Исследование ошибки в моделях категоризации

Описание

Кластеризация - разбиение множества объектов на подмножества, называемые кластерами. Кластеризация, будучи математическим алгоритмом, имеет широкое применение во многих сферах: начиная с таких естественно научных областей как биология и физиология, и заканчивая маркетингом в социальных сетях и поисковой оптимизацией.

Основная идея *кластерного анализа* (clustering, cluster analysis) заключается в том, чтобы разбить объекты на группы или кластеры таким образом, чтобы внутри группы эти наблюдения были более похожи друг на друга, чем на объекты другого кластера. При этом мы заранее не знаем на какие кластеры необходимо разбить наши данные. Это связано с тем, что мы обучаем модель на неразмеченных данных (unlabeled data), то есть без целевой переменной, компонента y. Именно поэтому в данном случае говорят по машинное обучение без учителя (Unsupervised Learning). Существует множество алгоритмов кластеризации, и при использовании различных методов кластерного анализа для одной и той же совокупности могут быть получены различные варианты разбиения. Существенное влияние на характеристики кластерной структуры оказывают набор признаков, по которым осуществляется классификация и тип выбранного алгоритма.

Результат будет зависеть от того, что мы выберем в качестве "меры качества"или функционала. Наилучшим по выбранному функционалу следует считать такое разбиение, при котором достигается его экстремальное (минимальное или максимальное) значение.

Наиболее распространенными являются следующие функционалы:

- Сумма квадратов расстояний до центров классов.
  При использовании этого критерия стремятся получить такое разбиение совокупности объектов на k кластеров, при котором значение F минимально.

- Сумма внутриклассовых расстояний между объектами.
  В этом случае наилучшим следует считать такое разбиение, при котором достигается минимальное значение F. Объекты, попавшие в один кластер, близки между собой по значениям тех переменных, которые использовались для классификации.

- Суммарная внутриклассовая дисперсия.
  В данном случае разбиение, при котором сумма внутриклассовых (внутри групповых) дисперсий будет минимальной, следует считать оптимальным.

Задание

1. Используя датасет dataset1.cvs, проведите предобработку данных. Визуализируйте их.
2. Реализуйте алгоритм кластеризации методом ближайшего соседа, методом дальнего соседа.
3. Проведите разбиение на кластеры, выбрав в качестве метрики качества функционалы, перечиленные выше. Визуализируйте полученный результат.
4. **Определите тип распределения в каждом кластере.**