Clustering basics: Takeaways

Syntax

- Computing the Euclidean distance in Python:

```
from sklearn.metrics.pairwise import euclidean_distances

euclidean_distances(votes.iloc[0,3:], votes.iloc[1,3:])
```

- Initializing the KMeans class from scikit-learn:

```
from sklearn.cluster import KMeans

kmeans_model = KMeans(n_clusters=2, random_state=1)
```

- Calculating the distance between observations and the clusters:

```
senator_distances = kmeans_model.fit_transform(votes.iloc[:, 3:])
```

- Computing a frequency table of two or more factors:

```
labels = kmeans_model.labels_

print(pd.crosstab(labels, votes["party"]))
```

Concepts

- Two major types of machine learning are supervised and unsupervised learning. In supervised learning, you train an algorithm to predict an unknown variable from known variables. In unsupervised learning, you're finding patterns in data as opposed to making predictions.
- Unsupervised learning is very commonly used with large data sets where it isn't obvious how to start with supervised machine learning. It's a good idea to try unsupervised learning to explore a data set before trying to use supervised machine learning models.
- Clustering is one of the main unsupervised learning techniques. Clustering algorithms group similar rows together and is a key way to explore unknown data.
- We can use the Euclidean distance formula to find the distance between two rows to group similar rows. The formula for Euclidean distance is:

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

  where q_n and p_n are observations from each row.

- The k-means clustering algorithm uses Euclidean distance to form clusters of similar items.

Resources

- Documentation for sklearn.cluster.KMeans
- Unsupervised Machine learning
- Redefining NBA Basketball Positions