Решающие деревья и случайные леса

Практические задания для самостоятельного выполнения

Задание 1.

1. Используя модуль ***datasets*** библиотеки ***Scikit-learn*** сгенерировать модельный набор данных для задачи многоклассовой классификации с двумя информативными признаками и числом классов, равным трем. Обеспечить воспроизводимость результатов.
2. Выполнить визуализацию сгенерированных облаков точек (точки, отображающие объекты разных классов, должны быть выведены разными цветами).
3. Выполнить разбиение набора данных, полученного в п. 1, на обучающую и тестовую выборки в соотношении 70/30.
4. Создать модель решающего дерева с параметрами по умолчанию (обеспечить только воспроизводимость результатов) и обучить ее на обучающей выборке.
5. Получить предсказания обученной модели для объектов тестовой выборки. Оценить качество классификации с помощью метрики ***accuracy*** как на обучающей, так и на тестовой выборке; дать интерпретацию полученных оценок.
6. Изучить зависимость обобщающей способности модели от глубины дерева. Для этого выполнить следующие действия.
   6.1. Создать модель решающего дерева глубиной 1 и обучить ее на обучающей выборке. Получить предсказанные моделью метки классов для объектов обучающей и тестовой выборок. Оценить качество классификации с помощью метрики ***accuracy*** на обучающей и тестовой выборках.

   Выполнить визуализацию полученных результатов: изобразить на графиках разделяющие линии классов (залить области, относимые моделью к каждому из трех классов, разными цветами); на этих же графиках отобразить облака точек, представляющие объекты выборки. Выполнить такое построение отдельно для обучающей и тестовой выборки. В заголовках графиков вывести информацию о том, какая именно выборка визуализирована (обучающая или тестовая) и оценку качества классификации на этой выборке.
   6.2. Создать модель решающего дерева глубиной 2 и выполнить с ней действия, описанные в п. 6.1.
   6.3. Создать модель решающего дерева глубиной 3 и выполнить с ней действия, описанные в п. 6.1.
7. Проанализировать все результаты, полученные в п. 5 и 6. Сделать выводы. Создать отчет: описание и оценка качества каждой модели, выводы по результатам исследования.
8. Создать модель решающего дерева с ограничением на число объектов в листе (задать не более 3 объектов). Обучить эту модель на обучающей выборке. Получить предсказанные моделью метки классов для объектов

обучающей и тестовой выборок. Оценить качество классификации с помощью метрики ***accuracy*** на обучающей и тестовой выборках. Сопоставить результаты с полученными в п. 5 и 6, сделать выводы и добавить их в отчет.

Задание 2.

Рассматривается (в упрощенном варианте) задача, размещенная на **kaggle**: https://www.kaggle.com/c/bioresponse. Задача состоит в том, чтобы по данным характеристикам молекулы определить, будет ли дан биологический ответ (biological response). Исходные данные: https://www.kaggle.com/c/bioresponse/data. Для анализа следует использовать данные из файла *train.csv*. Каждая строка описывает одну молекулу.

1. Импортировать данные из файла *train.csv* в объект *Pandas DataFrame* и вывести несколько первых записей (для контроля корректности импорта и получения представления о наборе данных). Вывести также размерность полученного датафрейма.
2. Значения целевого признака (наличие/отсутствие биологического ответа) находятся в столбце *Activity*. Отделить эти значения от остальных данных, сохранив их в отдельном объекте.
3. Определить соотношение классов в имеющемся наборе данных. Для этого вывести доли записей, относящихся к каждому из классов, в общем количестве записей. Сделать вывод о сбалансированности выборки.
4. Проанализировать возможности модели с деревьями небольшой глубины.
    а. Создать случайный лес с 20 деревьями, каждое из которых имеет глубину не более 2. При настройке параметров модели включить использование подхода ***out-of-bag***. Обучить полученную модель и вывести ***OOB*** – оценку качества обученного алгоритма.
    б. Создать аналогичные случайные леса с 50, 10, 150 и 200 деревьями. Обучить эти модели и для каждой из них получить оценку ***OOB***.
    в. Построить график зависимости оценки качества случайного леса от количества базовых алгоритмов. Сделать выводы.
5. Проанализировать возможности моделей с деревьями большей глубины: выполнить все действия, перечисленные в п. 4, для случайных лесов с деревьями глубины 10.
6. Сопоставить результаты, полученные при выполнении пп. 4 и 5. Сделать выводы о предсказательной способности рассмотренных моделей.

Задание 3 (дополнительно).

Продолжение анализа ситуации, описанной в задании 5 по линейной регрессии (в файле 4_2_Линейная регрессия.docx).

1. Выполнить (если это еще не сделано) пп. 1 – 7 задания 5 по линейной регрессии.
2. Для получения прогнозов вместо модели линейной регрессии использовать случайный лес с числом деревьев, равным 50, и глубиной каждого дерева, равной 20. Организовать раздельную обработку бинарных, числовых и

категориальных признаков:

- бинарные признаки не нужно преобразовывать (только отделить);
- числовые признаки – отделить и отмасштабировать (метод ***StandardScaler*** из модуля ***preprocessing***);
- категориальные признаки – отделить и применить бинарное кодирование (результат – булева матрица);
- после обработки столбцы собрать вместе, причем после сборки порядок столбцов должен сохраняться (для этого можно использовать трансформер ***FeatureUnion***);
- для преобразованного набора данных создать модель «случайный леес».

Указание: можно использовать ***Pipeline***, созданный при выполнении задания 5, заменив в нем линейную регрессию на случайный лес.

3. Обучить случайный лес на обучающих данных. Оценить качество полученной модели на тестовых данных с помощью метрики *MAE*. Сравнить результат с результатом, полученным при выполнении задания 5.
4. Для получения наглядного представления об адекватности модели выполнить визуализацию: вывести облако точек в координатах «истинные значения целевой функции» – «предсказанные моделью значения». Проанализировать полученные результаты, сделать выводы.

   Указание: ясно, что при хорошей предсказательной способности модели облако точек должно располагаться вдоль биссектрисы первого координатного угла (прогнозируемые моделью результаты близки к истинным значениям целевой функции).

   Сделать выводы.