

Profiling Clusters & Scoring New Observations

Cluster Profiling

- *Profiling* can be defined as the generation of descriptions of the derived clusters from the input variables.
- These descriptions are the class label for a cluster.
- At least two (related) varieties of profiling exist:
 - Comparing the derived cluster means.
 - Comparing the derived means against a class.

Comparing the derived cluster means

1. Generate a variable indicating to which cluster each observation belongs.
2. Call PROC MEANS to generate the centroids that define each cluster.
3. Perform ANOVA or MANOVA on the input variables to test for significant differences between the cluster centroids.

MANOVA

- If the multivariate tests are found to be significant, it is still not known which variables serve as the basis for the difference.
- If the multivariate tests are found not to be significant, the groups may still significantly differ on the individual variables that define them.

Compare Known to Derived Clusters

Another way to profile clusters is to determine whether the clusters obtained match the clusters you expected to find.

1. Assign each observation to a class of interest.
2. Cluster the observations and construct a confusion matrix of the derived groups.
3. Use a chi-square test to determine whether the derived clusters and the given class are associated.
4. Determine which input variables differ significantly between clusters.

Scoring New Observations

- Use Cluster Analysis to generate the cluster centroids.
- Score new observations against the established cluster definitions, that is, centroids.

Scoring with the FASTCLUS Procedure

- **Step 1:** Perform cluster analysis and save the centroids.

```
PROC FASTCLUS ... OUTSTAT=<centroids>;
```

- **Step 2:** Load the saved centroids and score a new file.

```
PROC FASTCLUS ... INSTAT=<centroids> OUT=<scored>;
```

PROC FASTCLUS uses the entered centroids to calculate the distances between the new observations and the given centroids, and assigns each observation to the closest (that is, most similar) centroid.

Scoring with the CLUSTER Procedure

- **Step 1:** Generate the clusters and output to PROC TREE.

```
PROC CLUSTER METHOD=<method> ...  
OUTTREE=<tree>;
```

- **Step 2:** Generate cluster assignments.

```
PROC TREE DATA=<tree>  
OUT=<treeout> N=<nclusters>;
```

The TREE procedure assigns a cluster number to each observation.

continued...

Scoring with the CLUSTER Procedure

- **Step 3:** Sort the observations by cluster number and calculate centroids.

```
PROC SORT DATA =<treeout>;  
PROC MEANS DATA=<treeout>;  
  OUTPUT OUT=<centroids>;
```

- **Step 4:** Load the saved centroids and score a new file.

```
PROC FASTCLUS DATA=<newdata>  
  MAXCLUSTERS=<nclusters>  
  SEED=<centroids>  
  MAXITER=0 OUT=<scored>;
```