Лабораторная работа № 7

Тема: Метод главных компонент (PCA).

1. Для работы используйте данные для классификации с большим количеством параметров (не менее 30).
2. Обучите модель случайного леса (Random Forest). Рассчитайте точность.
3. Сократите количество параметров датасета одним из способов Feature Selection (из модуля sklearn.feature_selection), например, параметры с низкой дисперсией.
4. На сокращенном датасете обучите модель случайного леса (Random Forest). Рассчитайте точность.
5. К исходному (большому) датасету примените метод PCA, найдите 2 главные компоненты.
6. Визуализируйте данные по этим двум компонентам.
7. Обучите модель случайного леса на полученной модели PCA с двумя компонентами. Оцените точность и время.
8. Из графика зависимости отклонения модели от количества главных компонент (см. пример по ссылке https://habr.com/ru/companies/ods/articles/325654/) найдите такое количество главных компонент, чтобы оставить 90% дисперсии исходных данных..
9. Рассчитайте модель с определенным в п.6 количеством компонент. Оцените точность и время.

Вопросы:

1. Что означает уменьшение размерности в машинном обучении?
2. Какие методы входят в Feature Selection? Как они работают?
3. Расскажите принцип работы метода PCA.
4. Что означает понятие главная компонента?
5. Что означает термин Ансамбли в контексте машинного обучения?
6. Как работают алгоритмы стекинг, бэггинг, бустинг?
7. Объясните метод Random Forest. Какой алгоритм асамблирования в нем используется?