Clustering

Two groups

Five groups

Clustering on features

- Function
- Color
- Width
- Height
- Weight
- Price
- Availability in local warehouse

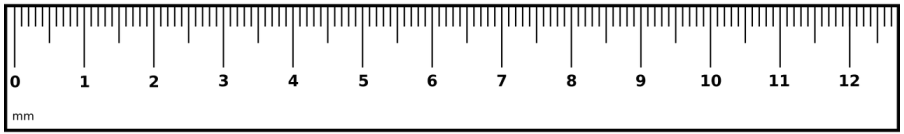
⋮

Clustering:

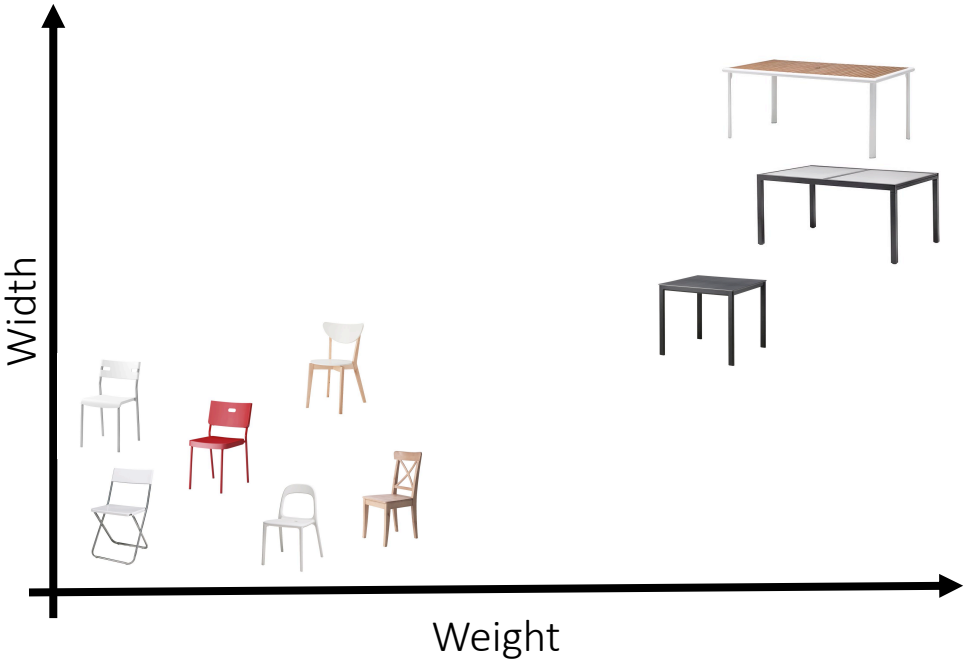
Grouping things by how similar they are

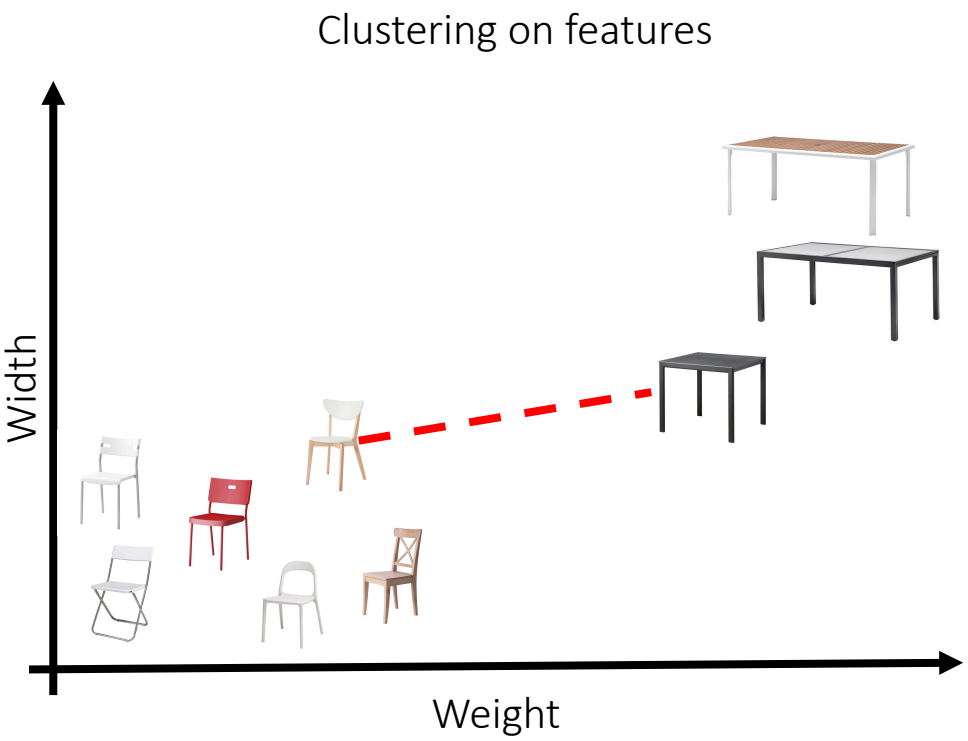
How do we quantify ”similarity”?

How do we quantify ”similarity”?

