МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ В МАШИННОМ ОБУЧЕНИИ

Нарвойш П.Ю., студент

Институт информационных технологий
Белорусского государственного университета информатики и радиоэлектроники
г. Минск, Республика Беларусь

Мацкевич И.Ю. – ст. препод. каф. ФМД

В работе описывается процесс подготовки данных для алгоритмов машинного обучения, рассматривается отбор признаков для обучения модели. Обнаружение мультиколлинеарности анализируется при расчете коэффициента корреляции Пирсона.

Машинное обучение – обширный подраздел искусственного интеллекта, изучающий методы построения алгоритмов, способных обучаться. Одним из типов машинного обучения является обучение по прецедентам, или индуктивное обучение, основанное на выявлении общих закономерностей по частным эмпирическим данным.

Общая постановка задачи обучения по прецедентам звучит так: Дано конечное множество прецедентов (объектов, ситуаций), по каждому из которых собраны (измерены) некоторые данные. Данные о прецеденте называют также его описанием. Совокупность всех имеющихся описаний прецедентов называется обучающей выборкой. Требуется по этим частным данным выявить общие зависимости, закономерности, взаимосвязи, присущие не только этой конкретной выборке, но вообще всем прецедентам, в том числе тем, которые ещё не наблюдались [1]. В дальнейшем под понятием машинного обучения будет иметься в виду обучение по прецендентам.

Машинное обучение сильно зависит от данных, от их построения и подбора признаков. Признаком называется результат измерения некоторой характеристики объекта и именно на основании признаков и строятся предсказания в моделях. Часто наборы данных, с которыми приходится работать, содержат большое количество признаков, число которых может достигать нескольких сотен и даже тысяч. При построении модели машинного обучения не всегда понятно, какие из признаков действительно для неё важны (т.е. имеют связь с целевой переменной), а какие являются избыточными (или шумовыми). Есть много причин, по которым включение тех или иных признаков в модель может привести к неудовлетворительным результатам. Одна из них – мультиколлинеарность.

Мультиколлинеарность – это явление, при котором одна из входных переменных статистической модели (например, логистической регрессии) линейно зависит от других входных переменных, т.е. между ними наблюдается сильная корреляция.

При этом различают полную коллинеарность, которая означает наличие функциональной (тождественной) линейной зависимости и частичную или просто мультиколлинеарность – наличие сильной корреляции между факторами. Если полная коллинеарность приводит к неопределенности значений параметров, то частичная мультиколлинеарность приводит к неустойчивости их оценок. Неустойчивость выражается в увеличении статистической неопределенности — дисперсии оценок. Это означает, что конкретные результаты оценки могут сильно различаться для разных выборок несмотря на то, что выборки однородны [2].

Обнаружить мультиколлинеарность можно, например, при помощи коэффициента корреляции Пирсона, вычислив его для каждой пары признаков. Коэффициент корреляции Пирсона характеризует существование линейной зависимости между двумя величинами и рассчитывается по формуле:

$$r_{xy} = \frac{\sum_{i=1}^{m}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{m}(x_i - \bar{x})^2 \sum_{i=1}^{m}(y_i - \bar{y})^2}} = \frac{cov(x,y)}{\sqrt{s_x^2 s_y^2}}, \quad (1)$$

где x, y – выборки длинной m, $\bar{x}, \bar{y}$ – выборочные средние, s_x^2, s_y^2 – выборочные дисперсии [3].

Рассчитаем коэффициент для всех пар признаков тестовых данных. Отбор признаков был произведен на тестовом наборе данных «Рак молочной железы». Набор данных взят с сайта https://www.kaggle.com/datasets/uciml/breast-cancer-wisconsin-data?datasetId=180&sortBy=voteCount. Для работы с данными используется язык программирования Python.

Результат расчета коэффициента корреляции Пирсона и визуализация полученных результатов при помощи тепловой карты представлены на рисунке 1.

Можно увидеть, что некоторые признаки сильно коррелируют между собой, то есть значения коэффициента Пирсона либо равен единице, либо очень близок к ней. Достаточно оставить один из таких признаков, что уменьшит объем данных для обработки алгоритмами, а также позволит получать более интерпретируемые результаты. Пороговое значение коэффициента для удаления признаков устанавливается опытным путем и зависит от конкретной задачи.

