Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

ОТЧЕТ ПО ЛАБОРАТОРНОЙ РАБОТЕ №1

по курсу «Машинное обучение»

Основы машинного обучения

Студент:
гр. 758641
Ярош Г.И.

Проверил:
Заливако С. С.

Минск, 2019

СОДЕРЖАНИЕ

ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ 3
1. Цель 3
2. Набор данных notMNIST 3
3. Проверка данных на сбалансированность 3
4. Удаление повторяющихся изображений 5
5. Классификация изображений 5
6. Вывод 6
СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ 7

ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

1. Цель

Изучить основы машинного обучения на задаче классификации изображений букв из набора данных notMNIST с помощью логистической регрессии.

2. Набор данных notMNIST

В данной лабораторной работе проводилось исследование возможности классификации изображений из набора данных notMNIST.

Набор данных состоит из двух выборок: обучающей и тестовой. Обучающая выборка содержит около 500 тысяч черно-белых изображений разрешением 28x28 пикселей. Данные изображения разделены по папкам, изображения в каждой из которых соответствуют первым 10 буквам латинского алфавита A…J. Тестовая выборка содержит порядка 19 тысяч изображений. Пример изображений можно увидеть на рисунке 1.

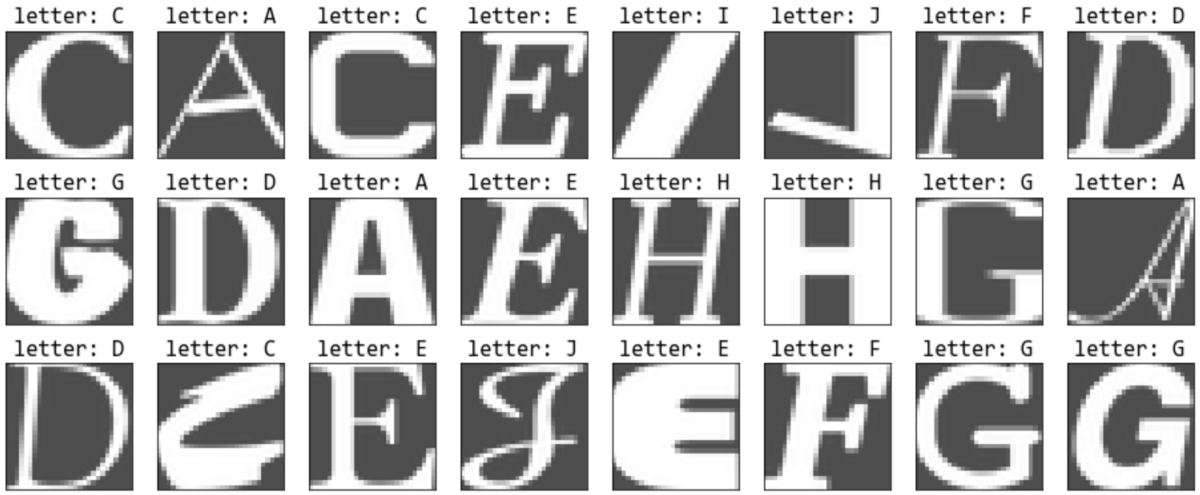

Рис. 1. Пример изображений из архива notMNIST

3. Проверка данных на сбалансированность

Для успешной классификации желательно, чтобы количество изображений по каждому классу в выборке было примерно одинаково. В таблице 1 и на рисунке 2 отображены распределения классов в тестовой выборке.

A	B	C	D	E	F	G	H	I	J
52909	52911	52912	52911	52912	52912	52912	52912	52912	52911

Таблица 1. Распределение классов в обучающей выборке.