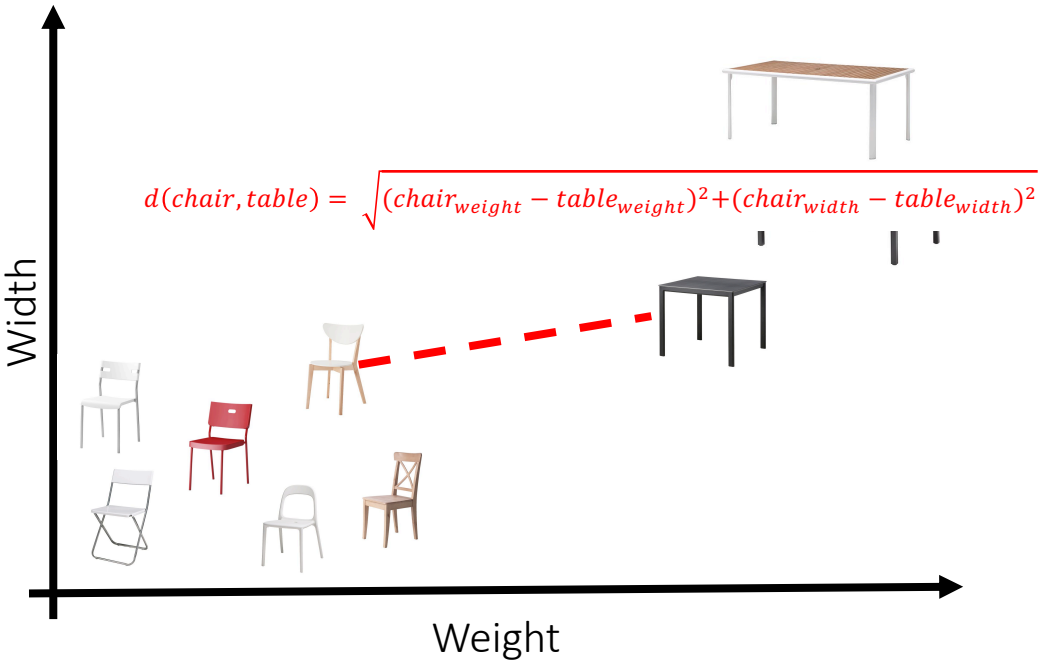


Clustering on features

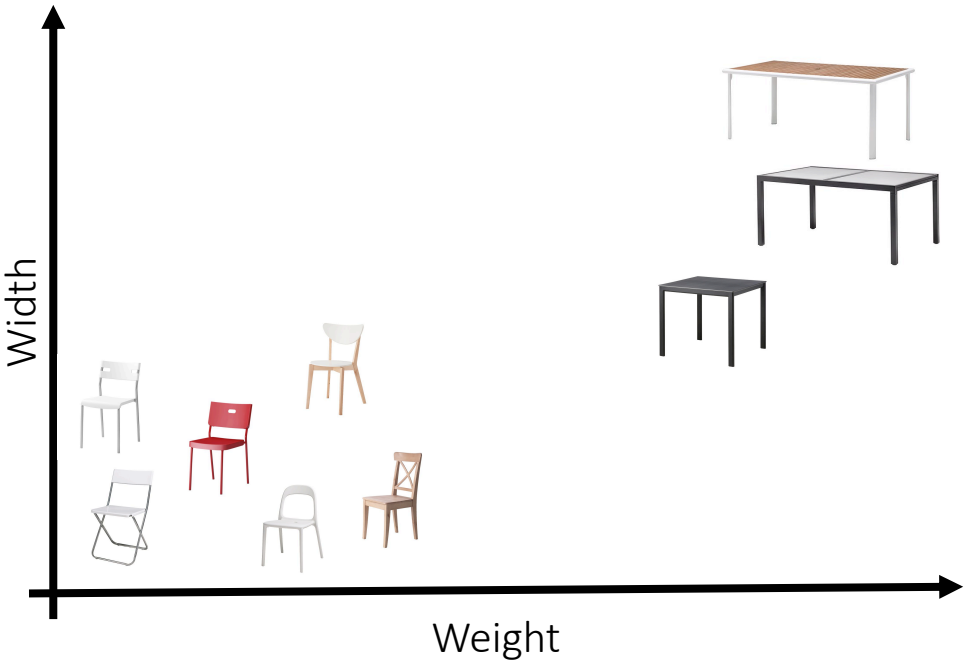



Clustering on features
Width
Weight

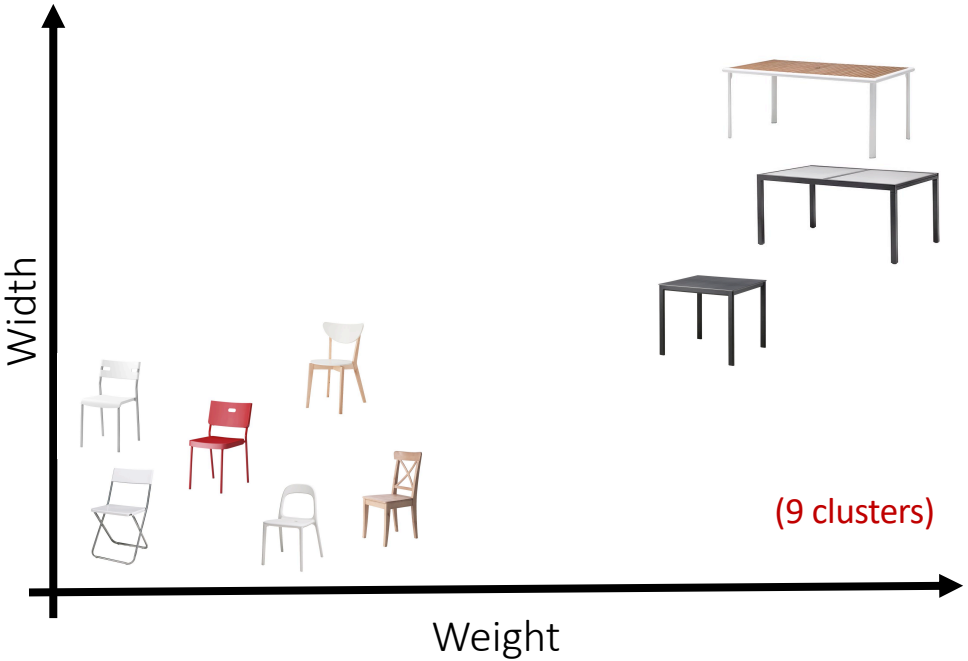
Clustering on features



Agglomerative hierarchical clustering

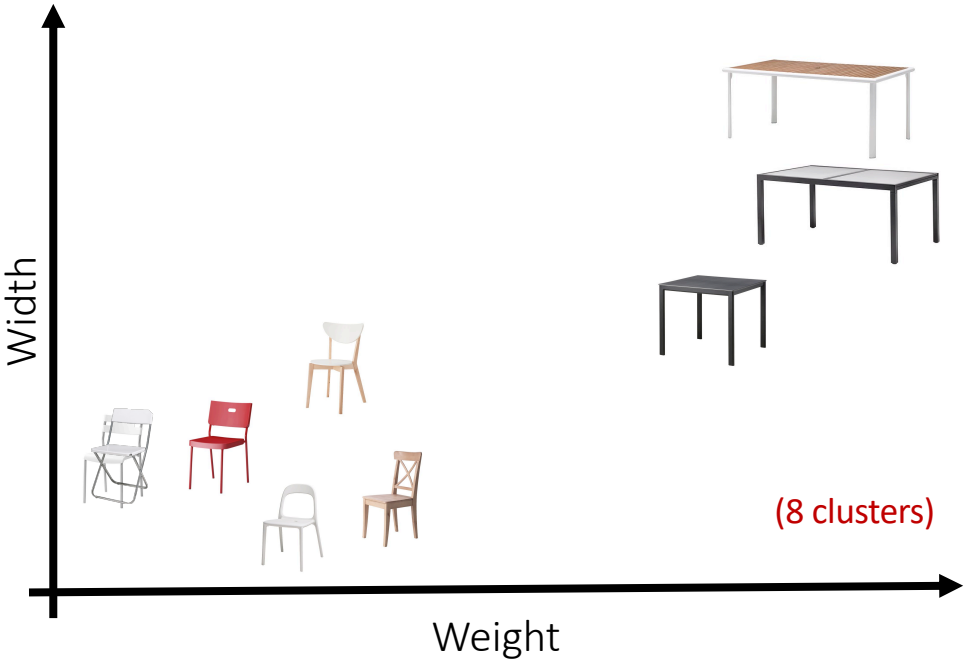


