

Московский авиационный институт
(Национальный исследовательский университет)

Факультет: «Информационные технологии и прикладная математика»

Кафедра: 806 «Вычислительная математика и программирование»

Дисциплина: «Машинное обучение»

Лабораторная работа № 1, 2

Студент: Камеш Михаил

Группа: М80-307Б-18

Преподаватель: Ахмед Самир Халид

Москва, 2021

Постановка задания

1) Найти себе набор данных (датасет) для следующей лабораторной работы и проанализировать его. Выявить проблемы набора данных, устранить их. Визуализировать зависимости, показать распределение некоторых признаков. Реализовать алгоритмы К ближайших соседа с использованием весов и Наивный Байесовский классификатор и сравнить с реализацией библиотеки sklearn.

2) Необходимо реализовать алгоритмы машинного обучения. Применить данные алгоритмы на наборы данных, подготовленных в первой лабораторной работе. Провести анализ полученных моделей, вычислить метрики классификатора. Произвести тюнинг параметров в случае необходимости. Сравнить полученные результаты с моделями реализованными в scikit-learn. Аналогично построить метрики классификации. Показать, что полученные модели не переобучились. Также необходимо сделать выводы о применимости данных моделей к вашей задаче.

Описание алгоритмов

1. Алгоритм n-ближайших соседей (KNN)

Для классификации каждого из объектов тестовой выборки необходимо последовательно выполнить следующие операции:

А) Вычислить расстояние до каждого из объектов обучающей выборки

Б) Отобрать k объектов обучающей выборки, расстояние до которых минимально

С) Класс классифицируемого объекта — это класс, наиболее часто встречающийся среди k ближайших соседей

При взвешенном способе во внимание принимается не только количество попавших в область определённых классов, но и их удалённость от нового значения. Для каждого класса j определяется оценка близости:

$$Q_j = \sum_{i=1}^n \frac{1}{d(x, a_i)^2},$$

где $d(x, a_i)$ — расстояние от нового значения x до объекта a_i .

У какого класса выше значение близости, тот класс и присваивается новому объекту.

2. Гауссовский Наивный Байесовский классификатор

Основная идея — построить классификатор в предположении того, что функция $p(X_i, C_j)$ известна для каждого класса и равна плотности многомерного нормального (гауссовского) распределения:

$$p(x_i | c_j) = \frac{1}{\sqrt{2\pi\sigma_{i,j}^2}} e^{-\frac{1}{2}\left(\frac{x_i - \mu_{i,j}}{\sigma_{i,j}}\right)^2} \text{ for } i = 1, 2 \text{ and } j = 1, 2, 3$$

где μ - среднее значение, а σ - стандартное отклонение, которое мы должны оценить по данным. Это означает, что мы получаем одно среднее значение для каждого признака i в паре с классом c .

3. Логистическая регрессия

Логистическая регрессия применяется для прогнозирования вероятности возникновения некоторого события по значениям множества признаков. Для этого вводится так называемая *зависимая переменная* y , принимающая лишь одно из двух значений — как правило, это числа 0 (событие не произошло) и 1 (событие произошло), и множество *независимых переменных* (также называемых признаками, предикторами или регрессорами) — вещественных x , на основе значений которых требуется вычислить вероятность принятия того или иного значения зависимой переменной. Как и в случае линейной регрессии, для простоты записи вводится фиктивный признак $x_0 = 1$

Делается предположение о том, что вероятность наступления события $y = 1$ равна:

$$\mathbb{P}\{y = 1 \mid x\} = f(z),$$

где $z = \theta^T x = \theta_0 + \theta_1 x_1 + \dots + \theta_n x_n$, x и θ — векторы-столбцы значений независимых переменных и параметров (коэффициентов регрессии) — вещественных чисел θ , соответственно, а $f(z)$ — так называемая *логистическая функция* (иногда также называемая сигмоидом или логит-функцией):