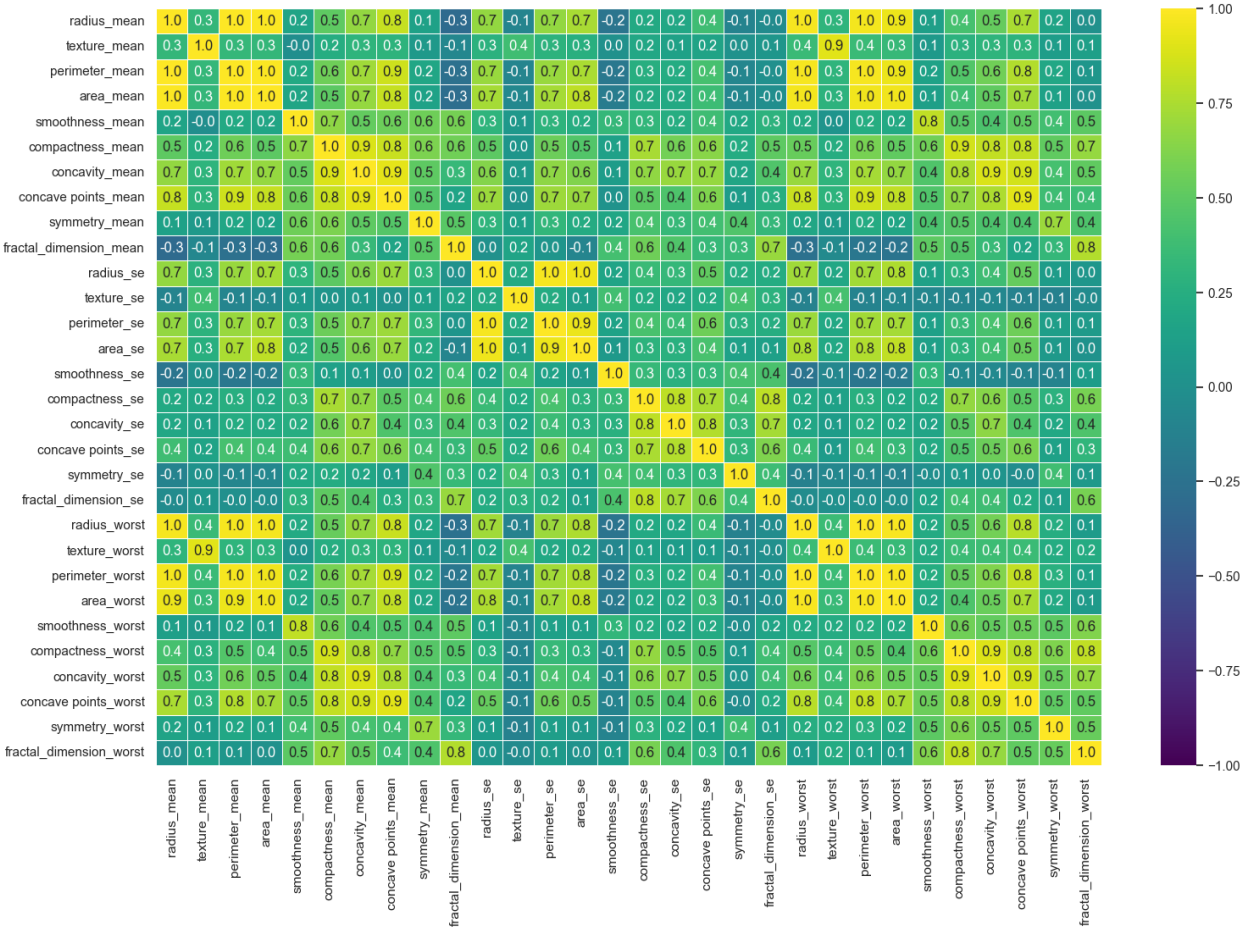

Рисунок 1 – Коэффициенты корреляции Пирсона для каждой пары признаков

Для сравнения работы алгоритмов машинного обучения на тестовых данных и очищенных тестовых данных была выбрана модель логистической регрессии.

Логистическая регрессия – это разновидность множественной регрессии, общее назначение которой состоит в анализе связи между несколькими независимыми переменными (называемыми также регрессорами или предикторами) и зависимой переменной.

В множественной линейной регрессии предсказывается непрерывная переменная со значениями на отрезке [0,1] при любых значениях независимых переменных. Это достигается применением следующего регрессионного уравнения (логит-преобразование, логистическое преобразование):

$$p = \frac{1}{(1+e^{-y})} \quad (2)$$

где p – вероятность того, что произойдет интересующее событие, y – стандартное уравнение регрессии.

Бинарная логистическая регрессия применяется в случае, когда зависимая переменная является бинарной (т.е. может принимать только два значения) [4].

Исходные данные содержали в себе 30 признаков. После отбора признаков при помощи коэффициента корреляции Пирсона с пороговым значением коэффициента 0.95 их осталось 22. В результате модель, обученная на отобранных данных, превзошла модель, обученную на исходном наборе данных, на 0,003. При этом была удалена почти треть признаков, что не привело к ухудшению работы алгоритма.

Избыточные признаки влияют на сложность модели машинного обучения, поэтому уменьшение количества признаков при сохранении точности модели приводит к уменьшению времени тренировки модели, что положительно сказывается на обработке данных, состоящих из большого количества признаков.

Список использованных источников:

1. Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [электронный ресурс]. – Режим доступа: http://www.machinelearning.ru. – Дата доступа: 02.04.2023.

2. Мультиколлинеарность [Электронный ресурс]. – Режим доступа: https://wiki.loginom.ru/articles/multicollinearity.html. – Дата доступа: 03.04.2023

3. Гришкина, Т.Е. Корреляционный анализ: метод. Указания / Т.Е.Гришкина. – Благовещенск: АмГУ, 2021. – 31 с.

4. Логистическая регрессия и ROC-анализ – математический аппарат [Электронный ресурс]. – Режим доступа: https://loginom.ru/blog/logistic-regression-roc-auc. – Дата доступа: 06.04.2023