Agglomerative hierarchical clustering



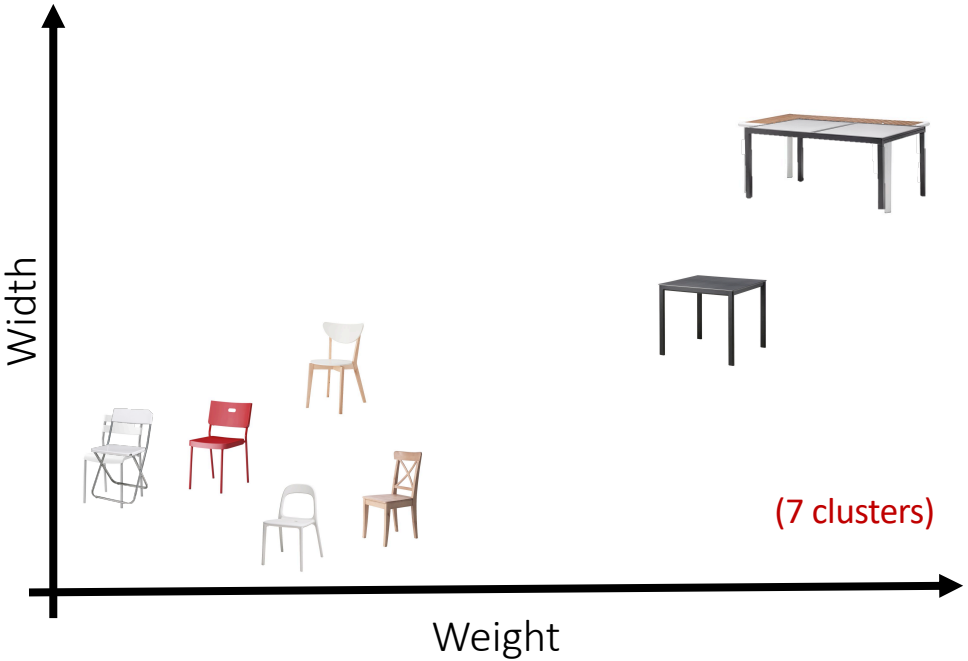
Clustering on features

Clustering on features

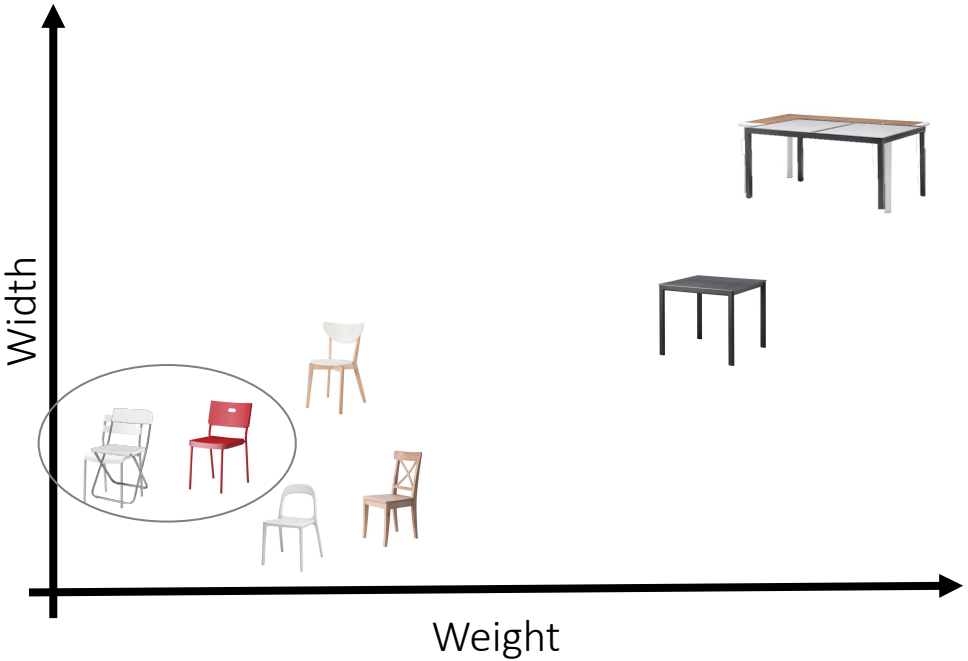


Clustering on features

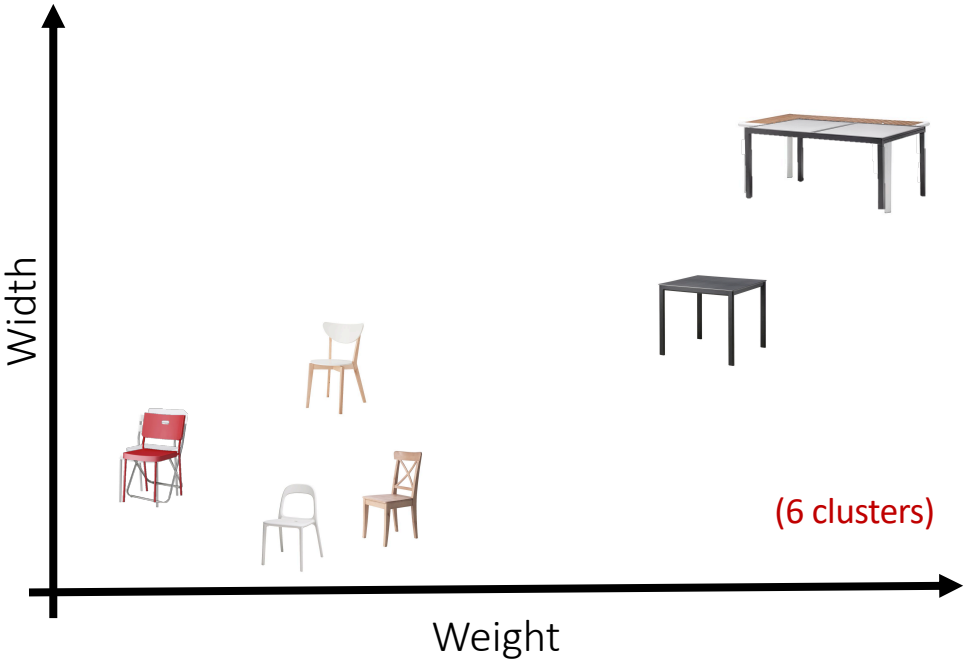
Clustering on features



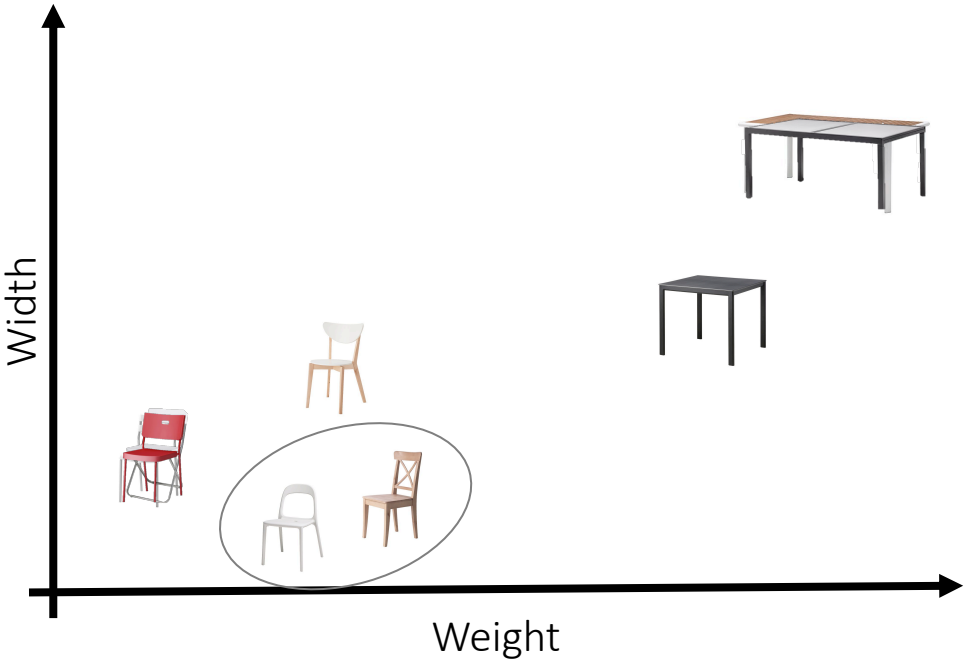
