Working with Categorical Data: Test Your Knowledge

Let's do a quick test! You must answer at least 4 questions correctly to pass this quiz.

1. Which of the following are examples of categorical data? (Choose all that apply)

Choose as many answers as you see fit.

- [ ] Number of pages in a book
- [x] Telephone number
  - ✓ Telephone numbers are categorical data, as their values cannot be meaningfully multiplied.
- [x] Type of french fries (curly, crinkle-cut, steak-cut, waffle)
  - ✓ Type of french fries is categorical data, as this feature accepts one of four categories.
- [x] Star rating (1 to 5 stars) for a restaurant, where 1 star indicates "poor" and 5 stars indicates "excellent"
  - ✓ Star ratings are categorical data, as their values cannot be meaningfully multiplied.

2. True or False: Machine labels are generally considered more desirable than labels provided by human raters.

- ○ True
- ◉ False
  - ✓ Labels provided by humans are often called "gold labels," and are generally considered to be more desirable than machine-generated labels.

3. You are training a model on a training dataset that includes the feature `eye_color`, which can be one of the following six values: `amber`, `blue`, `brown`, `gray`, `green`, `hazel`. Which of the following are valid encodings for an `eye_color` value of `blue`? (Choose all that apply)

Choose as many answers as you see fit.

- [x] [0, 1, 0, 0, 0, 0]
  - ✓ This is a valid one-hot encoding for blue. The array item corresponding to blue has a value of 1, and all other items in the array have a value of 0.
- [x] [1]
  - ✓ This is a valid sparse representation for blue, representing just the array item corresponding to blue (assuming a zero-indexed array).
- [ ] [1, 2, 3, 4, 5, 6]
- [ ] [0, 1]
- [ ] [1, 0, 2, 3, 4, 5]

4. In which of the following scenarios would it make sense to apply feature hashing?

- ◉ The number of categorical feature values is very large.
  - ✓ In cases where a categorical feature has a large number of possible feature values (also called a high number of dimensions), one-hot encoding is typically a bad choice. Instead, you may want to use hashing or embedding.
- ○ The number of categorical feature values is very small.
- ○ The model is being trained offline.
- ○ All the possible values of the categorical feature can be enumerated in advance.

5. You are performing a feature cross of the following two categorical features:

- `apple_color`, which takes one of these four values: `green`, `red`, `white`, or `yellow`
- `apple_texture`, which takes one of these two values: `crisp` or `mushy`

How many entries are in the resulting feature-cross vector?

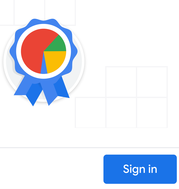
- ○ 1
- ○ 2
- ○ 6
- ◉ 8
  - ✓ In a feature cross of apple_color and apple_texture, each possible color value is paired with each possible texture value, resulting in 8 entries in the feature-cross vector.

Results

You scored 5 out of 5. Congratulations! You have passed this quiz.

You've unlocked a new milestone: **Machine Learning Crash Course: Categorical data!**

Sign in to claim your badge.



Share

Sign in

Was this helpful?