Инструкция по рецензированию

Файлы решения

KNN.ipynb (/solution_file/16557/KNN.ipynb)

Рецензии

1.2. Чем отличаются поверхности, полученные при числе соседей 1 и 10? Объясните, чем вызваны данные отличия.

2.0 балла Показано и дано объяснение, почему большее число соседей в среднем дает более гладкие границы	Рецензия №1 Рецензия №2 Рецензия №3
1.0 балл Отмечено, что при 10 соседях граница более гладкая. Доказательства и рассуждения не приведены	---
0.0 баллов Нет верного ответа	---

1.3. Объясните, почему наблюдается сильное отклонение разделяющей поверхности от прямой x=0 при значениях y<−10 и y>10. Дайте строгое математическое обоснование наблюдаемого явления

3.0 балла Доказано, что при разном масштабе признаков различна их степень влияния на ответ алгоритма. Например, рассмотрена Евклидова метрика и показано, что если масштаб одного признака больше в K раз масштаба другого, то этот признак будет в K^2 раз иметь больший вес при классификации	Рецензия №3
1.5 балла Даны рассуждения про масштаб признаков, но нет четкого доказательства	---
0.0 баллов Нет верного ответа	Рецензия №1 Рецензия №2

1.4. Начертите разделяющие поверхности для KNN, обученного на нормализованных признаках. Используйте функцию $plot\_knn\_bound$

1.0 балл Граница начерчена для обоих нормализаторов	Рецензия №1 Рецензия №2 Рецензия №3
0.0 баллов Хотя бы одна граница не начерчена	---

1.5. Пусть дано произвольное число точек на плоскости. Представим, что каждая точка - это отдельный класс. Пусть на данной выборке был обучен $1-NN$ классификатор. Чем с геометрической точки зрения являются разделяющие поверхности этого классификатора? Требуется математически строго, однозначно и с полным обоснованием определить геометрическое место точек разделяющих поверхностей.

4.0 балла Для произвольного числа точек доказано, что разделяющими поверхностями будут являться пересечения срединных перпендикуляров между каждой парой точек. Можно начать доказательство с пары точек, для которых утверждение очевидно, и далее обобщить по индукции для произвольного числа точек	Рецензия №3
3.0 балла Доказательство приведено, но содержит незначительные неточности	---
1.0 балл Сказано про срединные перпендикуляры и/или Диаграмму Вороного и/или Триангуляцию Делоне, но не приведено строгое/корректное доказательство	Рецензия №2
0.0 баллов Нет верного ответа	Рецензия №1

2.2. Разбейте выборку на обучающую (75%) и тестовую (25%) с помощью функции *train_test_split*. Используйте параметр random_state=42! Не забудьте перемешать данные перед разбиением (см. параметры функции). Запустите кросс-валидацию на 3 фолдах с помощью реализованных вами функций *kfold_split*, *knn_cv_score*. В качестве метрики используйте *r2_score*.

1.0 балл Все пункты задания выполнены верно	Рецензия №1 Рецензия №2 Рецензия №3
0.0 баллов Задание выполнено не полностью и/или с ошибками	---

2.3. Какой наибольший *r2_score* удалось достичь на валидации? Какие закономерности вы видите? Обучите модель с наилучшими параметрами на всей обучающей выборке, измерьте *r2_score* на тестовой выборке.

3.0 балла 1) Получена метрика вблизи значения 0.7. 2) Отмечено, что нормализация значительно увеличивает качество. 3) Модель обучена на всей обучающей выборке и оценена на тестовой	Рецензия №1 Рецензия №2 Рецензия №3
2.0 балла Один из пунктов не выполнен или выполнен неверно	---
1.0 балл Два пункта не выполнены или выполнены неверно	---

0.0 баллов
Три пункта не выполнены или выполнены неверно

3.1. Найдите оптимальные параметры обучения модели. Осуществлять перебор параметров следует по заданной ниже сетке. Используйте реализованные вами функции *kfold_split*, *knn_cv_score*. В качестве метрики используйте *accuracy_score*

2.0 балла
Найдены оптимальные параметры. Качество на кросс-валидации порядка 0.76-0.78

Рецензия №1
Рецензия №2
Рецензия №3

0.0 баллов
Иначе

3.2. Какой метод предобработки данных в среднем дает наилучший результат? Почему?

2.0 балла
TfidfVectorizer дает лучшее качество в среднем. Если токен встречается в большом числе документов, то он менее информативен для определения класса конкретного документа и, наоборот. TF_IDF учитывает это, CountVectorizer - нет.

Рецензия №1
Рецензия №2
Рецензия №3

1.0 балл
Показано, что TfidfVectorizer дает лучшее качество в среднем, но не сказано почему.

0.0 баллов
Иначе

3.3. Какая метрика близости позволяет в среднем достичь наилучшее качество? Почему?

2.0 балла
Косинусная близость в среднем лучше. Приведены рассуждения о том, что сонаправленность векторов встречаемости токенов важнее чем разность их величин

Рецензия №1
Рецензия №2
Рецензия №3

1.0 балл
Показано, что косинусная близость в среднем лучше, но не сказано почему.

0.0 баллов
Иначе

3.4. Начертите график зависимости метрики качества от числа соседей. Метрику следует усреднить по всем параметрам, кроме числа соседей. Сделайте выводы о наблюдаемых зависимостях

2.0 балла Начерчен график. С ростом числа соседей метрика в среднем падает. Если обратить внимание на названия классов-тем, то можно заметить, что есть довольно близкие по смыслу темы. Вероятно, большее число соседей добавляет в целевую переменную шум от близких темах, что приводит к падению качества	Рецензия №3
1.0 балл Начерчен график. С ростом числа соседей метрика в среднем падает. Не приведено объяснение	Рецензия №1 Рецензия №2
0.0 баллов Иначе	---

3.5. Оцените точность вашей лучшей модели на тестовой части датасета. Отличается ли оно от качества, полученного на кросс-валидации? Почему?

3.0 балла 1) Финальная модель обучена на всей обучающей выборке, качество на тесте в районе 0.67 или выше, что значительно меньше качества на валидации. 2) Показано, что распределение целевых переменных обучающей и тестовой выборок совпадают. 3) Сделан вывод о том, что проблема в разном распределении токенов обучающей и тестовой выборок	Рецензия №2 Рецензия №3
2.0 балла В ответе присутствуют два пункта из трех, либо приведены рассуждения про переобучение и выполнен пункт 1	Рецензия №1
1.0 балл В ответе присутствует один пункт из трех, либо приведены рассуждения про переобучение	---
0.0 баллов Иначе	---

Оцените, насколько понятным для вас был код автора

0.0 баллов Все понятно	Рецензия №1 Рецензия №2 Рецензия №3
0.0 баллов В целом понятно, были неясные места	---
0.0 баллов Есть, куда расти	---

Оцените, насколько понятным для вас было оформление ноутбука автора в целом

0.0 баллов Все понятно	Рецензия №1 Рецензия №2 Рецензия №3
0.0 баллов В целом понятно, были неясные места	---
0.0 баллов Есть, куда расти	---

Комментарии

Рецензия №1

Рецензия №2

Рецензия №3

Топ