Clustering on features



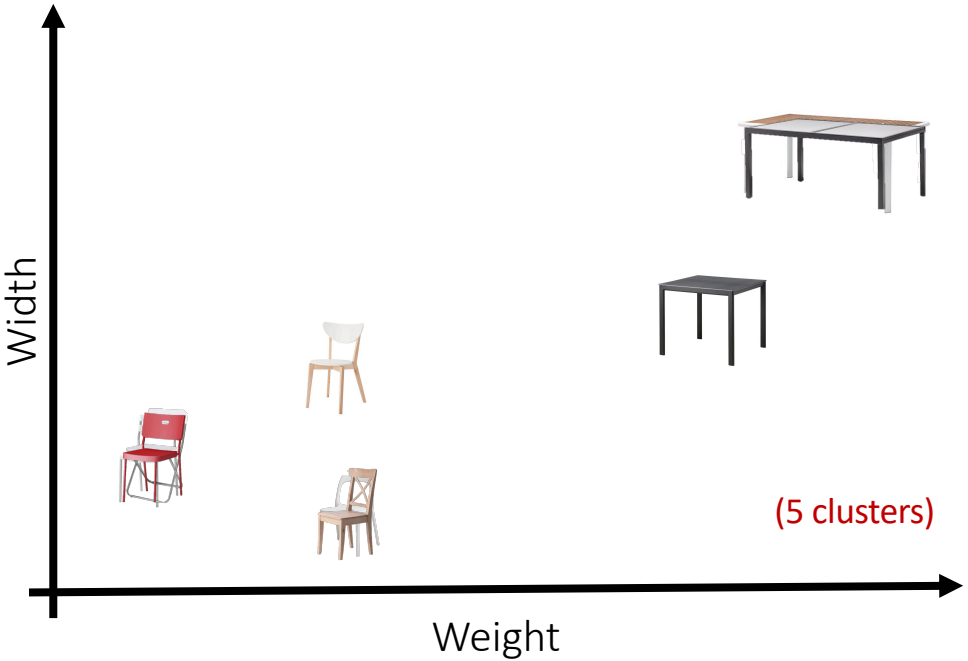
Clustering on features



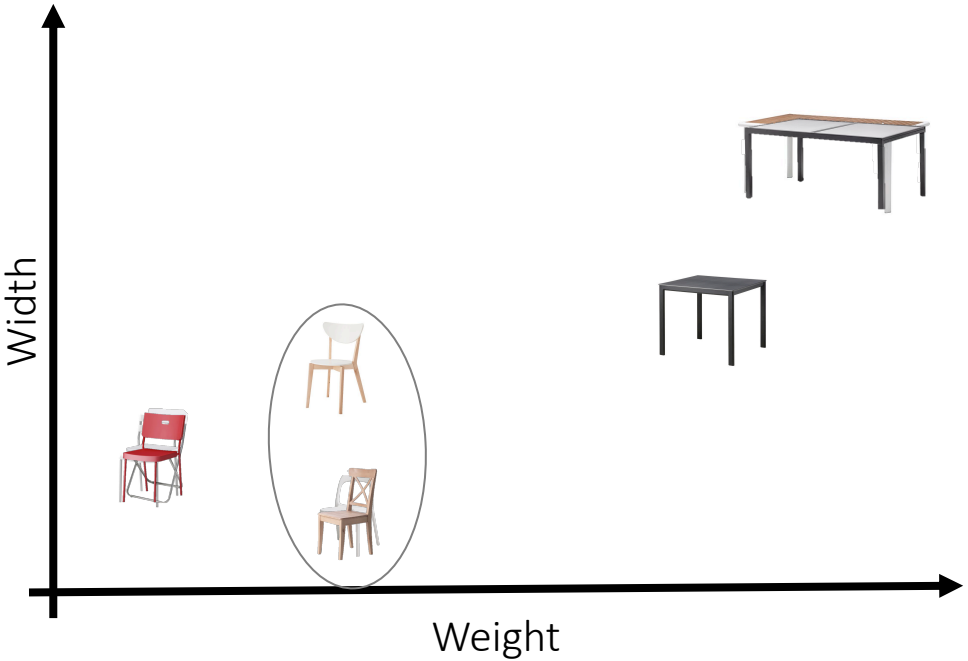
Clustering on features



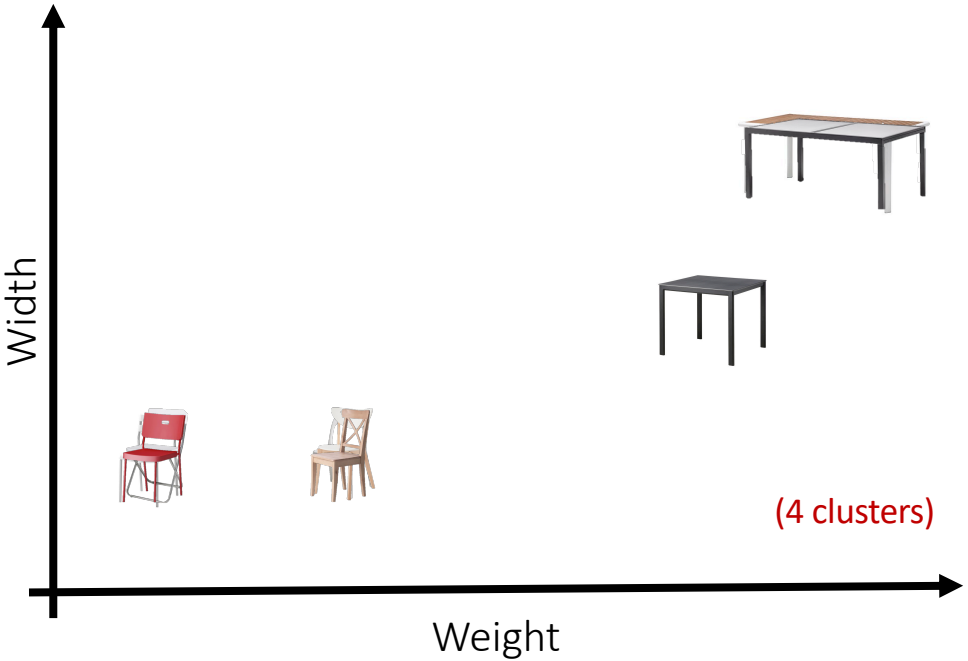
Clustering on features



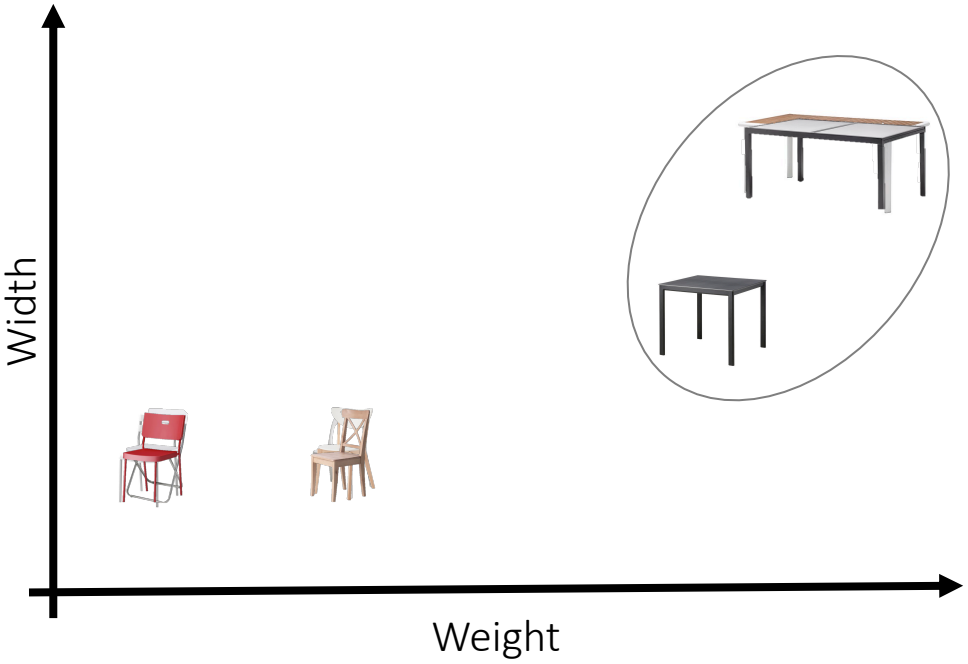
Clustering on features



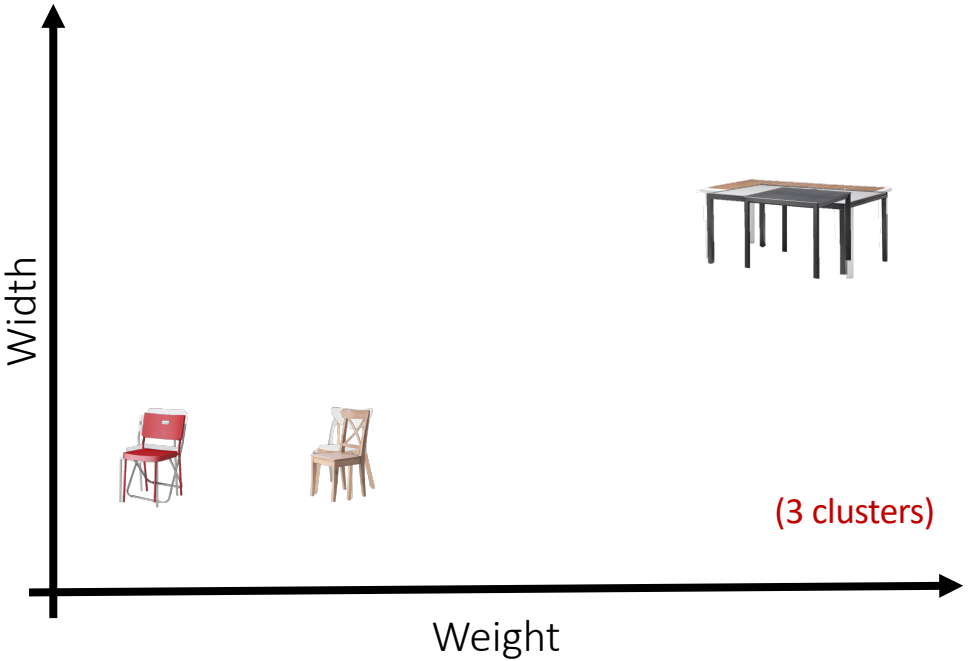