$$f(z) = \frac{1}{1 + e^{-z}}.$$

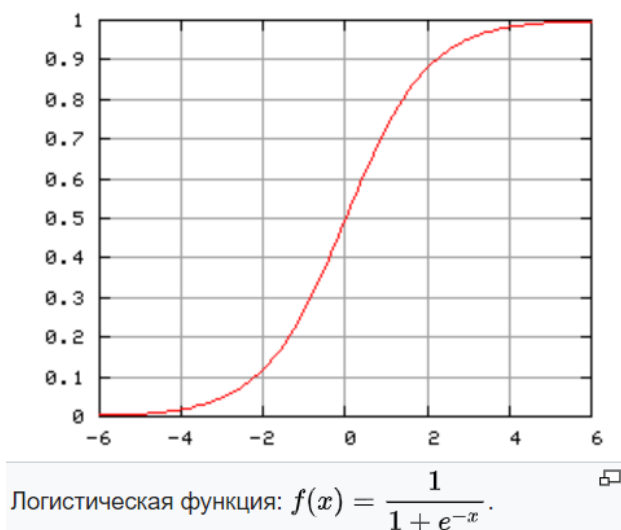
Так как y принимает лишь значения 0 и 1, то вероятность принять значение 0 равна:

$$\mathbb{P}\{y = 0 \mid x\} = 1 - f(z) = 1 - f(\theta^T x).$$

Для краткости функцию распределения при заданном x можно записать в таком виде:

$$\mathbb{P}\{y \mid x\} = f(\theta^T x)^y (1 - f(\theta^T x))^{1-y}, \quad y \in \{0, 1\}.$$

Фактически, это есть распределение Бернулли с параметром, равным $f(\theta^T x)$.



4. Дерево решений

Дерево решений — в основном жадное, нисходящее, рекурсивное разбиение. Энтропия — это мера случайности или неопределенности. Уровень энтропии колеблется от 0 до 1. Для меры энтропии используют примесь Джини. Узел чистый, если все его выборки принадлежат одному и тому же классу, в то время как узел с множеством выборок из разных классов будет иметь Джини ближе к 1.

$$G = 1 - \sum_{k=1}^n p_k^2$$

Каждый узел делит выборку таким образом, что примесь Джини у детей (точнее, среднее значение Джини у детей, взвешенных по их размеру) сводится к минимуму. Рекурсия останавливается, когда, достигается максимальная глубина, или когда нет разделения, которое может привести к двум детям, чище, чем их родитель.

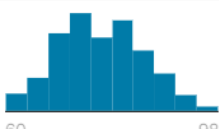
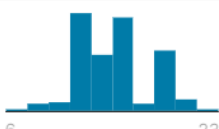

5. Случайный лес

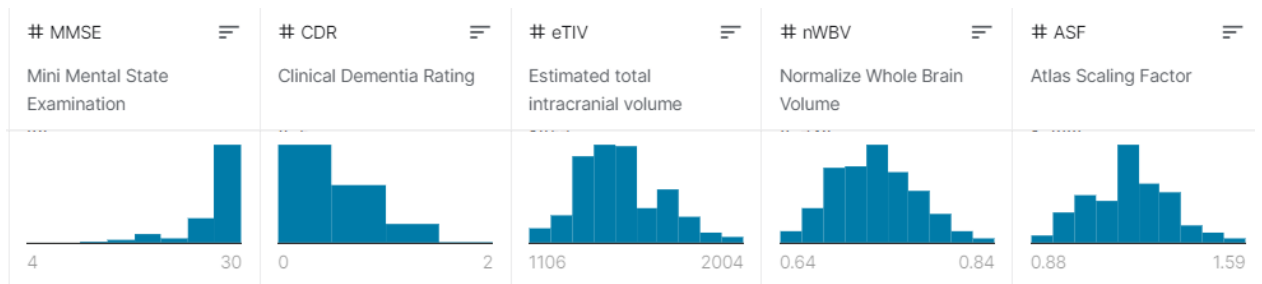
RF (random forest) — это множество решающих деревьев. В задаче регрессии их ответы усредняются, в задаче классификации принимается решение голосованием по большинству. Все деревья строятся независимо по следующей схеме:

- Выбирается подвыборка обучающей выборки размера `samplesize` (м.б. с возвращением) — по ней строится дерево (для каждого дерева — своя подвыборка).
- Для построения каждого расщепления в дереве просматриваем `max_features` случайных признаков (для каждого нового расщепления — свои случайные признаки).
- Выбираем наилучшие признак и расщепление по нему (по заранее заданному критерию). Дерево строится, как правило, до исчерпания выборки (пока в листьях не останутся представители только одного класса), но в современных реализациях есть параметры, которые ограничивают высоту дерева, число объектов в листьях и число объектов в подвыборке, при котором проводится расщепление.

