Quiz M1.03

Note: For each question **make sure you select all of the correct options—** there may be more than one! Don't forget to use the sandbox notebook if you need.

Question 1 (1/1 point)

How are categorical variables represented?

- [ ] a) categorical feature is only represented by non-numerical data
- [x] b) categorical feature represents a finite number of values called categories
- [x] c) categorical feature can either be represented by numerical or non-numerical values

✔

Select all answers that apply

You have used 1 of 2 submissions

Question 2 (1/1 point)

An ordinal variable:

- [ ] a) is a categorical variable with a large number of different categories
- [x] b) can be represented by integers or string labels
- [x] c) is a categorical variable with a meaningful order

You have used 1 of 2 submissions

Question 3 (1/1 point)

One-hot encoding:

- [ ] a) encodes each column with string-labeled values into a single integer-coded column
- [ ] b) transforms a numerical variable into a categorical variable
- [x] c) creates one additional column for each possible category
- [x] d) transforms string-labeled variables using a numerical representation

✔

Select all answers that apply

You have used 1 of 2 submissions

Question 4 (1/1 point)

Assume we have a dataset where each line describes a company. Which of the following columns should be considered as a meaningful **numerical feature** to train a machine learning model to classify companies:

- [ ] a) the sector of activity ("construction", "retail", "energy", "insurance"...)
- [ ] b) the phone number of the sales department
- [x] c) the number of employees ✔

- [ ] e) the post code of the head quarters

✔

Select all answers that apply

EXPLANATION

Solution: c) d)

The number of employees (an integer count) and the profits (expressed in a given currency, possibly with a decimal representation) are both quantities and can meaningfully be treated as numerical features.

The sector of activity is typically represented by a string identifier with a choice among a fixed list of possible values. It is therefore a canonical example of a nominal categorical value and therefore has a no numerical interpretation.

A phone number can be represented by an integer number but cannot be interpreted as a quantity. Also note: while phone numbers could be treated as categorical values, they are typically unique to each companies and therefore have no predictive value (for instance they would never on both sides of a train/test split). Such columns with unique identifiers are typically just dropped from the feature list of machine learning pipelines.

A post code does not represent a quantity either and the relative order of post codes is often arbitrary. Therefore it would not make sense to treat this column as a numerical feature. While a post code is typically unique to a specific geographic area, several companies in the database can share the same post code. One could therefore decide to treat it as a categorical variable. It could also be possible to extract parts of the post code (e.g. the two or three leadning digits) to extract categorical variable that represent coarser administrative regions.

Other variables such a credit rating ("AAA", "AA", "A", "B", "C"...) can sometimes be treated as a numerical variable (once converted to an integer) to express the ordering information (ordinal variable)

You have used 1 of 2 submissions

YOUR EXPERIENCE

According to you, this whole 'Handling categorical data' lesson was:

- ○ **Too easy, I got bored**
- ○ **Adapted to my skills**
- ○ **Difficult but I was able to follow**
- ○ **Too difficult**

Submit

To follow this lesson, I spent:

- ○ **less than 30 minutes**
- ○ **30 min to 1 hour**
- ○ **1 to 2 hours**
- ○ **2 to 4 hours**
- ○ **more than 4 hours**
- ○ **I don't know**

Submit

FORUM (EXTERNAL RESOURCE)

Topic	Replies	Last reply
About the M1. Quiz M1.03 category brospars 27 Oct	0	27 Oct brospars

There are no more M1. Quiz M1.03 topics. Ready to start a new conversation?

Terms and conditions