Clustering on features



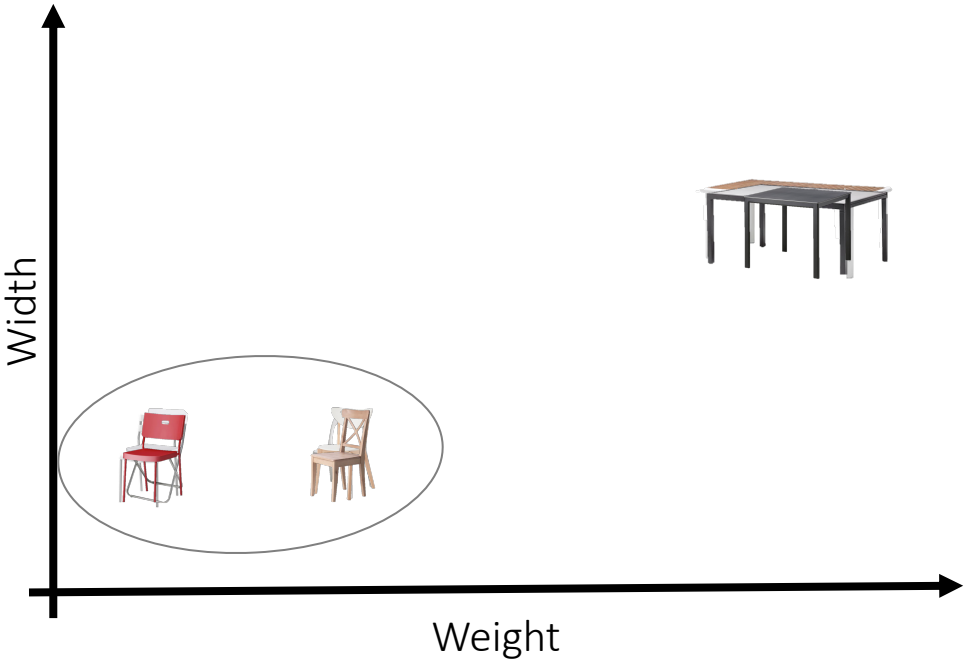
Clustering on features



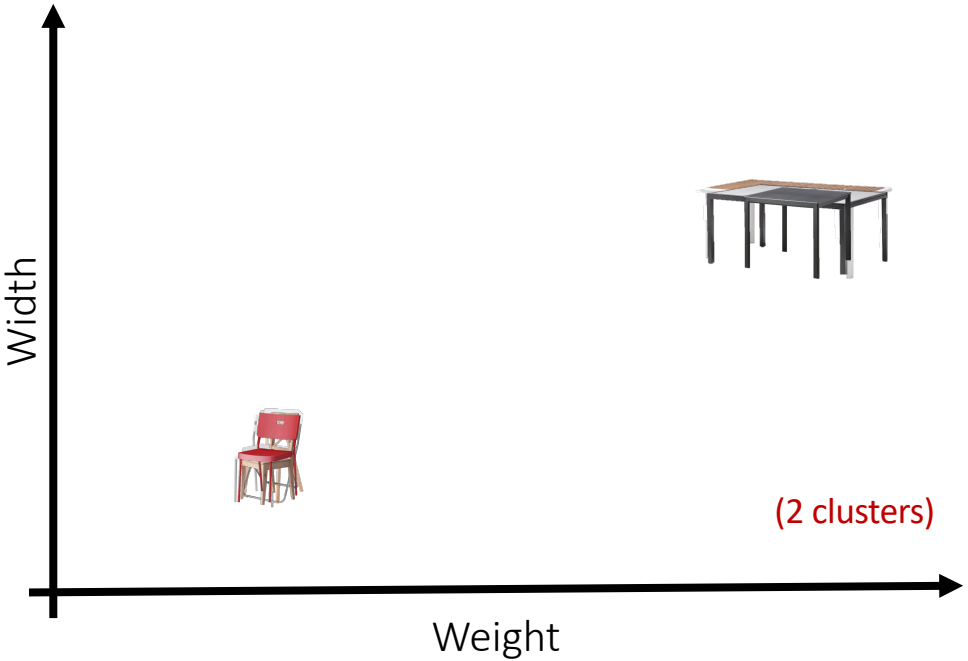
Clustering on features



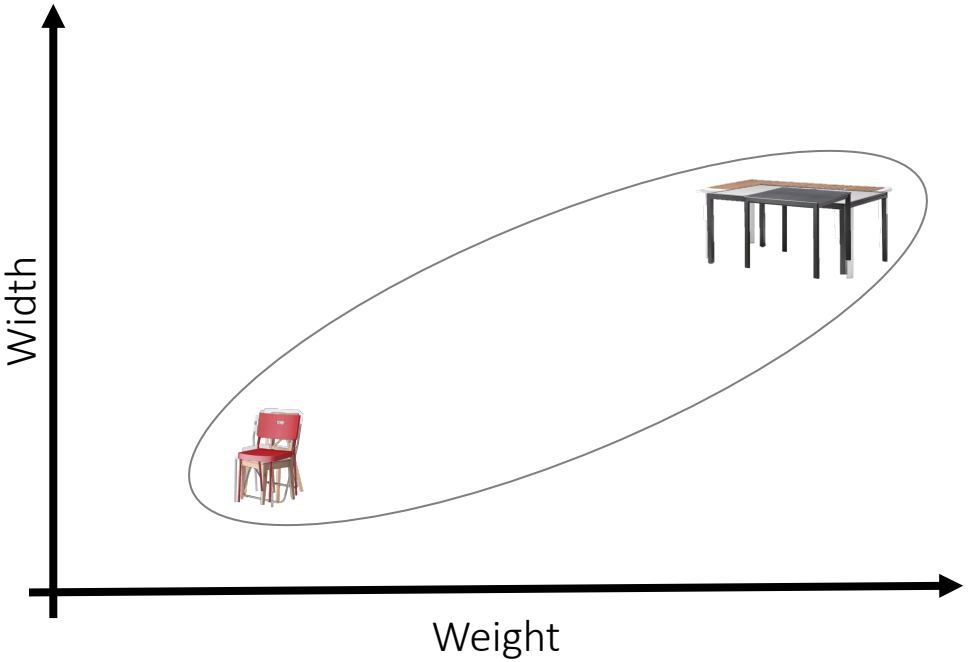
Clustering on features



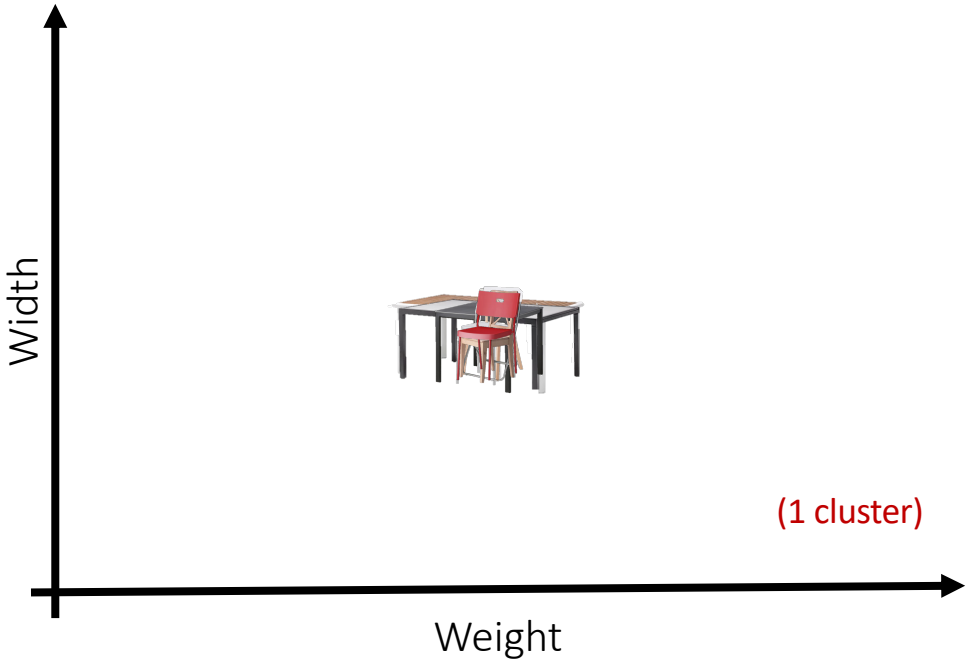
Clustering on features



