Лабораторная работа № 3

Бинарная классификация. Библиотека sklearn

Пример выполнения приведен в файле *DT_KNN.html*

1. Выберите любой доступный датасет подходящий для бинарной классификаци (т.е. в нем присутствует вектор меток, который имеет два значения)

Хорошо подходят датасеты по различным болезням, где надо определить есть болезнь/нет болезни или выдача кредитов – выдавать/не выдавать, или отток клиентов мобильных операторов – уйдут/не уйдут.

В папке с примерами выложены 2 датасета, которыми тоже можно воспользоваться: citrus.csv и diabetes.csv.

В файле citrus.csv вектором меток является столбец name со значениями orange/ grapefruit.

В diabetes.csv вектором меток является столбец Outcome со значениями 0/1.

2. Проанализируйте исходные данные, при необходимости заполните пропуски или удалить не важную информацию. Категориальные признаки замените на числовые

3. Выделите из данных вектор меток У и матрицу признаков Х.

4. Разделите набор данных на обучающую и тестовую выборки (*train_test_split()*)

5. На обучающей выборке обучите **модель дерева решений** *DecisionTreeClassifier()*.

6. **Оцените точность** модели, рассчитайте матрицу ошибок (*confusion matrix*) и метрики качества (*Accuracy, Precision, Recall*).

7. **Улучшите модель** путем подбора наилучших гиперпараметров модели при помощи функции *GridSearchCV()* (для дерева решений *max_depth* и *max_features*).

8. **Оцените точность** модели после улучшения, сравните с точностью до улучшения.

9. Проделайте п.5-8 для методов **k-ближайших соседей и случайного леса.**

10. Сделать выводы о лучшей модели для данного датасета.

11*. Визуализируйте полученную модель дерева решений (при визуализации желательно уменьшить глубину дерева, что бы рисунок был читаемым, или сохранить в отдельный файл)

Вопросы:

1. Сформулируйте задачу классификации?
2. Что означает обучение с учителем?
3. Зачем разделять обучающую выборку?
4. Что означает переобученная модель? Как с этим бороться?
5. Что означает обобщающая способность моделей машинного обучения?
6. Объясните значения в матрице ошибок, как она рассчитывается?
7. Что показывают *accuracy, precision и recall*?
8. Что означает понятие ансамбль в машинном обучении?
9. Расскажите о методе случайного леса.