3.5 Semi-supervised clustering

3.5.1 Clustering with Constraints

The constraints can be enforced during the clustering:

> *Each time an observation is (re-)assigned, the constraints is verified and the (re-)assignment halted in case violations occur.*

3.5.1.1 Types of Constraints #Costraints

Three types of contraints:

- **Observation-level constraints:** Set for individual observations;
- **Cluster-level constraints:** Defined at the level of the cluster:
  1. Minimum separation or δ constraint;
  2. ε-constraint.
- **Negative background information:** find a clustering which is different from a given clustering;
- **Other constraints:** Requirement to have balanced clusters, whereby each cluster contains the same amount of observations.

OBSERVATION-LEVEL CONSTRAINTS #OBSERVATIONLEVELCONSTRAINTS

If the fraud behavior of only a few observations is known, they can then be forced into the same cluster. Two types of observation-level constraints can be identified:

- **Must-link constraint:** enforces that two observations should be assigned to the same cluster;
- **Cannot-link constraint:** enforces that two observations should be assigned to different clusters.

CLUSTER-LEVEL CONSTRAINTS #CLUSTERLEVELCOSTRAINTS

Can be subdivided in:

- **Minimum separation or δ constraint:** specifies that thedistance between any pair of observations in two different clusters must be at least δ;
- **ε-constraint:** specifies that each observation in a cluster with more than one observation must have another observation within a distance of at most ε.

3.5.1.2 One-Class SVM #OneClassSVM

> *The goal of **SVMs** (**S**upport **V**ector **M**achines) is to filter out outliers in the dataset, in order to have less skewed clusters.*

This is done by dividing the data with an hyperplane. The point that lie between the origin and the hyperplane (outliers) are discarded, while al the other data is kept.

Outliers: observations that lie below the hyperplane, closest to the origin. Normal observations lie above the hyperplane

⚠ Outliers will return a positive value for: $f(x) = sign(w^\tau \cdot \varphi(x) \cdot \rho)$.

HYPERPLANE #HYPERPLANE

One-class SVMs aim at solving the following optimization (hyperplane):

$$
\begin{aligned}
&min : [\tfrac{1}{2} \textstyle\sum_{i=1}^{N} w_i - \phi + \tfrac{1}{u \cdot n} \textstyle\sum_{i=1}^{n} e_i] \\
&subject\ to : \ w^\tau \cdot \varphi(t) \geq \rho - e_k,\ k = 1 \ldots n\ , \\
&\qquad\qquad\quad e_k \geq 0\ .
\end{aligned}
$$

This distance is maximized by minimizing the first part in the objective function, while the second part of the objective function accounts for the errors and/or the outliers.
The constraints force the majority of observations to lie above the hyperplane.

3.5.2 Evaluating and Interpreting Clustering Solutions #ClusteringEvaluation

🔑 Evaluating a clustering solution is not a trivial task because no universal criterion exists.

Statistical Perspective

One evaluation metric could be the **S**um of **S**quared **E**rrors (**SSE**):

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} dist^2(x, m_i)\ .$$

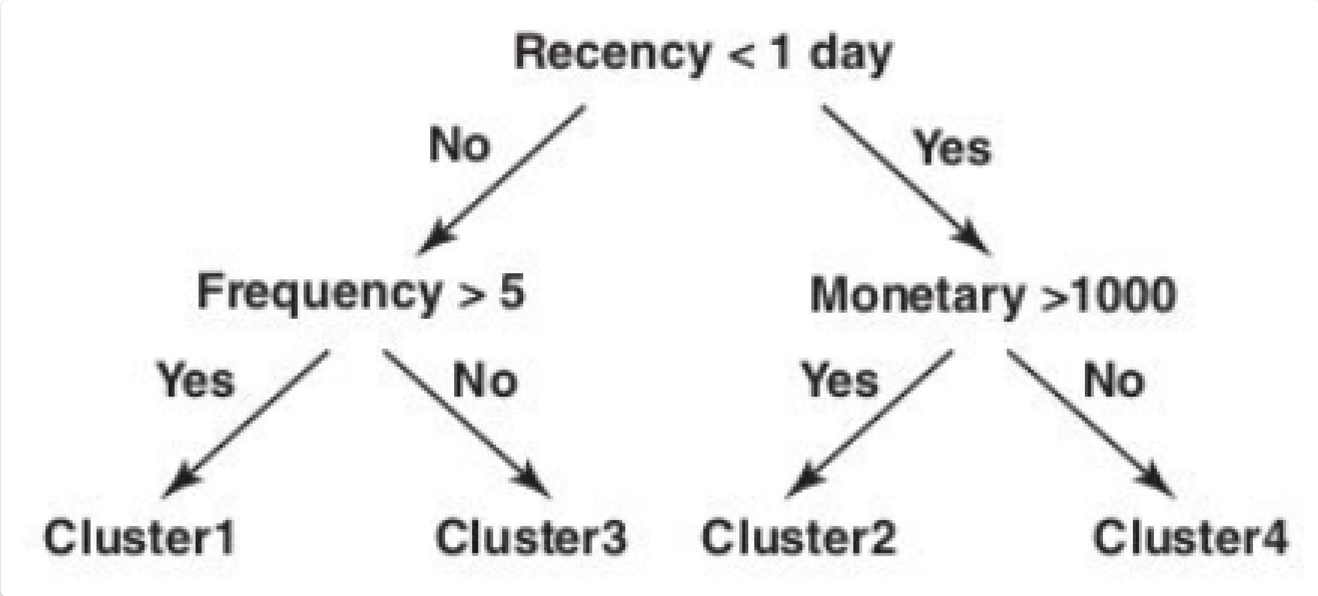
where K represents the number of clusters and m_i the centroid (e.g., mean) of cluster i.
==> When comparing two clustering solutions, the one with the lowest SSE can then be chosen.

Graphical Evaluation

Explore data and graphically compare cluster distributions with population distributions across all variables on a cluster-by-cluster basis.

Decision Trees

Given a clustering solution, build a decision tree with the ClusterID as the target variable.



White-Box Solutions

White-box supervised or predictive techniques can be used to explain the solution from a black-box descriptive analytics exercise.

Next chapter: [Predictive Analytics for Fraud Detection]