СЕССИЯ 1

Исходные файлы:

1) test_ses.csv (Тестовый набор данных)
2) train_ses.csv (Обучающий набор данных)
3) site.pkl (Словарь сайтов)
4) Машинное обучение – C1.pdf (Инструкция к 1 сессии)

Результаты работы:

1) Data.zip (Предобработанные данные)
2) Report_C1.html, Report_C1.ipynb (Отчет о проделанной работе)
3) Readme.txt (Дополнительные комментарии)

ВВЕДЕНИЕ

На этом чемпионате вам предстоит решить задачу определения злоумышленника по его поведению в сети Интернет. По последовательности из десяти веб-сайтов, посещенных подряд одним и тем же человеком, мы будем идентифицировать этого человека. Идея такая: пользователи Интернета по-разному переходят по ссылкам, и это может помогать их идентифицировать (кто-то сначала в почту, потом про футбол почитать, затем новости, социальная сеть, потом, наконец, – работать, кто-то – сразу работать, если это возможно).

Набор данных содержит информацию о сеансах просмотра пользователями, в которых:

- site_i - это идентификаторы сайтов в этом сеансе (в соответствии с словарем site.pkl);
- time_j - это отметки времени посещения соответствующего сайта;
- target - принадлежит ли эта сессия злоумышленнику;

На этой сессии необходимо выполнить подготовку данных к анализу и построению моделей.

ЗАДАНИЕ

1.1 Подготовка обучающей и тестовой выборок

Необходимо выполнить подготовку данных для дальнейшего описательного анализа и построения прогнозных моделей. Следует выполнить загрузку и преобразование всех необходимых данных. Данные необходимо очистить и привести к приемлемому формату

1.2 Работа с разреженным форматом данных

Сформировать мешок сайтов. То есть необходимо создать новые матрицы, в которых строкам будут соответствовать сессии из 10 сайтов, а столбцам – индексы сайтов. На пересечении строки и столбца будет стоять число – количество раз, которое встретился сайт в сессии номер N.

Сериализуйте полученные матрицы для дальнейшего применения и возможного улучшения модели.

1.3 Подготовка отчета

Подготовьте отчет о проделанной работе по итогам сессии, в котором будут представлены результаты, выводы и обоснования выбора по каждому разделу задания. Результаты работы должны состоять из отчетов в формате .html и исходников с возможностью перекомпиляции. Архив Data.zip должен содержать все результаты выполнения модуля, а также все необходимые файлы для запуска и проверки участков кода. В файле Readme.txt необходимо описать содержимое результирующих файлов архива Data.zip.