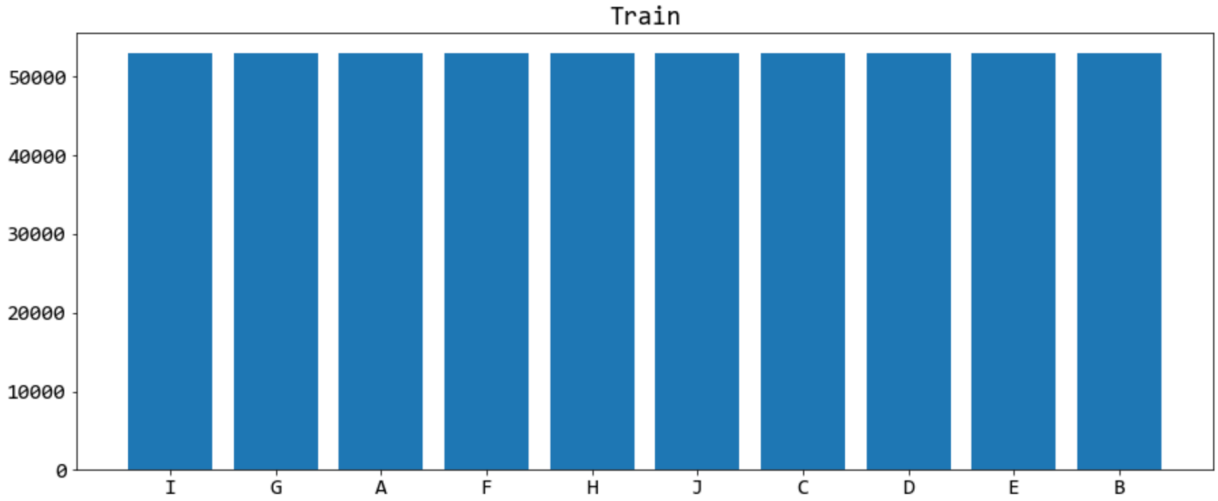

Рис. 2. Распределение классов в обучающей выборке.

Соответственно, распределение классов в тестовой выборке приведено в таблице 2 и на рисунке 3.

A	B	C	D	E	F	G	H	I	J
1872	1873	1873	1873	1873	1872	1872	1872	1872	1872

Таблица 2. Распределение классов в обучающей выборке.

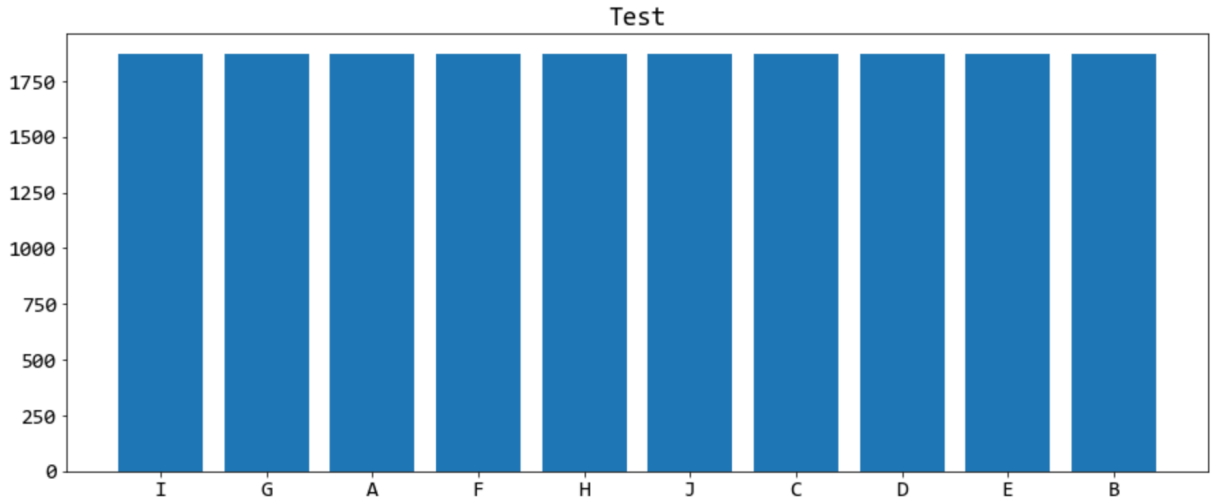

Рис. 3. Распределение классов в тестовой выборке.

