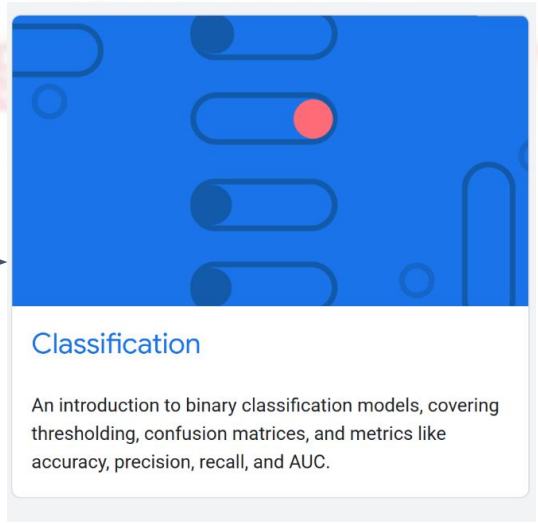
TECHCRUSH ARTIFICIAL INTELLIGENCE BOOTCAMP

Facilitator: Hammed Obasekore
September 17th, 2025

Recap



Categorical data

Data that has a specific set of possible values.
For instance,

- Different varieties of rice
- Street names
- Binned numbers
- Discretized number

Likewise, features that contain integer values are as categorical data instead of numerical data. For example, consider a postal code feature in which the values are integers.

Categorical data

Encoding means converting categorical or other data to numerical vectors that a model can train on.

- encode as a vocabulary: categorical feature has a low number of possible categories

Feature name	# of categories	Sample categories
snowed_today	2	True, False
skill_level	3	Beginner, Practitioner, Expert
season	4	Winter, Spring, Summer, Autumn
day_of_week	7	Monday, Tuesday, Wednesday
planet	8	Mercury, Venus, Earth

Categorical data

- Index numbers, that is convert each string to a unique index number

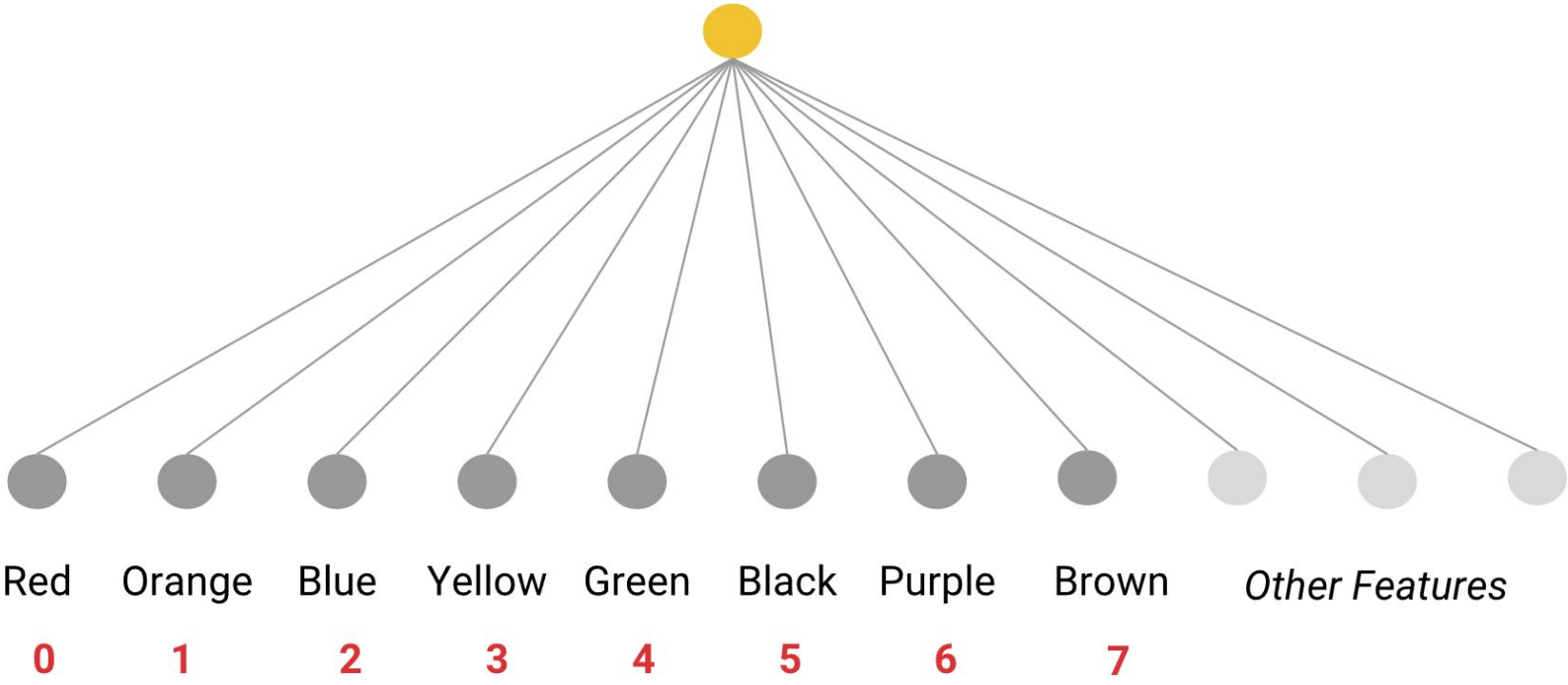


Figure 2. *Indexed features.*

Understanding Data

Categorical data

One-hot encoding
convert each index number to its one-hot encoding.
In a one-hot encoding:

Each category is represented by a vector (array) of N elements, where N is the number of categories.
Exactly one of the elements in a one-hot vector has the value 1.0;
all the remaining elements have the value 0.0.

