Глоссарий к Теме 9
Кластерный анализ

Кластерный анализ – предназначен для разбиения исходных данных на поддающиеся интерпретации группы, таким образом, чтобы элементы, входящие в одну группу были максимально «схожи», а элементы из разных групп были максимально «отличными» друг от друга.

Евклидово расстояние – это наименьшее расстояние между x и y. В двух- или трёхмерном случае — это прямая, соединяющая данные точки.

Расстояние Чебышева (Chebychev) – вычисление расстояния как максимума абсолютного значения разности между элементами.

Расстояние Минковского (Minkowski) – равно корню r-ой степени из суммы абсолютных разностей пар значений взятых в r-ой степени.

Межгрупповая связь (Between-groups linkage) – дистанция между кластерами, которая является средним значением всех расстояний между всеми возможными парами точек из обоих кластеров.

Внутригрупповая связь (Within-groups linkage) – дистанция между двумя кластерами рассчитывается на основании всех возможных пар наблюдений, принадлежащих обоим кластерам, причём учитываются также и пары наблюдений, образующиеся внутри кластеров.

Ближайший сосед (Nearest neighbor) – дистанция между двумя кластерами определяется как расстояние между парой наблюдений, расположенных друг к другу ближе всего, причём каждое наблюдение берётся из своего кластера.

Самый дальний сосед (Furthest neighbor) – дистанция между двумя кластерами определяется как расстояние между самыми удалёнными друг от друга значениями наблюдений, причём каждое наблюдение берётся из своего кластера.

Центроидная кластеризация (Centroid clustering) – в обоих кластерах рассчитываются средние значения переменных относящихся к ним наблюдений. Затем расстояние между двумя кластерами рассчитывается как дистанция между двумя осредненными наблюдениями.

Медианная кластеризация (Median clustering) – тот же центройдный метод, но центр объединенного кластера вычисляется как среднее всех объектов.

Метод Уорда (Ward-Method) – сначала в обоих кластерах для всех имеющихся наблюдений производится расчёт средних значений отдельных переменных. Затем вычисляются квадраты евклидовых расстояний от отдельных наблюдений каждого кластера до этого кластерного среднего значения. Эти дистанции суммируются. Потом в один новый кластер объединяются те кластеры, которые дают наименьший прирост общей суммы дистанций.

Центроиды – средние значения объектов, содержащихся в кластере, по каждой из переменных.