УДК 004.85

Барков Денис Владимирович,
магистрант, Ангарский государственный технический университет,
e-mail: barkov.dev@yandex.ru
Сенотова Светлана Анатольевна,
к.т.н., доцент, Ангарский государственный технический университет,
e-mail: sveta-senotova@mail.ru

КОДИРОВАНИЕ КАТЕГОРИАЛЬНЫХ ПРИЗНАКОВ В НЕЙРОННЫХ СЕТЯХ

Barkov D.V., Senotova S.A.

ENCODING OF CATEGORIAL FEATURES IN NEURAL NETWORKS

Аннотация. Исследована актуальность и области применения машинного обучения, рассмотрен один из алгоритмов машинного обучения – нейронные сети, а также один из процессов подготовки данных перед получением математической модели – кодирование категориальных признаков с помощью метода целевого кодирования. Реализован алгоритм кодирования на языке программирования Python.

Ключевые слова: машинное обучение, нейронные сети, анализ данных, кодирование, программирование.

Abstract. The relevance and areas of application of machine learning are investigated, one of the machine learning algorithms - neural networks, as well as one of the data preparation processes before extracting a mathematical model - the coding of categorical features using the target coding method is considered. Implemented a coding algorithm in the Python programming language.

Keywords: Machine learning, neural networks, data analysis, coding, programming.

Одной из самых захватывающих технологий современности является машинное обучение (далее МО). Крупнейшие компании, такие как: Google, суперкомпьютеры которых обрабатывают миллионы поисковых запросов пользователей за счет применения алгоритмов МО, Facebook, Apple, Amazon, IBM и многие другие, инвестируют огромные деньги в разработку методов анализа данных.

Ключевой характеристикой как машинного, так и человеческого обучения является обобщение (общий вывод) – то есть, способность применять полученные в процессе тренировок знания к новым, ранее не встречавшимся образцам. Например, ребенку необходимо определить, кто изображен на картинке, собака или кошка. Ему поочередно показывают картинки, и в зависимости от правильного ответа картинка кладется в одну из стопок. Чем больше продолжается процесс итерации, тем выше эффективность распознавания. Самое главное, что ребенка необязательно специально учить отличать собаку от кошки, изображенную на картинке, человеческое сознание обладает встроенными механизмами классификации, ему требуются лишь образцы. Научившись работать с картинками, ребенок сможет распознать практически любое изображение кошки или собаки, не говоря уже о реальных животных.

Разумеется, что процесс получения знаний человеком превосходит своей сложностью самые совершенные алгоритмы машинного обучения, но у вычис-

лительной машины есть ряд своих преимуществ в виде большей производительности обработки, запоминания и извлечения информации.

Методология МО предлагает для получения знаний из данных следующую альтернативу – постепенное улучшение качества прогнозных моделей и принятие решений вместо того, чтобы в ручном режиме строить модели на основе анализа больших объёмов данных.

МО хорошо зарекомендовало себя в широком спектре задач: реклама товара, обнаружение мошенничества, наблюдение за производством в реальном времени, медицинская диагностика, распознавание личности, анализ рукописного текста, тональности текстов и голоса.

Один из способов интеллектуального анализа данных – это нейронные сети (далее НС), которые представляют собой математическую модель и ее программное воплощение, построенную по принципу организации и функционирования биологических НС – сетей нервных клеток живого организма. Можно сказать, что нейросеть – это машинная интерпретация мозга человека, в котором передают информацию в виде электрических импульсов миллионы нейронов.

Нейрон – это вычислительная единица, получающая информацию и производящая над ней простые вычисления, затем передающая ее дальше другим нейронам.

Подразделяются нейроны на следующие типы: входные, скрытые, выходные и ситуативные нейроны смещения (необходимы для получения выходного результата, путем сдвига графика функции активации, для попадания в точку, отвечающую за решение, их выходы всегда равны единице). В случае, когда НС состоит из большого количества нейронов, вводится понятие «слой». Следовательно, есть входной слой, на котором каждый нейрон получает в качестве входного параметра признак (предикторную переменную), например, возраст или среднюю заработную плату заемщика. Количество входных нейронов прямо пропорционально количеству предикторных переменных.

Далее они передают информацию в скрытый слой, где скрытые нейроны обрабатывают информацию с помощью синапсов (связей), у которых есть всего один параметр – вес. Именно благодаря ему входная информация изменяется, когда передается от одного нейрона к другому. Нейрон, у которого вес, соответствующий входному признаку, будет больше, будет доминировать в следующем нейроне. Совокупность весов НС – это своеобразный мозг всей системы, благодаря этим весам информация обрабатывается, строятся зависимости и превращаются в результат.

Расчет необходимого количества синаптических весов НС может быть осуществлен с помощью формулы 1, которая вытекает из теоремы Колмогорова – Арнольда-Хехт-Нильсена:

$$\frac{N_y \cdot Q}{1+log_2 \cdot Q} \leq N_w \leq N_y \cdot \left(\frac{Q}{N_x}+1\right) \cdot \left(N_x + N_y + 1\right) + N_y, \quad (1)$$

где N_x – количество входных предикторных переменных, определяющих число нейронов входного слоя;

N_y – количество нейронов выходного слоя, определяемое числом выходных переменных (в случае одобрения кредита – одна);

Q – размерность обучающей выборки (заявления 50 клиентов);

N_w – необходимое число синаптических связей.

