РАБОТА С МОДЕЛЬЮ КЛАССИФИКАЦИИ

СЛУЧАЙНЫЙ ЛЕС

Задание основано на применении пакета Python Scikit-learn для расширения области знаний по работе с логическим методом, таким как случайный лес (Random Forests).

Цели практики:

- овладеть навыками работы со случайным лесом
- научиться решать задачи регрессии с помощью случайного леса
- освоить навык подбора параметров для случайного леса

Начало работы

В библиотеке scikit-learn случайные леса реализованы в классах **sklearn.ensemble.RandomForestClassifier** (для классификации) и **sklearn.ensemble.RandomForestRegressor** (для регрессии).

Обучение модели производится с помощью функции **fit**, построение прогнозов – с помощью функции **predict**. Число деревьев задается с помощью поля класса **n_estimators.**

Пример:

```
import numpy as np
from sklearn.ensemble import RandomForestRegressor
X = np.array([[1, 2], [3, 4], [5, 6]])
y = np.array([-3, 1, 10])
clf = RandomForestRegressor(n_estimators=100)
clf.fit(X, y)
predictions = clf.predict(X)
```

В рамках задания необходимо вычислить качество предсказаний на тестовой выборке. Для этого применяется метрика **R2** – по сути, это среднеквадратичная ошибка **(MSE**), нормированная на отрезок **[0, 1]** и обращенная так, чтобы ее наилучшим значением была единица. Ее можно вычислить с помощью функции **sklearn.metrics.r2_score**.

Первым аргументов является список правильных ответов на выборке, вторым – список предсказанных ответов.

Пример:

```
from sklearn.metrics import r2_score
print r2_score([10, 11, 12], [9, 11, 12.1])
```

Задание

ВАЖНО: при решении заданий необходимо следить за изменением качества случайного леса в зависимости от количества деревьев в нем.

1. Загрузите данные из файла **abalone.csv** – датасет, в котором требуется предсказать возраст ракушки (число колец) по физическим измерениям.

2. Преобразуйте признак **Sex** в числовой: значение **F** должно перейти в ***-1;*** **I** – ***в 0,*** **M** – ***в 1***. При применении Pandas, можно использовать конструкцию иже:

data['Sex'] = data['Sex'].map(lambda x: 1 if x == 'M' else (-1 if x == 'F' else 0))

3. Разделите содержимое файлов на признаки и целевую переменную.
В последнем столбце записана целевая переменная, в остальных — признаки.

4. Обучите случайный лес (**sklearn.ensemble.RandomForestRegressor**) с различным числом деревьев: **от 1 до 50 (random_state=1).**
Для каждого из вариантов оцените качество работы полученного леса на кросс-валидации по 5 блокам. Используйте параметры **"random_state=1"**и "**shuffle=True"** при создании генератора кросс-валидации **sklearn.cross_validation.KFold.** (при появлении затруднений, обратитесь к предыдущим практическим задачам).
В качестве меры качества воспользуйтесь коэффициентом детерминации (sklearn.metrics.r2_score).

5. Определите, при каком минимальном количестве деревьев случайный лес показывает качество на кросс-валидации **выше 0.52.** Это количество и будет ответом на задание.

6. Обратите внимание на изменение качества по мере роста числа деревьев. Ухудшается ли оно?

Ответ на каждое задание — текстовый файл, содержащий ответ в первой строчке (#.txt). Обратите внимание, что отправляемые файлы не должны содержать перевод строки в конце.

ТЕОРЕТИЧЕСКИЕ ОСНОВЫ

Случайный лес представляет собой модель классификации, которая объединяет несколько решающих деревьев в одну композицию, улучшая тем самым качество их работы и обобщающую способность.

Каждое дерево строится независимо от остальных, при этом обучение происходит на случайном подмножестве обучающей выборки для обеспечения их различности. Для уменьшения схожести деревьев выбирается оптимальный признак для разбиения не из всех доступных, а лишь из случайного подмножества признаков.

Прогнозы, полученные отдельными деревьями, объединяются путем усреднения. Особенность случайного леса заключается в том, что он не поддается переобучению при увеличении количества деревьев в композиции. Это достигается благодаря независимости деревьев друг от друга: добавление нового дерева не усложняет модель, а лишь снижает уровень шума в прогнозах, поскольку деревья не взаимодействуют между
https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html