Making dummy variables with dummy_cols()

Jacob Kaplan

2024-08-03

Dummy variables (or binary variables) are commonly used in statistical analyses and in more simple descriptive statistics. A dummy column is one which has a value of one when a categorical event occurs and a zero when it doesn't occur. In most cases this is a feature of the event/person/object being described. For example, if the dummy variable was for occupation being an R programmer, you can ask, "is this person an R programmer?" When the answer is yes, they get a value of 1, when it is no, they get a value of 0.

We'll start with a simple example and then go into using the function `dummy_cols()`. You can also use the function `dummy_columns()` which is identical to `dummy_cols()`.

Imagine you have a data set about animals in a local shelter. One of the columns in your data is what animal it is: dog or cat.

animals
dog
dog
cat

To make dummy columns from this data, you would need to produce two new columns. One would indicate if the animal is a dog, and the other would indicate if the animal is a cat. Each row would get a value of 1 in the column indicating which animal they are, and 0 in the other column.

animals	dog	cat
dog	1	0
dog	1	0
cat	0	1

In the function dummy_cols, the names of these new columns are concatenated to the original column and separated by an underscore.

animals	animals_dog	animals_cat
dog	1	0
dog	1	0
cat	0	1

With an example like this, it is fairly easy to make the dummy columns yourself. `dummy_cols()` automates the process, and is useful when you have many columns to general dummy variables from or with many categories within the column.

```
fastDummies_example <- data.frame(numbers = 1:3,
                      gender  = c("male", "male", "female"),
                      animals = c("dog", "dog", "cat"),
                      dates   = as.Date(c("2012-01-01", "2011-12-31",
                                          "2012-01-01")),
                      stringsAsFactors = FALSE)
knitr::kable(fastDummies_example)
```

numbers	gender	animals	dates
1	male	dog	2012-01-01
2	male	dog	2011-12-31
3	female	cat	2012-01-01

The object **fastDummies_example** has two character type columns, one integer column, and a Date column. By default, `dummy_cols()` will make dummy variables from factor or character columns only. This is because in most cases those are the only types of data you want dummy variables from. If those are the only columns you want, then the function takes your data set as the first parameter and returns a data.frame with the newly created variables appended to the end of the original data.

```
results <- fastDummies::dummy_cols(fastDummies_example)
knitr::kable(results)
```

numbers	gender	animals	dates	gender_female	gender_male	animals_cat	animals_dog
1	male	dog	2012-01-01	0	1	0	1
2	male	dog	2011-12-31	0	1	0	1
3	female	cat	2012-01-01	1	0	1	0

In some situations, you would want columns with types other than factor and character to generate dummy variables. For example, a column of years would be numeric but could be well-suited for making into dummy variables depending on your analysis. Use the *select_columns* parameter to select specific columns to make dummy variables from.

```
results <- fastDummies::dummy_cols(fastDummies_example, select_columns = "numbers")
knitr::kable(results)
```

numbers	gender	animals	dates	numbers_1	numbers_2	numbers_3
1	male	dog	2012-01-01	1	0	0
2	male	dog	2011-12-31	0	1	0
3	female	cat	2012-01-01	0	0	1

The final option for `dummy_cols()` is *remove_first_dummy* which by default is FALSE. If TRUE, it removes the first dummy variable created from each column. This is done to avoid multicollinearity in a multiple regression

model caused by included all dummy variables. The “first” dummy variable is the one at the top of the rows (i.e. the first value that is not NA).

```
results <- fastDummies::dummy_cols(fastDummies_example, remove_first_dummy = TRUE)
knitr::kable(results)
```

numbers	gender	animals	dates	gender_male	animals_dog
1	male	dog	2012-01-01	1	1
2	male	dog	2011-12-31	1	1
3	female	cat	2012-01-01	0	0