Encoding Data Mining Big Data Data Science Word Definitions, Terminology, and Jargon Machine Learning

What is one hot encoding and when is it used in data science?

8 Answers

Håkon Hapnes Strand, Data Scientist

Updated Dec 20, 2016

One hot encoding transforms categorical features to a format that works better with classification and regression algorithms.

Let's take the following example. I have seven sample inputs of categorical data belonging to four categories. Now, I could encode these to nominal values as I have done here, but that wouldn't make sense from a machine learning perspective. We can't say that the category of "Penguin" is greater or smaller than "Human". Then they would be ordinal values, not nominal.

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

1	Human	1
2	Human	1
3	Penguin	2
4	Octopus	3
5	Alien	4
6	Octopus	3
7	Alien	4

What we do instead is generate one boolean column for each category. Only one of these columns could take on the value 1 for each sample. Hence, the term one hot encoding.

Sample	Human	Penguin	Octopus	Alien
1	1	0	0	0
2	1	0	0	0
3	0	1	0	0
4	0	0	1	0
5	0	0	0	1
6	0	0	1	0
7	0	0	0	1

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

learning libraries can take care of it.

54.1k Views · 247 Upvotes

Related Questions More Answers Below

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Jotham Apaloo, Machine learning, Linux, GTD

Updated Apr 19, 2015 · Upvoted by Ricardo Vladimiro, Game Analytics and Data Science Lead @ Miniclip

Edit: To answer the 'when is it used' portion of the question: Categorical variables are intentionally (for censorship) or implicitly encoded as numerical variables in order to be used as features in any given model.

e.g. [house, car, tooth, car] becomes [0,1,2,1].

This imparts an ordinal property to the variable, i.e. house < car < tooth.

As this is ordinal characteristic is usually not desired, one hot encoding is necessary for the proper representation of the distinct elements of the variable.

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

perhaps unordered factors in it?

One hot encoding transforms:

a single variable with n observations and d distinct values,

to

to d binary variables with n observations each. Each observation indicating the presence (1) or absence (0) of the dth binary variable.

e.g. [house, car, tooth, car] becomes
[[1,0,0,0],
[0,1,0,1],
[0,0,1,0]]

I'm sure there's a nice matrix expression for this, but it doesn't readily come to mind.

41k Views · 60 Upvotes

Nouroz Rahman, BSc Electrical and Electronics Engineering, Bangladesh University of Engineering and Technology (2016)

Answered Mar 17

one hot encoding is assigning 1 to working feature and 0's to other idle features.
Mathematically this is very easy to understand:

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

```
4  [0.0 5.0 0.0]  // one_hot(1)
```

In data science, this is a VERY powerful tool for classification problems, however should be useful for regression and clustering as well but since I have used it only for classification until now, could only say about it.

Qualitatively, it engages all features and tells which is present, and which is absent for a particular set of output.

Mathematically, one hot encoding produces a balanced matrix, which is easy to understand during complex computations inside algorithms.

12.1k Views · 6 Upvotes

Related Questions

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

How is machine learning used in data science?

For how long will the data science industry continue to be hot?

In what professions can one use data?

How does one control the output generated by an LSTM with one-hot encoding?

How are ANOVA methods used in Data Science?

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?

Related Questions

Do we need to apply one-hot encoding to use xgboost?

When converting categorical variables to a numeric representation, how is one-hot encoding different from dummy variable encoding?

In scikit-learn what is the best way to handle categorical features of high cardinality? Using one hot encoder seems to blow up my feature spa...

Does it make a difference in my model if I simply change the data type of a categorical variable to category instead of doing one hot encoding...

Which is better for data analysis: R or Python? Is R still a better data analysis language than Python? Has anyone else used Python with Panda...

Is the hype around data science / big data over?

Are banks using data science? If so, how?

Which one is the future Big Data or Data Science?

What is the best way to one hot encode an array of categorical variables?

How is data science used to fight crime?