

Домашнее задание по теме «Иерархическая кластеризация»

Формулировка задания

Выполните подготовку данных для решения задачи кластеризации. Проведите кластеризацию алгоритмом AgglomerativeClustering. Обратите внимание на подбор гиперпараметров `n_clusters`, `metric`, `linkage`. Проведите кластеризацию методом DBSCAN. Обратите внимание на подбор параметров `eps`, `min_samples`.

Для кластеризации использовать библиотеку `scikit-learn`.

Результирующий код должен быть читаемым, с единой системой отступов и адекватными названиями переменных.

Описание плана работы

- 1) Загрузите данные из дополнительных материалов или по ссылке:
<https://www.kaggle.com/datasets/youssefaboelwafa/clustering-penguins-species>
- 2) EDA(Exploratory Data Analysis) можно использовать из задачи Снижения размерности. Обратите внимание на выбросы и требования к нормализации данных.
- 3) Примените метод кластеризации AgglomerativeClustering с подбором гиперпараметров в цикле.
- 4) Постройте дендрограмму по кластерам. Сделайте вывод.
- 5) Сохраните модель в файл `joblib`.
- 6) Подготовьте данные для кластеризации с целевым признаком `sex` или пол пингвина. Примените метод кластеризации AgglomerativeClustering с подбором гиперпараметров. Оцените качество кластеризации методом `adjusted_rand_score` для 2 кластеров из метрик библиотеки `scikit-learn`. Сделайте вывод.
- 7) Попробуйте построить для исходных данных модель DBSCAN с подбором гиперпараметров в цикле.
- 8) Оцените качество кластеризации методом силуэта или любым другим. Сделайте вывод.

Перечень необходимых инструментов

- Python
- scikit-learn
- pandas
- venv
- Jupiter Notebook
- IDE VS Code
- GigaIDE

Форма предоставления результата

1. В поле ссылки загрузить ссылку на удаленный репозиторий с доступом для наставника.
2. В поле файла загрузить архив с папкой, в которой разместить отчет со скриншотами по заданию и решение задачи. Решение должно быть представлено в формате .ipynb или .py.

Шкала оценивания

- 1.0 – отлично
- 0.7–0.9 – хорошо
- 0.5–0.6 – удовлетворительно
- Менее 0.5 – задание не выполнено