ВМК

Задание 3. Композиции алгоритмов для решения задачи регрессии

Практикум 317 группы, 2019

Начало выполнения задания: 8 декабря 2019 года.
Мягкий дедлайн: **21 декабря 2019 года, 23:59.**

Формулировка задания

Данное задание направлено на ознакомление с алгоритмами композиций.

В задании необходимо:

1. Написать на языке Python собственную реализацию методов случайных лес и градиентный бустинг. Прототипы функций должны строго соответствовать прототипам, описанным в спецификации и проходить все выданные тесты. Задание, не проходящее все выданные тесты, приравнивается к невыполненному. При написании необходимо пользоваться стандартными средствами языка Python, библиотеками numpy, scipy и matplotlib. Библиотекой scikit-learn пользоваться запрещается, если это не обговорено отдельно в пункте задания.
2. Провести описанные ниже эксперименты с выданными данными.
3. Написать отчёт о проделанной работе (формат PDF). Отчёт должен быть подготовлен в системе LaTeX.

Список экспериментов

Эксперименты этого задания необходимо проводить на датасете данных о продажах недвижимости. Данные можно скачать по ссылке.

Эксперименты

1. Проведите минимальную обработку имеющихся данных. Разделите данные на обучение и контроль, переведите данные в numpy ndarray.
2. Исследуйте поведение алгоритма случайный лес. Изучите зависимость RMSE на отложенной выборке и время работы алгоритма в зависимости от следующих факторов:
   - количество деревьев
   - размерность подвыборки признаков для одного дерева
   - максимальная глубина дерева (+случай, когда глубина неограничена)
3. Исследуйте поведение алгоритма градиентный бустинг. Изучите зависимость RMSE на отложенной выборке и время работы алгоритма в зависимости от следующих факторов:
   - количество деревьев
   - размерность подвыборки признаков для одного дерева
   - максимальная глубина дерева (+случай, когда глубина неограничена)
   - выбранный learning_rate
     (каждый новый алгоритм добавляется в композицию с коэффициентом γ * learning_rate)

Требования к реализации

Прототипы всех функций описаны в файлах, прилагающихся к заданию.

Среди предоставленных файлов должны быть следующие модули и функции в них:

1. Модуль `ensembles.py` с реализациями случайного леса и градиентного бустинга. Алгоритмы должны соответствовать классическим реализациям, разобранным на лекции.

   Для одномерной оптимизации используйте функцию minimize_scalar. Разрешается использовать класс DecisionTreeRegressor из библиотеки scikit-learn.