For example, the following table shows the one-hot encoding for each color in `car_color`:

Feature	Red	Orange	Blue	Yellow	Green	Black	Purple	Brown
"Red"	1	0	0	0	0	0	0	0
"Orange"	0	1	0	0	0	0	0	0
"Blue"	0	0	1	0	0	0	0	0
"Yellow"	0	0	0	1	0	0	0	0
"Green"	0	0	0	0	1	0	0	0
"Black"	0	0	0	0	0	1	0	0
"Purple"	0	0	0	0	0	0	1	0
"Brown"	0	0	0	0	0	0	0	1

Categorical data

One-hot encoding

The following illustration suggests the various transformations in the vocabulary representation:

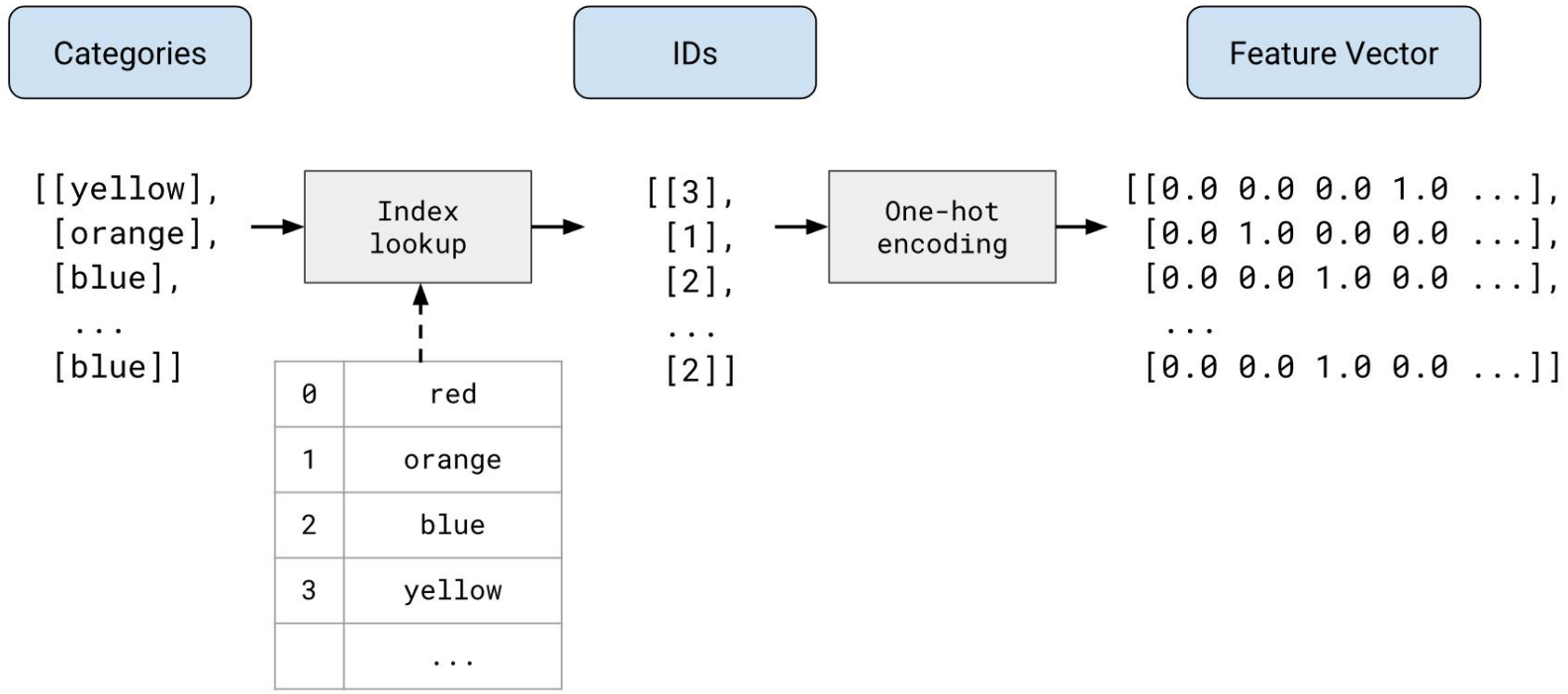


Figure 3. *The end-to-end process to map categories to feature vectors.*

Categorical data

Sparse representation
A feature whose values are predominantly zero (or empty) is termed a sparse feature.

Sparse representation means storing the position of the 1.0 in a sparse vector. For example, the one-hot vector for "Blue" is:

[0, 0, 1, 0, 0, 0, 0, 0]

Categorical data

Encoding high-dimensional categorical features

When the number of categories is high, one-hot encoding is usually a bad choice. Embeddings, detailed in a separate Embeddings module, are usually a much better choice. Embeddings substantially reduce the number of dimensions, which benefits models in two important ways:

- The model typically trains faster.
- The built model typically infers predictions more quickly. That is, the model has lower latency.

Some categorical features have a high number of dimensions, such as those in the following table:

Feature name	# of categories	Sample categories
words_in_english	~500,000	"happy", "walking"
US_postal_codes	~42,000	"02114", "90301"
last_names_in_Germany	~850,000	"Schmidt", "Schneider"

Questions?