

Сравнение KNN и SVM на плоских данных

Филютович Дмитрий

27 декабря 2021 г.

1 Вступление

Я хочу сравнить алгоритмы классификации KNN и SVM по нескольким параметрам: частоту ошибок, время обучения и время предсказания. Я провел сравнение на различных тестовых данных.

2 Немного о тестовых данных

Для создания данных я использовал библиотеку `sklearn.datasets`. Среди данных есть несколько "клякс "луны" с разным уровнем шума, круги с разным уровнем шума. Также я нашел реальный датасет о сердечных заболеваниях.

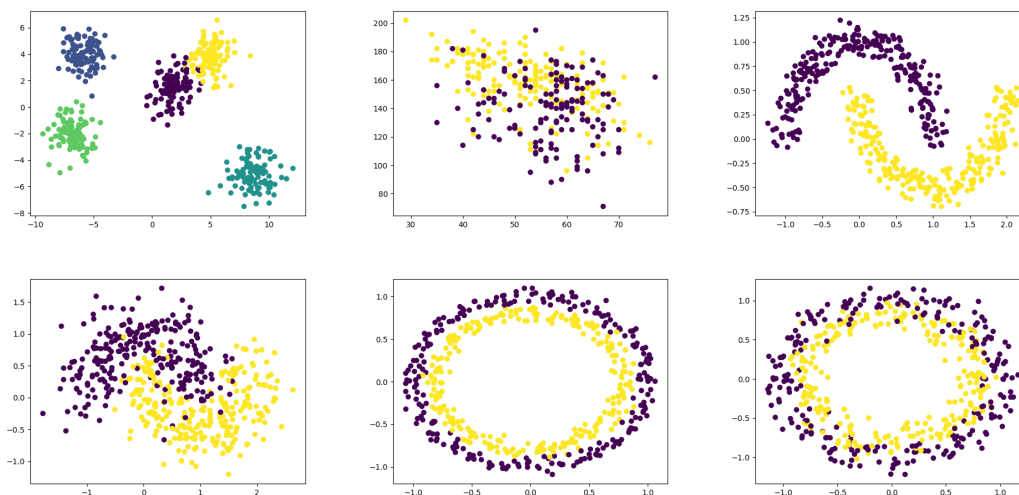


Рис. 1: Используемые тестовые данные

3 Исследование

3.1 Кляксы

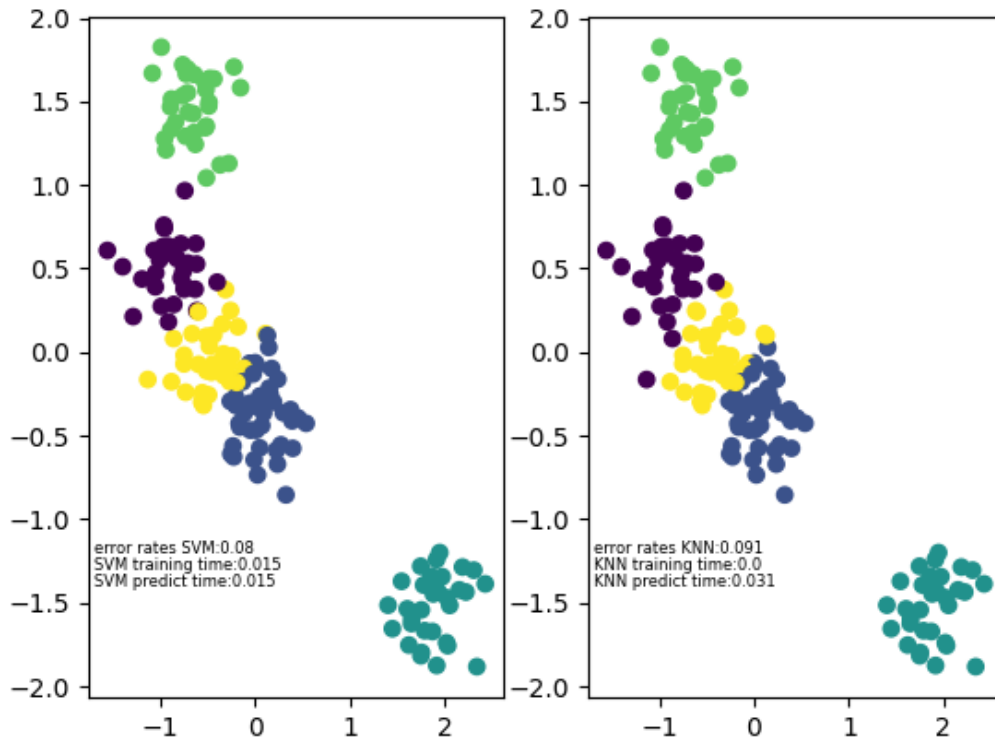


Рис. 2: Работа алгоритмов на кляксах

Как мы видим частота ошибок у обоих алгоритмов почти одинаковая. Здесь стоит заметить, что KNN почти не тратит время на обучения в отличие от SVM. А вот предсказывает быстрее алгоритм SVM.