Число нейронов скрытого слоя N может быть определено по формуле 2:

$$N = \frac{N_w}{N_x+N_y}. \quad (2)$$

Последним располагается выходной слой, который выводит результат. У каждого нейрона есть два основных параметра: входные и выходные данные, в случае входного нейрона вход равен выходу. В остальных на вход попадает суммарная информация всех нейронов с предыдущего слоя и обрабатывается.

Важно отметить, что нейроны оперируют числами в диапазоне от 0 до 1 или от -1 до 1, следовательно, категориальные признаки необходимо кодировать и нормализировать (для попадания в диапазон нейрона), а большие вещественные признаки просто нормализировать.

На рисунке 1 представлена схема простой нейросети, I – входные нейроны, H – скрытые нейроны, O – выходные нейроны, w_n – синапсы, B – bias (нейрон смещения).

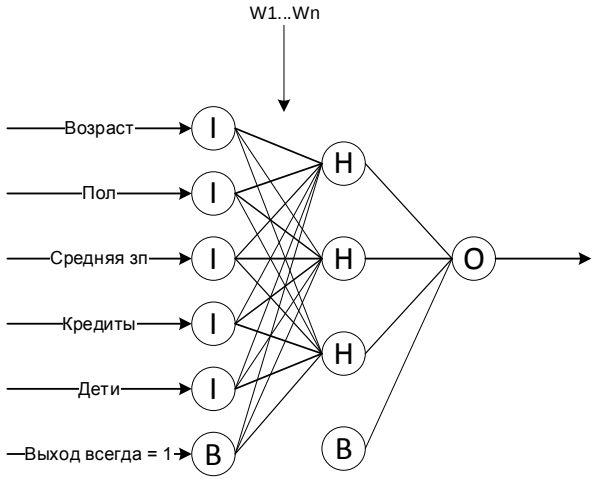

Рисунок 1 – Упрощенная схема нейронной сети

Сначала для реализации НС подготавливается тренировочный набор данных, который тщательно анализируется и сортируется, особенно категориальные признаки. Например, в предикторной переменной «регион проживания»,

можно обнаружить число, город, область, республику или край, которые к тому же, могут быть написаны с опечатками или с разными вариантами названий, в том числе устаревшими.

Следовательно, чтобы система могла оперировать значением данного признака, он обрабатывается (исправляются ошибки, различные варианты написания региона приводятся к одному) и кодируется с помощью целевого кодирования (target-based coding).

В целевом кодировании для каждого значения признака определяется среднее значение целевой переменной (одобрен = 1, не одобрен = 0) в наборе данных. Для применения кодировки необходимо выполнить следующие действия:

– Выбрать категориальную переменную, которую следует преобразовать;

– Подсчитать отношение категориального признака к метке «одобрен» (шаг 1 на рисунке 2);

– Подсчитать общее количество каждого значения признака (шаг 2 на рисунке 2);

– Найти частное от количества целей и суммы целей, затем подставить числа в таблицу вместо строк.

На рисунке 2 продемонстрирован алгоритм целевого кодирования. Из рисунка 2 видно, что значения Санкт-Петербург и Ангарск включают целевую переменную, что в итоге приведет к ухудшению прогноза системы.

Чтобы решить данную проблему, используется один из вариантов реализации целевого кодирования – сглаживание, которое можно применить, используя формулу 3:

$$smooth = \frac{(counts \cdot means + weight \cdot mean)}{(counts + weight)}, \quad (3)$$

где $counts$ – количество целей для каждого возможного значения;
$means$ – отношение количества целей к сумме целей каждого значения;
$mean$ – отношение количества целей к их общему количеству;
w$eight$ – константа, всегда равна 100:

Данный вариант реализации целевого кодирования позволяет более точно скорректировать значения для каждого признака. Это приведет к тому, что НС будет определять, у какого признака вес больше, даже если до сглаживания их кодированные значения были одинаковыми.

