable that can take on one of a limited of possible values among nominal categories.

Multi-field : several variables.

Categorical variables are very common, especially in medical records.

Introduction

Problems :

- High **cardinality**
- Algorithms (mainly) take **numerical** variables as input
- Need a **smart encoding** because the machine does not have context about an information

Encodings

Encodings: None (baseline)

As a baseline, categorical variables are simply removed.

Color	Age
Green	38
Blue	24
Red	21

Age
38
24
21

Table 1: Data with and without categorical variable

Encodings: Label

Each category is arbitrarily replaced by a **numerical value**.
Order: dataset, alphabetical or random.

Color	Color encoded
Green	0
Blue	1
Red	2
Blue	1

Table 2: Variable (left) and its label encoding (right)

Encodings: One-hot

Each category value is turned into a binary vector where all columns are equal to zero besides the category column.

Color	Green	Blue	Red
Green	1	0	0
Blue	0	1	0
Red	0	0	1
Blue	0	1	0

Table 3: Variable (left) and its one-hot encoding (right)

Issue : dimensionality

Encodings: One-hot with rare values

If occurence $<$ average occurence * coefficient
Then the category is replaced by “RARE”

Color	Green	Blue	RARE
Green	1	0	0
Blue	0	1	0
Blue	0	1	0
Blue	0	1	0
Green	1	0	0
Red	0	0	1
Blue	0	1	0
Green	1	0	0
Green	1	0	0
Purple	0	0	1

Table 4: Variable (left) and its one-hot encoding (right)

Encodings: Feature hashing

Feature hashing:

- Fix vector size
- For every categorical feature: use a hash function to map values to indices of the feature vector (and one-hot the indice)
- Sum all vectors into one

Encodings: Target

A numerical variable is defined as the target.

Each category is replaced by its mean target value.

Color	Target	Encoded Color
Green	12.5	13.25
Blue	7	8.83
Red	21	21
Blue	9.5	8.83
Blue	10	8.83
Green	14	13.25

Table 5: Variables (left) and target encoding (right)

Likelihood encoding: Target encoding on the principal component of the continuous variables.

Encodings: Frequency

Categories are replaced by their number of occurrences.

Color	Encoded Color
Green	2
Blue	3
Red	1
Blue	3
Blue	3
Green	2
Purple	1

Table 6: Variable (left) and its frequency encoding (right)

Note that frequencies can be transformed into probabilities with a normalization.

