Лабораторная №5. Выбор признаков

https://classroom.github.com/a/ACOGqPUf

Набор данных

- Выберите набор данных для задачи классификации текста.
- Желательно использовать набор данных, который вы получили в первой лабораторной работе, если он содержал полноценный текстовый признак.
- Если не удастся найти подходящий набор данных, можно взять набор данных SMS или castle-or-lock.
- Выберите целевую функцию ошибки или качества для задачи классификации, а также соответствующий способ валидации алгоритма классификации.

Алгоритмы

- Реализуйте 3 метода выбора признаков: встроенный, обёртку и фильтрующий.
- Выберите библиотечную реализацию 3-х методов выбора признаков: встроенного, обёртки и фильтрующего. При этом конкретные варианты методов должны отличаться от ваших реализаций.
- Реализуйте любой алгоритм кластеризации.

Задание

- Векторизуйте набор данных при помощи CountVectorizer или аналогов.
- Выведите 30 наиболее значимых признаков (слов) различными методами выбора признаков. Сравните полученные списки.
- Определите, как меняется качество работы различных (не менее трёх) классификаторов до и после выбора признаков каждым из методов. Выберите один метод выбора признаков.
- Кластеризуйте данные до и после выбора признаков. Оцените качество кластеризации любой внешней и внутренней мерой.
- Методами PCA и tSNE уменьшите размерность данных до и после выбора признаков. Визуализируйте данные и отметьте реальные классы, а также как их кластеризовал алгоритм кластеризации из предыдущего пункта.
-