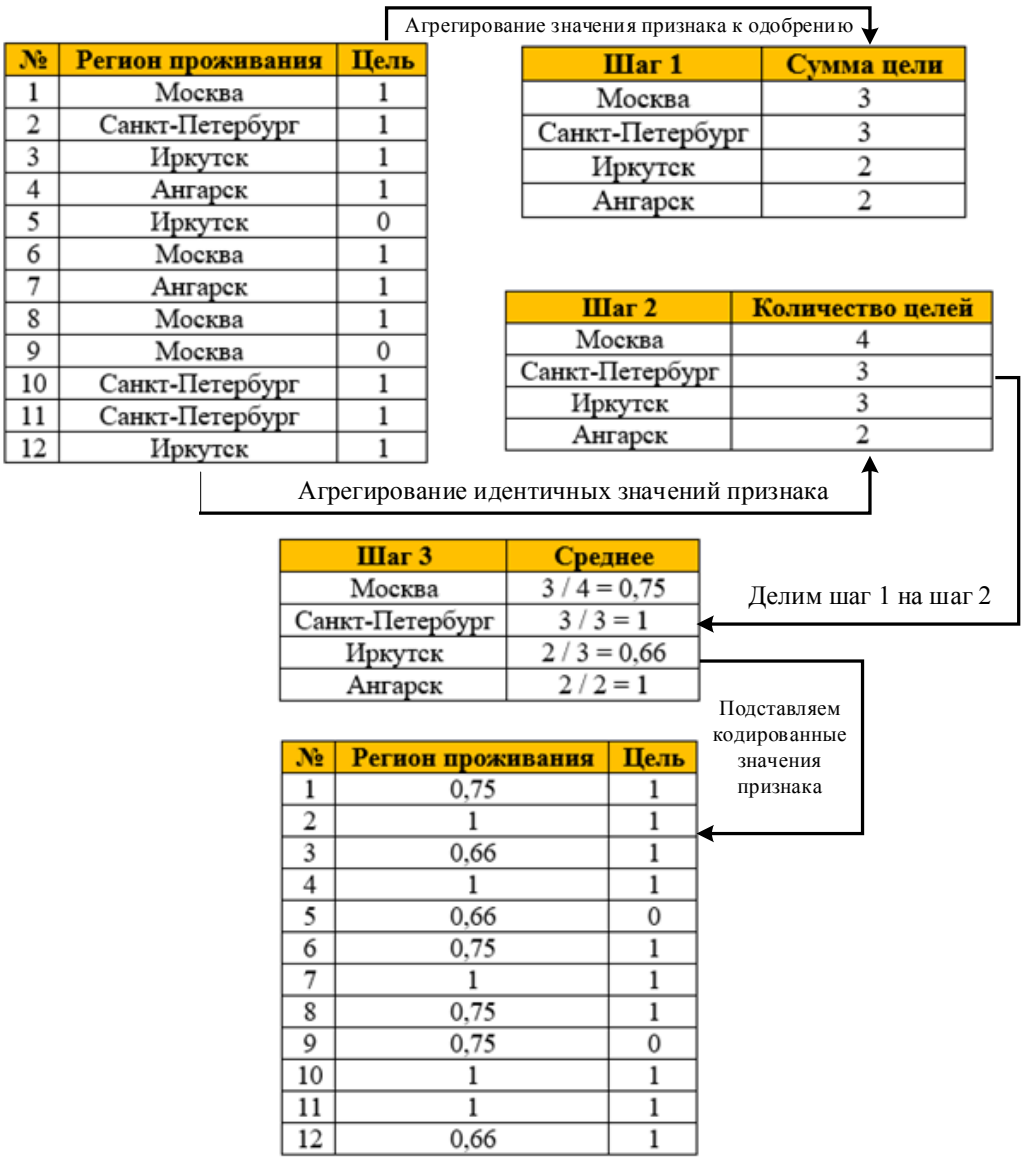

Рисунок 2 – Принцип работы алгоритма целевого кодирования

Далее представлен код на языке программирования Python, который реализует сглаживание значений категориальных признаков.

```
import pandas as pd #Импортируем библиотеку для обработки и анализа данных
#Инициализируем словарь и заполняем тестовыми данными
data = {'Регион проживания': ['Москва', 'Санкт-
Петербург', 'Иркутск', 'Ангарск', 'Иркутск', 'Москва', 'Ангарск', 'Москва', 'Москва',
'Санкт-Петербург', 'Санкт-Петербург', 'Иркутск']
'Цель': [1,1,1,1,0,1,1,1,0,1,1,1]}
#Заносим в таблицу данные из словаря
df = pd.DataFrame(data,columns=['Регион проживания', 'Цель'])
```

```
#Подсчитываем отношение количества целей к их общему количеству = 10/12 =
 0.8333333333333334
mean = df['Цель'].mean()
#Преобразуем категориальные значения в числовые, применяя функции
агрегирования к столбцу Цель
agg = df.groupby('Регион проживания')['Цель']. agg(['count', 'mean'])
counts = agg['count'] #количество целей для каждого возможного значения
means = agg['mean'] #отношение количества целей к сумме целей каждого
значения
weight = 100 #Константа вес
#Формула сглаживания
smooth = (counts * means + weight * mean) / (counts + weight)
#Печатаем в терминал результат
print(smooth)
```

На рисунке 3 представлен результат выполнения кода.

```
Регион проживания
Ангарск            0.836601
Иркутск            0.828479
Москва             0.830128
Санкт-Петербург    0.838188
dtype: float64
```

Рисунок 3 – Результат выполнения кода сглаживания значений

ЛИТЕРАТУРА

1. Рашка С. Python и машинное обучения: / пер. с англ., – М.: ДМК Пресс, 2017, – 418 с.: ил;
2. Хенрик Бринк., Джозеф Ричардс., Марк Феверолф. Машинное обучение – СПб.: Питер, 2017. – 336 с.: ил. – (Серия «Библиотека программиста»);
3. Pandas.docs / Информационная система [Электронный ресурс] – https://pandas.pydata.org / (Дата последнего обращения: 20.02.2020);