

Лабораторная работа №3

Вы используете исходные данные из лабораторной работы №1 и №2. Вариант также за вами сохраняется.

Вы решаете задачу кластеризации демографических данных по регионам.

Требования по отправке и оформлению лабораторной работы аналогичны требованию лабораторной работы №1 и №2.

Основные этапы выполнения работы:

1. Не забудьте удалить таргеты из предыдущих лабораторных работ из вашей выборки.
2. Нормирование (масштабирование) исходных данных. Обратите внимание, что данные (коэффициенты, числа) для нормализации (масштабирования) рассчитываются только на основе обучающей выборки. И затем уже применяются к тестовым данным.
3. С помощью библиотеки `sklearn` сделать `fit-predict` модели иерархической кластеризации. Произвести кластеризацию 3 раза – с каждым из типов связей, которые мы проходили на занятии (параметр `linkage`). Построить дендрограмму для каждого типа связи и определить оптимальное число кластеров по ней. Выберите наилучший вариант (по вашему мнению) и обоснуйте ваш выбор. Получите итоговые метки кластера для каждого объекта на основе наилучшего варианта и определенного вами по дендрограмме наилучшего числа кластеров.
4. С помощью библиотеки `sklearn` сделать `fit-predict` модели `k-средних`. Перебрать по сетке различные варианты числа кластеров. Для каждого посчитать метрику Дэвиса-Болдина. Определить оптимальное число кластеров на основе значений этой метрики (выбрать наилучший вариант кластеризации).
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.davies_bouldin_score.html
5. Посчитайте индекс Рэнда между наилучшей кластеризацией из п.3 и наилучшей кластеризацией из п. 4. Сделать вывод о близости выбранных вами вариантов на основе этого индекса.
https://scikit-learn.org/stable/modules/generated/sklearn.metrics.rand_score.html#sklearn.metrics.rand_score
6. Для одного из наилучших вариантов для каждого кластера посчитать среднее значение признаков в каждом кластере. Проинтерпретировать кластеры на основе различий между средними значениями признаков в различных кластерах (постараться дать «логичные» названия).