3.2 Луны

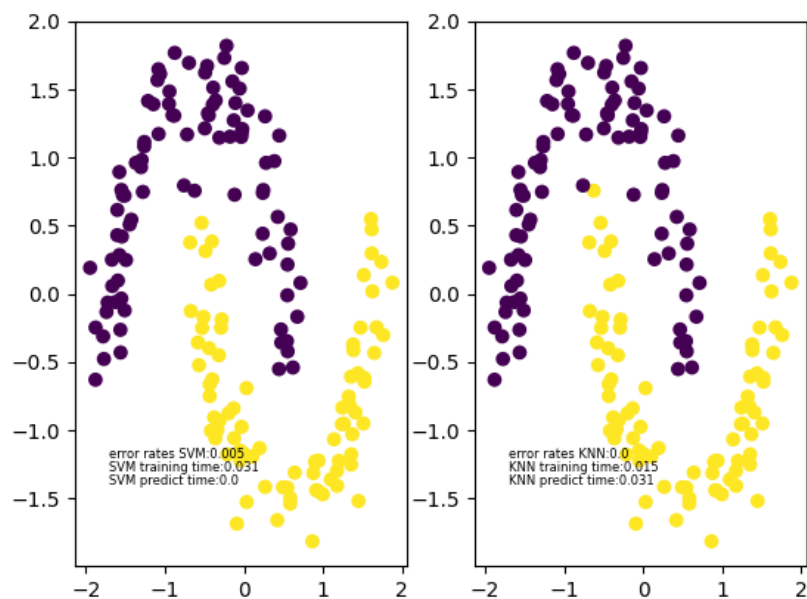


Рис. 3: Работа алгоритмов на лунах с малым шумом

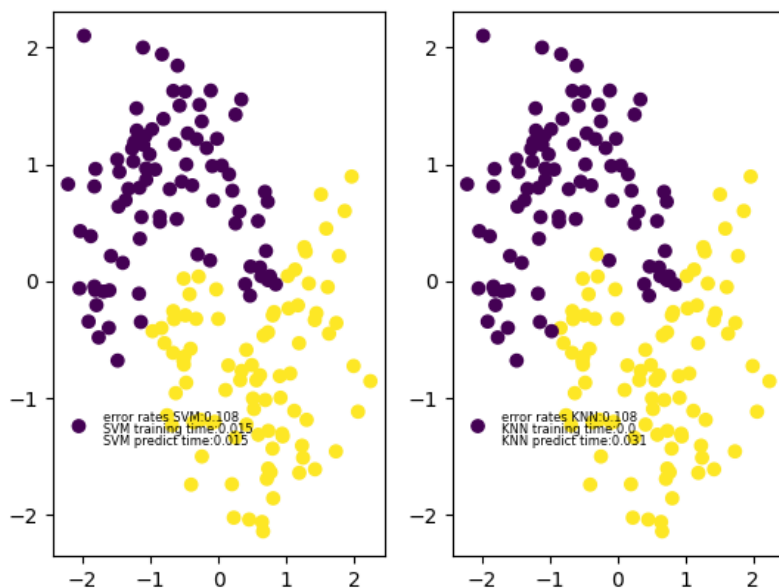


Рис. 4: Работа алгоритмов на лунах с большим шумом

На лунах с малым шумом оба алгоритма дают практически нулевую частоту ошибок. Также SVM обучается дольше, чем KNN, но предсказывает быстрее, чем KNN. На данных с большим шумом

результаты по времени практически не изменились, но частота ошибок значительно увеличилась у обоих алгоритмов.