Encodings: Deep category embedding

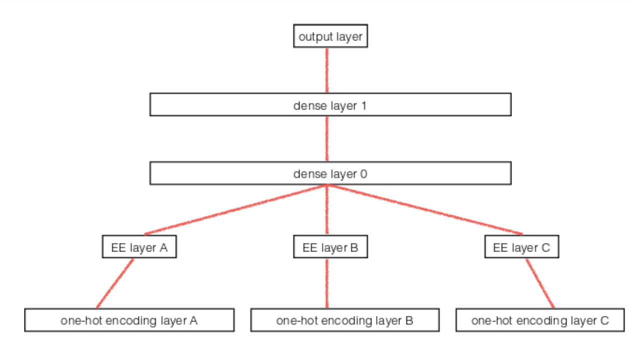


Figure 1: Deep category embedding scheme

Encodings: Word2Vec

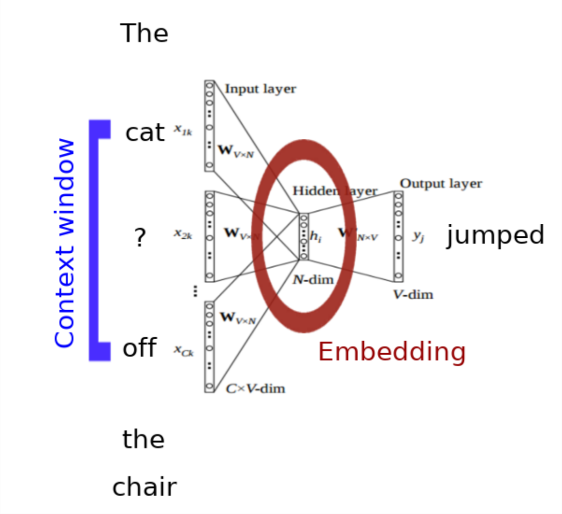


Figure 2: Word2Vec scheme

Encodings: Cat2Vec

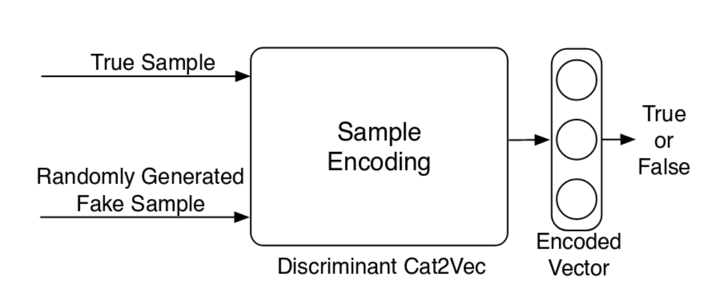


Figure 3: Cat2Vec training scheme

Benchmark

Data and model

Adult dataset

- Features about people: Age, gender, nationality, education, work hours per week, etc.
- Binary classification task: annual income less or greater than 50k
- 6 numerical, 8 categorical variables
- 50 000 instances

Models

- Logistic regression
- Random Forest with 100 estimators

Tuning: Rare one-hot

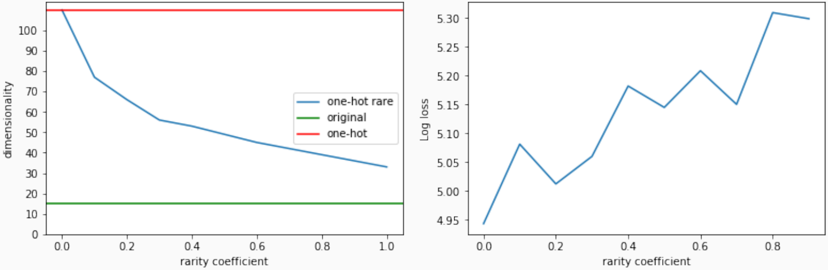


Figure 4: Dimensionality (left) and log loss (right) versus rarity coefficient

Chosen coefficient: 0.2

Tuning: Feature hashing

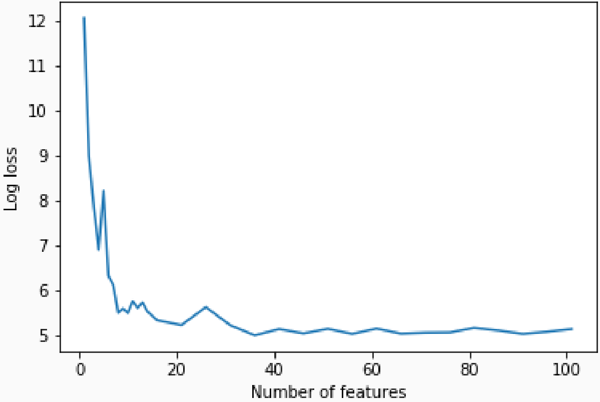


Figure 5: Log loss versus number of features

Chosen number of features: 20

Tuning: Target encoding

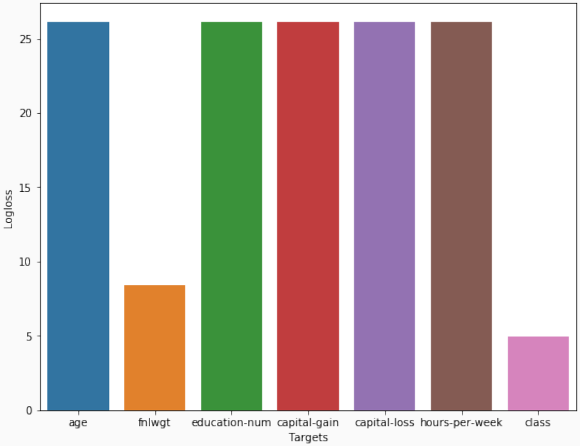


Figure 6: Log loss for each target column

Chosen target: class

Tuning: Word2Vec

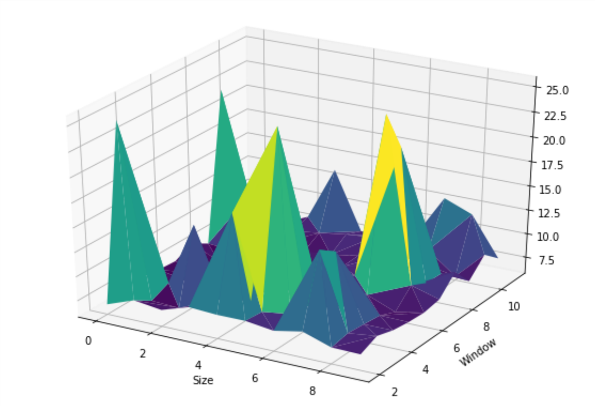


Figure 7: Log loss versus size and window

Chosen parameters: Size 8, window 10 (logloss = 6.19)

