ВМК МГУ

Задание для поднаучных Кравцовой О.А.

Основы методов математического прогнозирования, Кравцова Ольга Анатольевна, 2024

1 Характеристики задания

- **Тема задания:** машинное обучение, оптимизация моделей.
- **Телеграм автора задания:** @oakravts

2 Описание задания

Цель задания – проверить несколько методов машинного обучения в произвольной задаче прогнозирования.

Для подготовки к выполнению задания выполняющий должен выполнить следующие пункты:

- Ознакомиться с задачей машинного обучения. Можно воспользоваться вот этой статьёй на хабре, но лучше посмотреть вот эту лекцию Воронцова
- Ознакомиться с библиотекой scikit-learn и с основными методами работы с ней. Для начала можно воспользоваться данной инструкцией.
- Выбрать задачу, которую вы хотите решить: регрессия или классификация. Для выбранной задачи найти данные, с которыми вы будете работать. Разрешается пользоваться данными, которые уже есть в sklearn и готовы к использованию.
- Выбрать как минимум два метода из sklearn, которые поддерживают различные функции потерь. Для того, чтобы убедиться в этом, следует зайти на страницу с описанием метода (пример), и убедиться, что loss является одним из параметров. Обычно такими методами являются различные SGD* и GradientBoosting*.
- Выбрать как минимум два метода из sklearn, которые НЕ поддерживают различные функции потерь.

3 Решение задания

Для решения задания выполняющий должен выполнить следующие пункты

1. Подготовить данные для работы с ними: составить матрицу признаков (по которым делаются предсказания) и предсказываемые значения. В случае использования готового датасета этот шаг можно пропустить.

2. Разбить данные на обучающую часть и тестируемую. Можно воспользоваться данной функцией.

3. Выбрать метрику качества для вашей задачи. К примеру, для задачи классификации можно воспользоваться точностью, а в случае регрессии средней абсолютной ошибкой.

4. Взять методы, которые поддерживают различные функции потерь. Обучить данные методы с несколькими вариантами функции потерь (минимум 2 разные). Оценить качество методов на тестовом множестве с использованием метрики, выбранной в пункте 3. Проанализировать результаты и сделать по ним выводы.

5. Взять методы, которые НЕ поддерживают различные функции потерь. Воспользоваться методом перебора параметров по сетке. Сетку нужно составить с использованием как минимум 2 различных параметров метода (выбираются на выбор выполняющего). В сетку перебова в качестве scoring (метода оценки качества) следует взять функцию, которая не совпадает с целевой метрокой из 3го пункта, но совпадает с одной из функций потерь, использованной в методах из предыдущего пункта. Разрешается использовать готовые метрики.

6. Сделайте выводы по каждому из пунктов:

- Удалось ли Вам подобрать хорошую метрику качества? На чем основано ваше суждение?

- Какая функция потерь в 4 пункте дала вам лучшее значение целевой метрики. Как думаете, почему?
- Удалось ли вам выбрать хорошую метрику, используемую как scoring? Помогла ли она решить вашу задачу?

4 Критерии оценивания

Задание будет оцениваться по следующим критериям:

Понимание задачи. Требуется описать решаемую задачу: что у вас за выборка, какую задачу вы решаете, почему она возникла.

Выбор модели для решения задачи. Описание методов, которые вы используете для решения поставленной задачи.

Правильность и понятность кода. Будет оцениваться как верность выполнения задания, так и общее качество кода.

Визуализация результатов. Будет оцениваться как выполняющий смог представить полученный результат. Просто цифры не столь показательны, как, например, гистограмма, на которой собрано качество различных методов.