Lesson 26: Cluster interpretation

Once we have inferred the clusters, we would like to know what are the distinguishing features. For the zoo data set, we could, for instance, mark two clusters, and then ask for the features that distinguish among these. Having data marked with cluster identifiers takes us back to classification, and we can use any of visualization, model inference, or feature ranking techniques we have introduced there. Here, we will show how to use ranking to infer what features characterize the group of mammals when compared to a close cluster of other species.
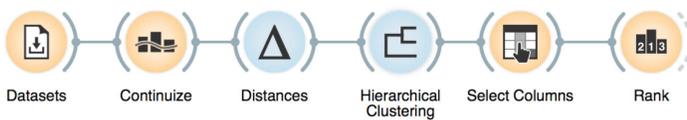
We load zoo data from Datasets, continuize the categorical features (this will be changed in Orange soon, Distances should automatically perform continuization), estimate the distances, and feed everything in Hierarchical Clustering. So far, nothing new.

In the hierarchical clustering, we choose two clusters. Note that Hierarchical Clustering adds cluster identifier as a meta feature; to make the data ready for classification-specific tasks, we need to promote cluster identifier into target variable (a class) by reassigning the feature types in Select Columns.
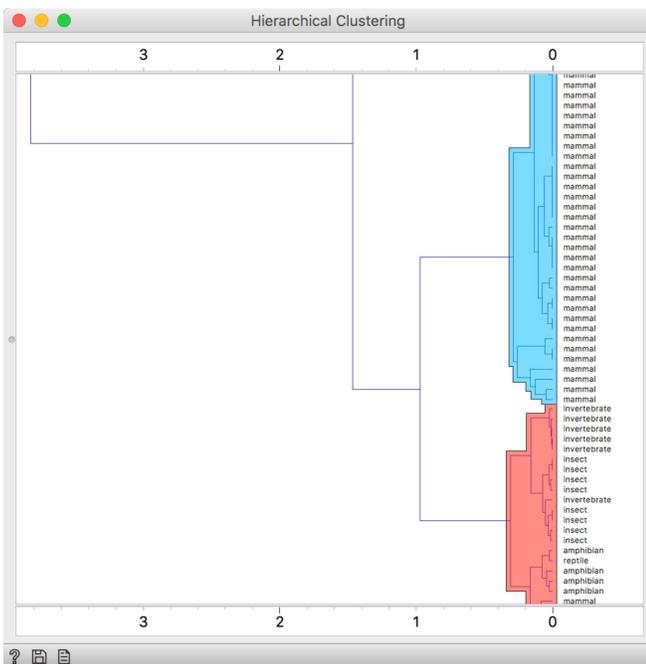
Use modifier keys (command) to select different branches of the dendrogram and mark them as separate clusters. Done correctly, Hierarchical Clustering will mark the branches with different colors.
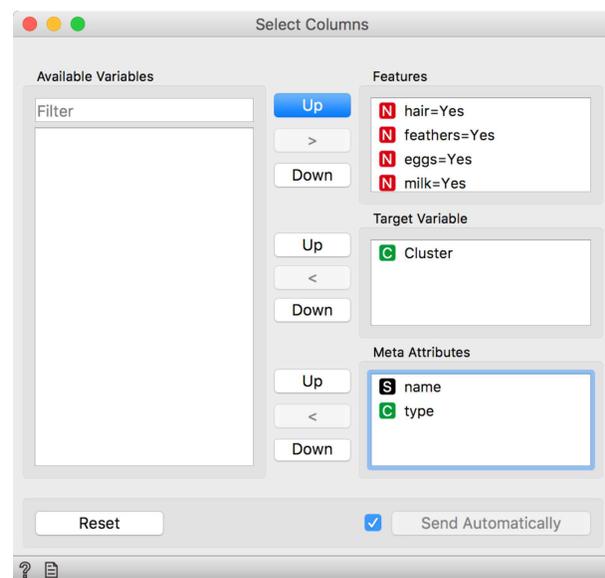
The data is now ready for classification-based analysis. Here, we used a rank widget.
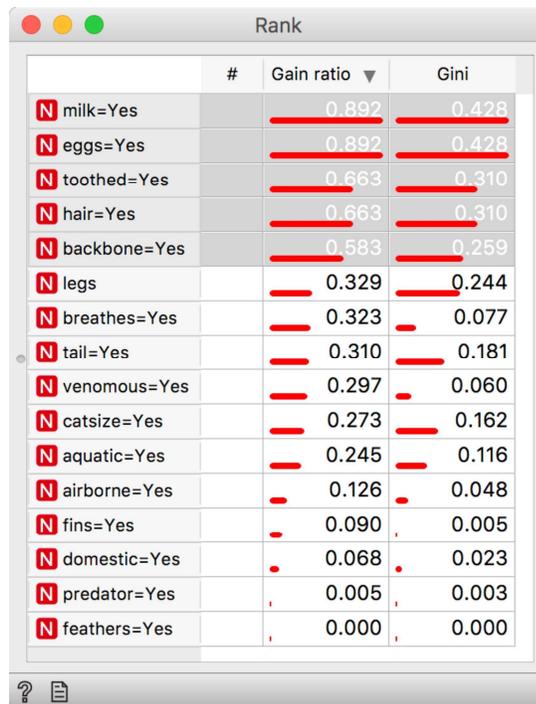
	#	Gain ratio ▼	Gini
N milk=Yes		0.892	0.428
N eggs=Yes		0.892	0.428
N toothed=Yes		0.663	0.310
N hair=Yes		0.663	0.310
N backbone=Yes		0.583	0.259
N legs		0.329	0.244
N breathes=Yes		0.323	0.077
N tail=Yes		0.310	0.181
N venomous=Yes		0.297	0.060
N catsize=Yes		0.273	0.162
N aquatic=Yes		0.245	0.116
N airborne=Yes		0.126	0.048
N fins=Yes		0.090	0.005
N domestic=Yes		0.068	0.023
N predator=Yes		0.005	0.003
N feathers=Yes		0.000	0.000