**Лабораторная работа 3**

Описание датасета

Данные содержат 830 записей и 6 атрибутов. Датасет содержит информацию для скрининга рака молочной железы.

Attributes

BI-RADS assessment: 1 to 5

Age: patient's age in years

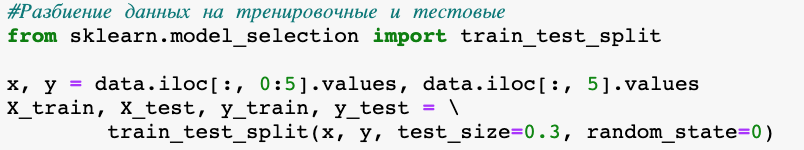
Shape: mass shape: round=1 oval=2 lobular=3 irregular=4

Margin: mass margin: circumscribed=1 microlobulated=2 obscured=3 ill-defined=4 spiculated=5

Density: mass density high=1 iso=2 low=3 fat-containing=4

Severity: benign=0 or malignant=1





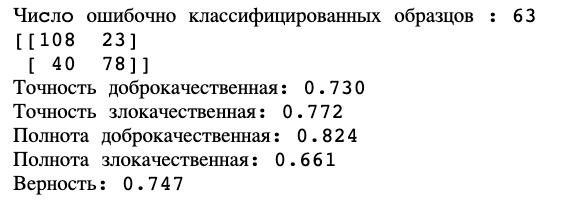
Матрица несоответствий - это просто квадратная таблица, которая сообщает о числе истинно положительных, истинно отрицательных, ложноположительных и ложноотрицательных. Хотя эти метрики можно легко вычислить вручную путем сравнения истинных и предсказанных меток классов, библиотека scikit-leaгn предлагает вспомогательную функцию confusion\_matrix предсказаний классификатора.

Точность ( P R E ) и полнота ( R E C ) - это метрики оценки качества, которые связаны с долями истинно положительных и истинно отрицательных исходов, при этом фактически полнота является синонимом доли истинно положительных исходов.

В диагностике опухоли, например, мы более заинтересованы в обнаружении злокачественных опухолей, для того чтобы помочь пациенту с соответствующим лечением. Однако также важно уменьшить число неправильно классифицированных как злокачественные (ложноположительные исходы) доброкачественных опухолей, с тем чтобы не беспокоить пациента без необходимости.

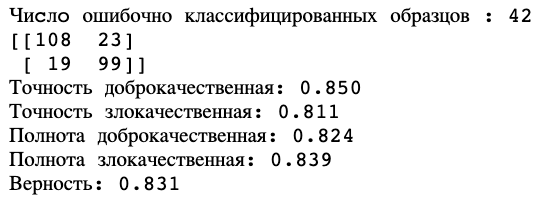
Персептрон

Используя объекты Perceptron, задавая темп обучения eta и число эпох n\_iter, т. е. проходов, по тренировочному набору. Если темп обучения слишком большой, то алгоритм промахнется по глобальному минимуму стоимости. Если темп обучения слишком малый, то для достижения сходимости алгоритм потребует большего количества эпох, что может замедлить обучение - в особенности для больших наборов данных. Кроме того, мы использовали параметр random\_state для воспроизводимости исходного перемешивания тренировочного набора данных после каждой эпохи.



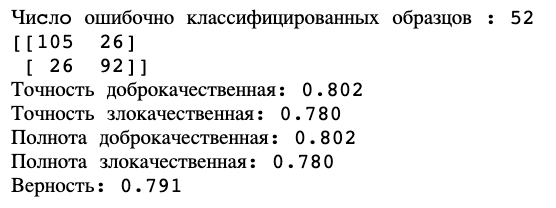
Дерево решений

Деревья решений могут создавать сложные границы решения путем деления пространства признаков на прямоугольники. Однако мы должны быть осторожными, поскольку чем глубже дерево решений, тем сложнее становится граница решения, что может легко закончиться переобучением. Теперь, используя scikit-learn, натренируем дерево решений с максимальной глубиной 3, применив в качестве критерия неоднородности энтропию. Хотя для целей визуализации может понадобиться шкалирование признаков, отметим, что масштабирование признаков не является необходимой составной частью алгоритмов деревьев решений.

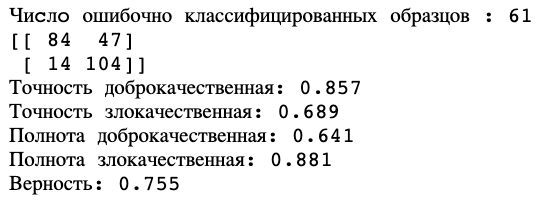


К-средних

Основное преимущество такого подхода с запоминанием состоит в том, что классификатор немедленно адаптируется по мере сбора новых тренировочных данных. Однако его оборотная сторона - вычислительная сложность классифицирования новых образцов - растет линейно вместе с числом образцов в тренировочном наборе данных в наихудшем случае, если только в наборе данных не очень много размерностей (признаков) и алгоритм не был реализован с использованием эффективных структур данных, таких как КD-деревья. Кроме того, мы не можем отбросить тренировочные образцы, поскольку никакого трестирующего шага нет. Вследствие этого, если мы работаем с большими наборами данных, пространство памяти может представлять серьезную проблему.



SVM



Результирующая средняя верность на перекрестно-проверочных данных дает хорошую оценку того, чего ожидать, если настроить гиперпараметры модели и затем использовать ее на ранее не встречавшихся данных. Например, мы можем применить подход с вложенной перекрестной проверкой для сравнения модели на основе SVM с классификатором на основе простого дерева решений; для простоты мы настроим лишь ее параметр глубины.

Вывод

На практике были изучены методы преобразования данных, а так же четыре алгоритма классификации

Мы узнали о разнообразных алгоритмах машинного обучения, которые используются для решения линейных и нелинейных задач. Мы убедились, что деревья решений особенно привлекательны, в случае если интерпретируемости модели уделено должное внимание.

Несмотря на то что методы опорных векторов являются мощными линейными моделями, которые могут быть распространены на нелинейные задачи методом ядерного трюка, для создания хороших прогнозов они требуют настройки большого числа параметров.

Классификатор на основе k ближайших соседей предлагает альтернативный подход к классификации - использовать ленивое обучение, которое позволяет делать прогнозы без какой-либо тренировки модели, но с более вычислительно затратным шагом прогнозирования.

Никакой алгоритм не сможет делать хороших прогнозов без информативных и отличительных признаков.