12 колонок

11 (первые) – исходные данные

1 (последний) – выходные данные

1599 значений

# Информация по столбцам

1. Фиксированная кислотность (4,6-15,9)
2. Летучая кислотность (0,12-1,58)
3. Лимонная кислота (0-1)
4. Остаточный сахар (0,9-15,5)
5. Хлориды (0,01-0,61)
6. Свободный диоксид серы (1-72)
7. Диоксид серы общий (6-289)
8. Плотность (0,99-1)
9. PH (2,74-4,01)
10. Сульфаты (0,33-2)
11. Алкоголь (8,4-14,9)
12. Качество (оценка от 0 до 10) (3-8)

# Общая информация по датасету

Контекст

Эти два набора данных относятся к красному и белому вариантам португальского вина "Винью Верде". Для получения дополнительной информации см. Ссылку [Cortez et al., 2009]. Из-за проблем с конфиденциальностью и логистики доступны только физико-химические (исходные) и сенсорные (выходные) переменные (например, нет данных о сортах винограда, марке вина, цене продажи вина и т. Д.).

Эти наборы данных можно рассматривать как задачи классификации или регрессии. Классы упорядочены и не сбалансированы (например, нормальных вин гораздо больше, чем отличных или плохих).

Этот набор данных также доступен в репозитории машинного обучения UCI, https://archive.ics.uci.edu/ml/datasets/wine+quality, я просто поделился им с kaggle для удобства. (Если я ошибаюсь и тип общедоступной лицензии запрещает мне это делать, я сниму это, если потребуется.)

Содержание

Для получения дополнительной информации прочтите [Cortez et al., 2009].

Входные переменные (на основе физико-химических тестов):

* 1 - фиксированная кислотность
* 2 - летучая кислотность
* 3 - лимонная кислота
* 4 - остаточный сахар
* 5 - хлориды
* 6 - свободный диоксид серы
* 7 - диоксид серы общий
* 8 - плотность
* 9 - pH
* 10 - сульфаты
* 11 - алкоголь

Выходная переменная (на основе сенсорных данных):

* 12 - качество (оценка от 0 до 10)

Подсказки

Что может быть интересным, помимо использования регрессионного моделирования, это установить произвольное ограничение для вашей зависимой переменной (качество вина), например, 7 или выше классифицируются как «хорошо / 1», а остальные - как «не хорошо / 0».

Это позволяет попрактиковаться в настройке гиперпараметров, например, алгоритмы дерева решений, смотрящие на кривую ROC и значение AUC.

Без какой-либо разработки функций или переоборудования вы сможете получить AUC 0,88 (даже без использования алгоритма случайного леса)

KNIME - отличный инструмент (GUI), который можно использовать для этого.

* 1 - File Reader (для csv) для узла линейной корреляции и интерактивной гистограммы для базового EDA.
* 2- File Reader для 'узла механизма правил', чтобы превратить 10-балльную шкалу в переменную dichtome (хорошее вино и отдых), код для вставки в механизм правил выглядит примерно так:
* $ quality $> 6.5 => "хорошо". ИСТИНА => "плохо"
* 3- Выход узла механизма правил на вход узла фильтра столбцов, чтобы отфильтровать исходную 10-балльную функцию (это предотвращает утечку)
* 4- Вывод узла фильтра столбцов на вход узла разделения (стандартное разделение поездов / тестов, например, 75% / 25%, выберите «случайный» или «стратифицированный»)
* 5- Разделение данных узла разделения выходных данных на входные данные Разделение данных на входные данные Узла обучения дерева решений и
* 6- Разделение тестовых данных узла разделения выходных данных на входной узел предиктора дерева решений
* 7- Выход узла обучающего дерева решений для входа узла дерева решений
* 8- Вывод дерева решений на вход ROC Node .. (здесь вы можете оценить свою модель на основе значения AUC)

Вдохновение

Используйте машинное обучение, чтобы определить, какие физико-химические свойства делают вино «хорошим»!

Благодарности

Этот набор данных также доступен в репозитории машинного обучения UCI, https://archive.ics.uci.edu/ml/datasets/wine+quality, я просто поделился им с kaggle для удобства. (Я ошибаюсь, и тип общедоступной лицензии запрещает мне это делать, я удалю это по первому запросу. Я не являюсь владельцем этого набора данных.

Пожалуйста, включите эту цитату, если вы планируете использовать эту базу данных: П. Кортез, А. Сердейра, Ф. Алмейда, Т. Матос и Дж. Рейс. Моделирование предпочтений вин путем извлечения данных из физико-химических свойств. В системах поддержки принятия решений, Elsevier, 47 (4): 547-553, 2009.

Соответствующая публикация

П. Кортез, А. Кердейра, Ф. Алмейда, Т. Матос и Дж. Рейс. Моделирование предпочтений вин путем извлечения данных из физико-химических свойств.

В системах поддержки принятия решений, Elsevier, 47 (4): 547-553, 2009.