Cluster Evaluation

[View all machine learning examples](https://www.mathworks.com/solutions/machine-learning/examples.html)

Clustering is used on unlabeled data to find natural groupings and patterns. Most clustering algorithms need the researcher to have prior knowledge of the number of clusters. When this information is not available, one can use cluster evaluation techniques to determine the number of clusters present in the data based on a specified metric. This example identifies clusters present in Fisher’s iris data.

Load Data

Fisher's iris data consists of measurements on the sepal length, sepal width, petal length, and petal width for 150 iris specimens.

clear

load fisheriris

X = meas;

y = categorical(species);

Evaluate Multiple Clusters from 1 to 10 to Find the Optimal Cluster

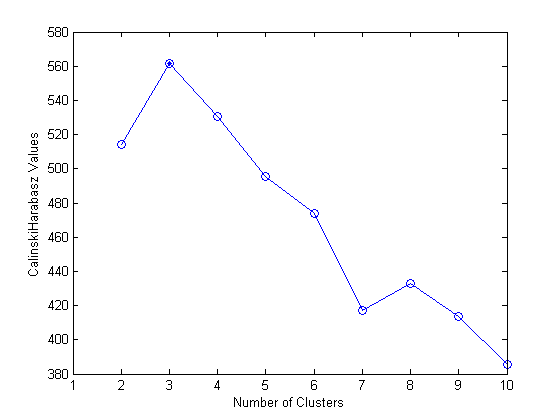
eva = evalclusters(X,'kmeans','CalinskiHarabasz','KList',[1:10]);

plot(eva)

disp(categories(y)')

Warning: Empty cluster created at iteration 1 during replicate 1.

'setosa' 'versicolor' 'virginica'



We can confirm the evaluation results since we know in advance that there are three species and, therefore, three clusters: *setosa, versicolor and virginica*

Dimensionality Reduction for Visualization

You may use principal component analysis to reduce the dimension of your data for visualization purposes. In this example, we will explore nonnegative matrix factorization, which (besides providing a reduction in the number of features) also guarantees that the features are nonnegative if your predictors are themselves nonnegative.

% Since none of our features are negative, lets use nnmf to confirm the 3

% clusters visually

Xred = nnmf(X,2);

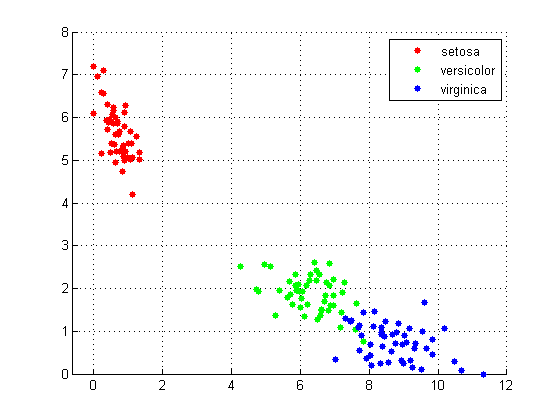
gscatter(Xred(:,1),Xred(:,2),y)

xlabel('Column 1')

ylabel('Column 2')

legend(categories(y))

grid on



Datasets and References

Fisher's iris data consists of measurements on the sepal length, sepal width, petal length, and petal width for 150 iris specimens. There are 50 specimens from each of three species. This dataset is shipped with [Statistics and Machine Learning Toolbox™](https://www.mathworks.com/products/statistics.html) .