Comparison: Number of features

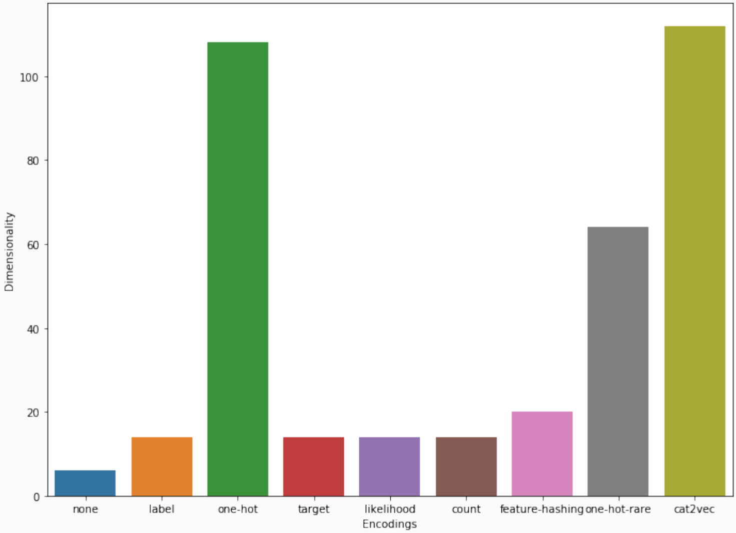


Figure 8: Number of features depending on encoding

Comparison: Computation time

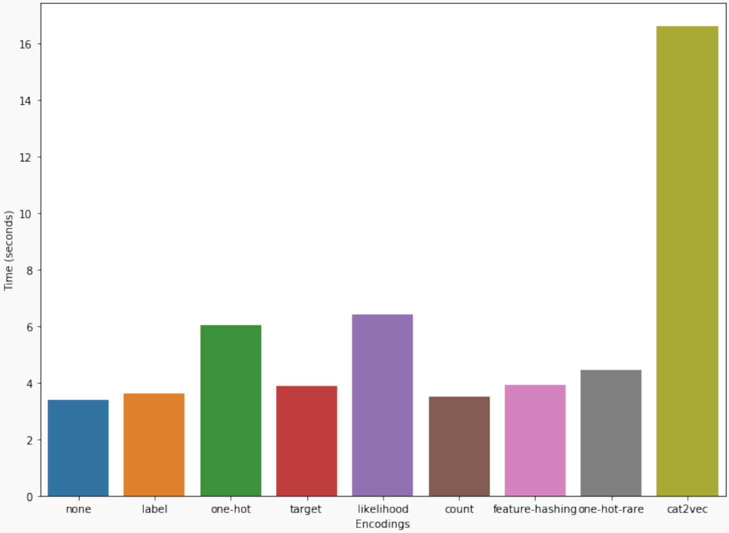


Figure 9: Computation time of encoding and classification

Comparison: Performance

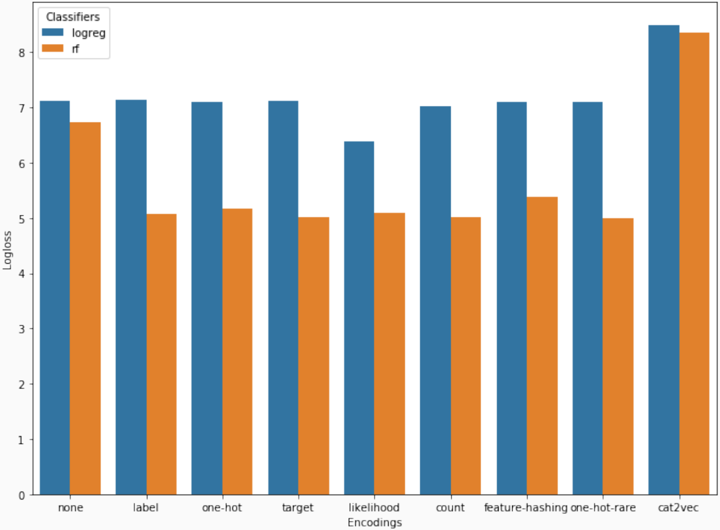


Figure 10: Encodings performance comparison

Conclusion

Conclusion

To go further:

- Try deep embeddings
- Evaluate on other datasets
- Evaluate on data generation

References

[1] Entity Embeddings of Categorical Variables, *Cheng Guo and Felix Berkhahn* (2016)

[2] Cat2Vec: Learning distributed representation of multi-field categorical data, *Ying Wen, Jun Wang, Tianyao Chen and Weinan Zhang* (ICLR 2017)

[3] Feature Hashing for Large Scale Multitask Learning, *Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford and Alex Smola* (2010)