Introduction to K-Nearest Neighbors: Takeaways

Syntax

- Randomizing the order of a DataFrame:

```
import numpy as np
np.random.seed(1)
np.random.permutation(len(dc_listings))
```

- Returning a new DataFrame containing the shuffled order:

```
dc_listings = dc_listings.loc[np.random.permutation(len(dc_listings))]
```

- Applying string methods to replace a comma with an empty character:

```
stripped_commas = dc_listings['price'].str.replace(',', '')
```

- Converting a Series object to a float datatype:

```
dc_listings['price'] = dc_listings['price'].astype('float')
```

Concepts

- Machine learning is the process of discovering patterns in existing data to make a prediction.
- In machine learning, a feature is an individual measurable characteristic.
- When predicting a continuous value, the main similarity metric that's used is Euclidean distance.
- K-nearest neighbors computes the Euclidean Distance to find similarity and average to predict an unseen value.
- Let q_1 to q_n represent the feature values for one observation, and p_1 to p_n represent the feature values for the other observation then the formula for Euclidean distance is as follows:

$$d = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \ldots + (q_n - p_n)^2}$$

- In the case of one feature (univariate case), the Euclidean distance formula is as follows:

$$d = \sqrt{(q_1 - p_1)^2}$$

Resources

- K-Nearest Neighbors
- Five Popular Similarity Measures