Используемый датасет

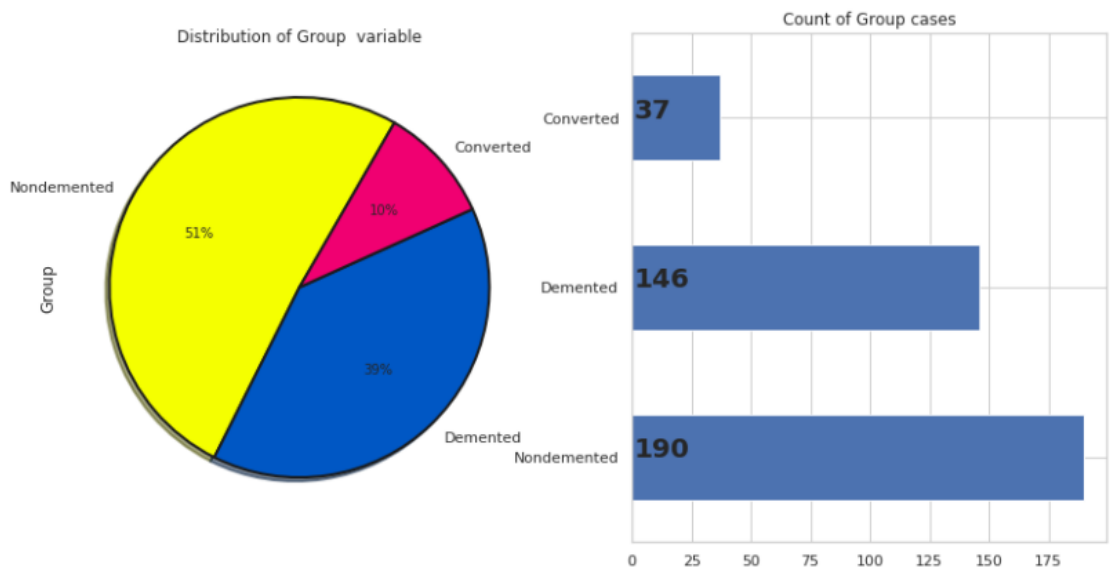
Признаки болезни альцгеймера (взяты с kaggle)

| ▲ Group | | ▲ M/F | | # Age | | # EDUC | | # SES | |
|-------------|-----|---------------|-----|---|--|--|--|---|--|
| Class | | Male - Female | | Age | | Years of Education | | Socioeconomic Status / 1-5 1 - Low 5 - High | |
| Nondemented | 51% | F | 57% |  | |  | |  | |
| Demented | 39% | M | 43% | | | | | | |
| Other (37) | 10% | | | | | | | | |



Group --> Class
 Age --> Age
 EDUC --> Years of Education
 SES --> Socioeconomic Status / 1-5
 MMSE --> Mini Mental State Examination
 CDR --> Clinical Dementia Rating
 eTIV --> Estimated total intracranial volume
 nWBV --> Normalize Whole Brain Volume
 ASF --> Atlas Scaling Factor

В данном датасете отсутствуют пропущенные значения, поэтому сразу можно выделить два значимых параметра – психическое состояние (MMSE) и объем мозга (eTIV). Разбиение на тестовые и тренировочные данные происходит случайным образом. Также разделим данные на две части.



Возьмем две группы пациентов по 100 человек – с альцгеймером и без в качестве даты.

Результаты:

Лабораторная 1

```

custom KNN accuracy = 0.8461538461538461
custom NB accuracy = 0.8461538461538461

sklearn KNN accuracy = 0.8846153846153846
sklearn NB accuracy = 0.8461538461538461
  
```

Лабораторная 2

```
sklearn log accuracy: 0.9565217391304348
custom log accuracy: 0.9565217391304348
LR train accuracy: 0.8248587570621468
sklearn Dtree accuracy: 0.9130434782608695
custom Dtree accuracy: 0.9565217391304348
DT train accuracy: 0.9096045197740112
sklearn RF accuracy: 0.9130434782608695
custom random forest accuracy: 0.8695652173913043
random forest train accuracy: 0.96045197740113
```

Выводы:

В результате выполнения работы были изучены различные алгоритмы машинного обучения. При повторной генерации тестовых данных вручную имплементированные алгоритмы иногда допускают на несколько ошибок больше, чем sklearn аналоги, но в общем работают с одинаковой точностью.

По применимости алгоритмов больше всего подходят логическая регрессия из-за разделения аргументов на 2 класса и маленького размера датасета.