Clustering on features



Clustering on features



How many clusters?

How many clusters?

Depends on what you are trying to achieve

and whether your features carry that information

Selecting appropriate clustering algorithm

Here we went through hierarchical (agglomerative) clustering.

Other notable algorithms include

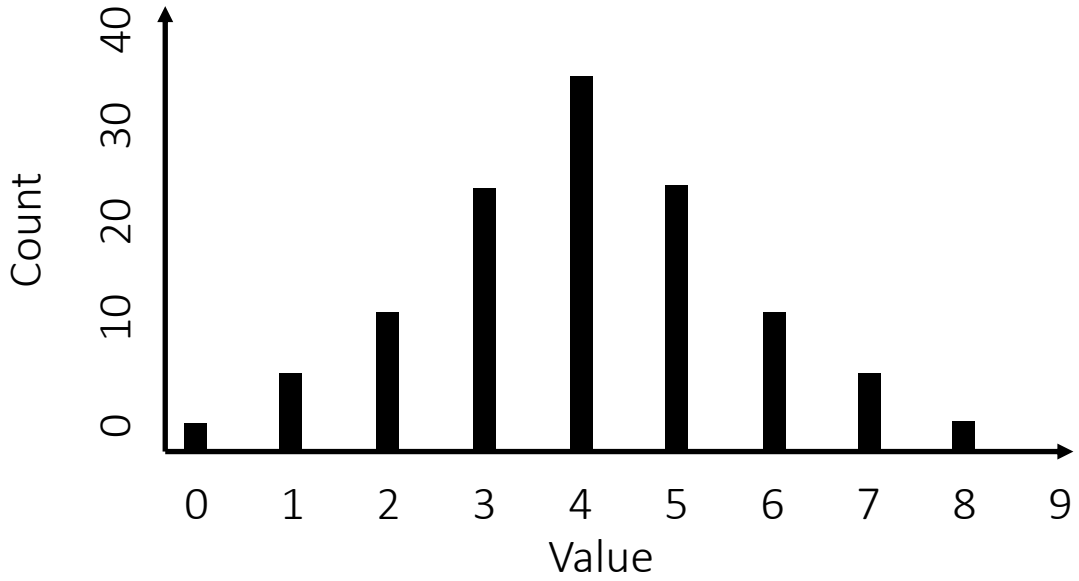
- *k*-means clustering
- Mean shift clustering
- DBSCAN
- Expectation–Maximization clustering

Selecting distance metrics

Here we used Euclidean distance, which will work for the vast majority of (normally distributed) expression data.