3.3 Круги

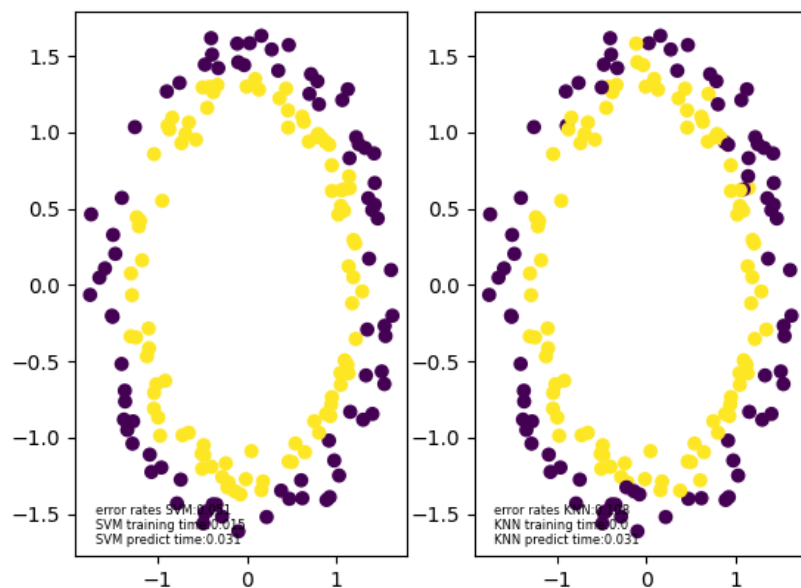


Рис. 5: Работа алгоритмов на кругах с малым шумом

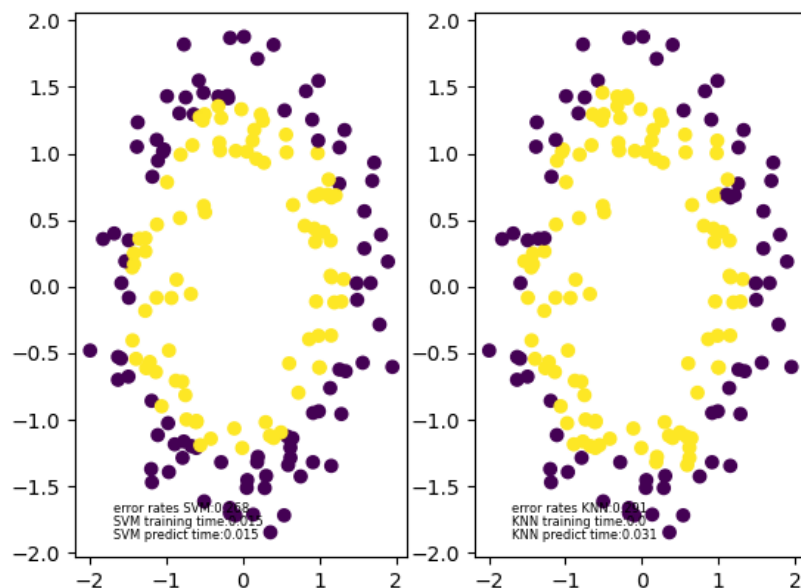


Рис. 6: Работа алгоритмов на кругах с большим шумом

На кругах с малым шумом частота ошибок SVM меньше, чем у KNN. Но на кругах SVM предсказывает также быстро, как и KNN. Если добавлять больше шума, то частота ошибок значительно возрастет. У обоих алгоритмов она в районе 30 процентов. Это не очень хороший результат. По времени соотношение остается тоже самое.

3.4 Реальный датасет

Этот датасет имеет два аргумента: возраст и максимально достигнутая частота сердцебиения.

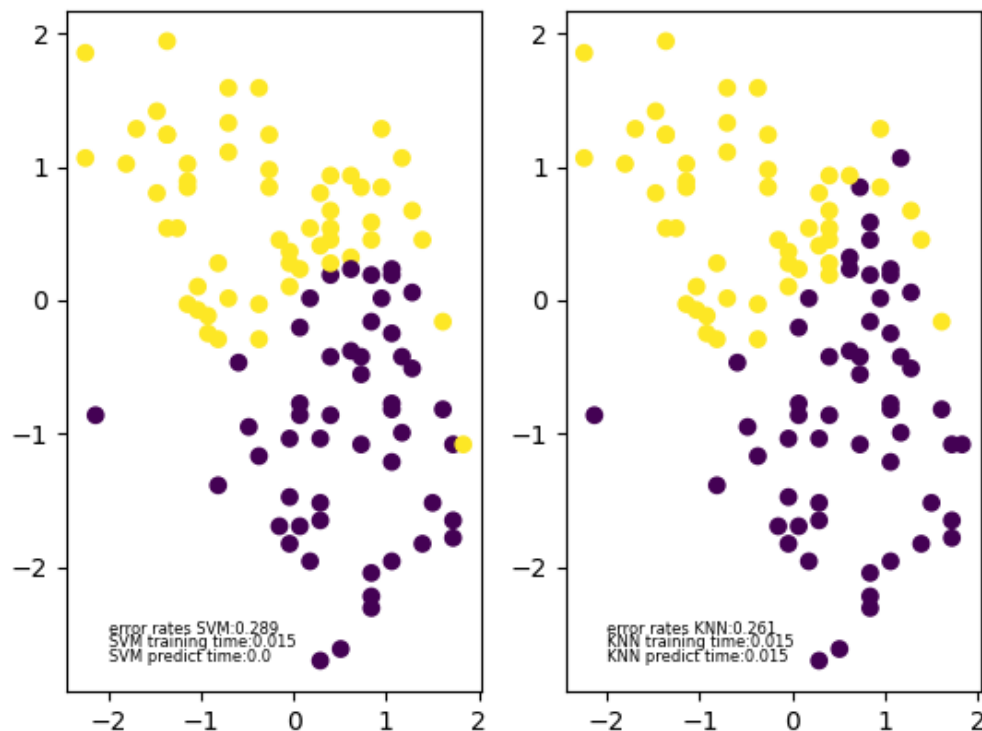


Рис. 7: Работа алгоритмов на реальном датасете

На реальном датасете частота ошибок высока. Это значит, что либо взаимосвязь между возрастом, максимальной частотой сердцебиения и сердечными заболеваниями слабая, либо оба алгоритма плохо справляются с этим датасетом.

3.5 Вывод

И SVM и KNN в среднем показывают одинаковую точность, но стоит заметить, что KNN обучается намного быстрее, чем SVM, но предсказывает результат медленнее.