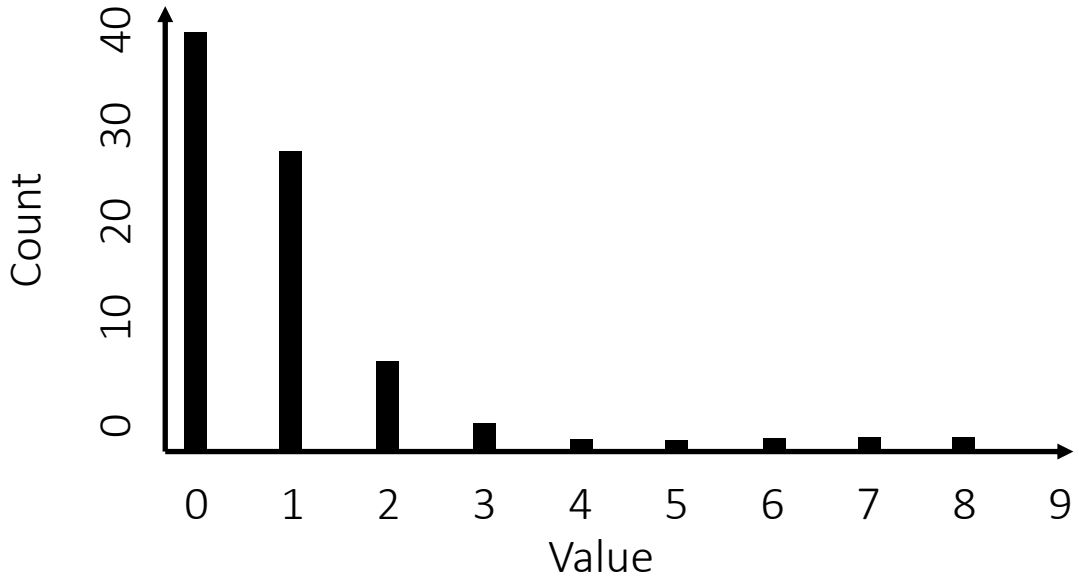
Value distributions

Normal distribution



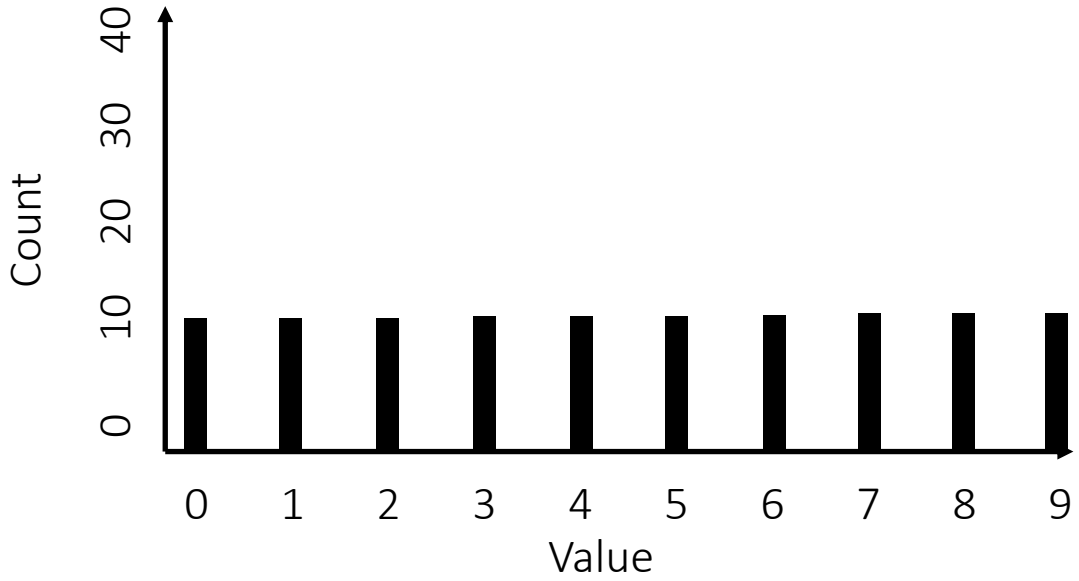
Value distributions

Negative binomial distribution



Value distributions

Value rank distribution



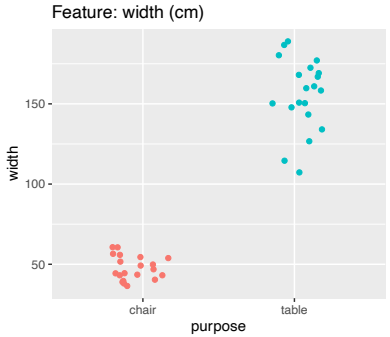
Selecting distance metrics

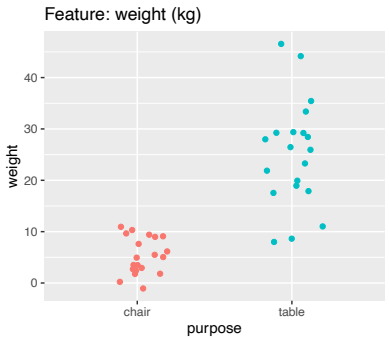
Here we used Euclidean distance, which will work for the vast majority of (normally distributed) expression data.

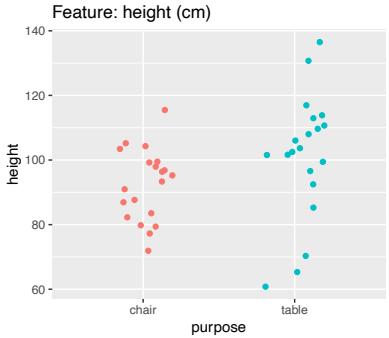
For data following a negative binomial distribution, Rao's distance is typically used.

For ranked data, the Kendall Tau distance is typically used.

Feature selection







Steps in clustering

- Selecting appropriate algorithm (there are many algorithms out there)
- Selecting appropriate distance metric (depending on the data distribution)
- Feature selection (optional)