Как мы видим из приведенных таблиц и рисунков, классы достаточно сбалансированы как и в обучающей, так и в тестовой выборках.

4. Удаление повторяющихся изображений

Для того, чтобы эффективно оценить обобщающую способность модели на тестовой выборке, необходимо, чтобы данные из тестовой выборки не пересекались с данными из обучающей выборки.

При проверке на пересечение выборок, обнаружилось, что 12213 изображений из обучающей выборки содержаться в тестовой. После удаления повторяющихся изображений распределение классов в тестовой выборке приняло следующий характер (рисунок 4 и таблица 3):

A	B	C	D	E	F	G	H	I	J
52322	52274	52213	52241	52290	52266	52276	52132	46651	52236

Таблица 3. Распределение классов в обучающей выборке после удаления пересечений с тестовой.

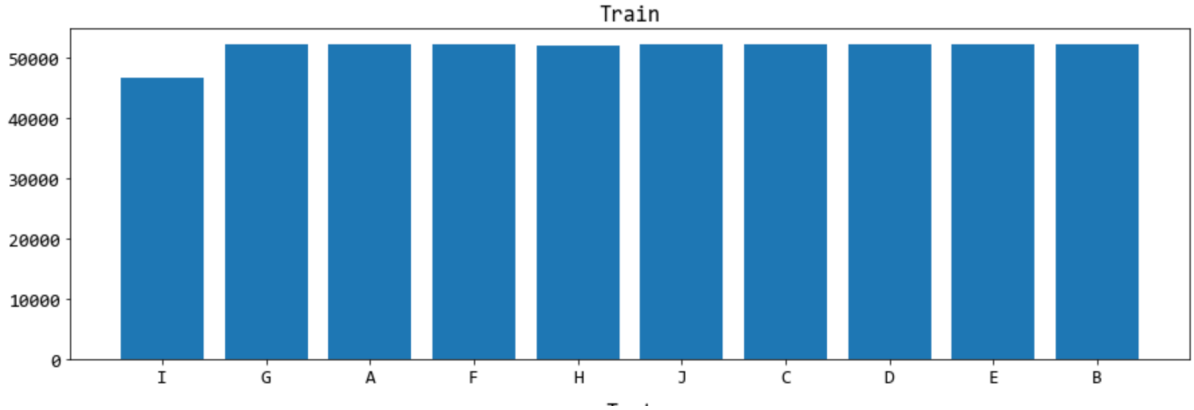

Рис. 4. Распределение классов в обучающей выборке после удаления пересечений с тестовой.

Как мы видим, после удаления пересечений класс I чувствительно потерял в представленности в обучающей выборке. Для остальных классов количество удаленных изображений оказалось несущественно.

5. Классификация изображений

Для решения задачи классификации исходные изображения были преобразованы в вектор x длиной 784. Также был сформирован вектор результатов y, где значение y_i соответствовало классу изображения x_i. Далее из обучающей выборки была выделена валидационная выборка размером в 10% от исходной.

В качестве классификатора мною была выбрана логистическая регрессия с мультиноминальной функцией потерь.

В качестве метрики эффективности классификации мною была выбрана категориальная точность, которая вычислялась по следующей формуле:

$$accuracy = \frac{1}{N}\sum_{i=0}^{N} 1(y_i^0 = y_i),$$

где N – размер тестовой выборки, y_i^0, y_i – исходное и предсказанные значения классов соответственно.

Обучение модели происходило в несколько этапов с изменяющимся размером обучающей выборки. На каждом этапе вычислялась точность предсказания модели на валидационной выборке. Зависимость точности классификации от размера обучающей выборки приведена на рисунке 5.

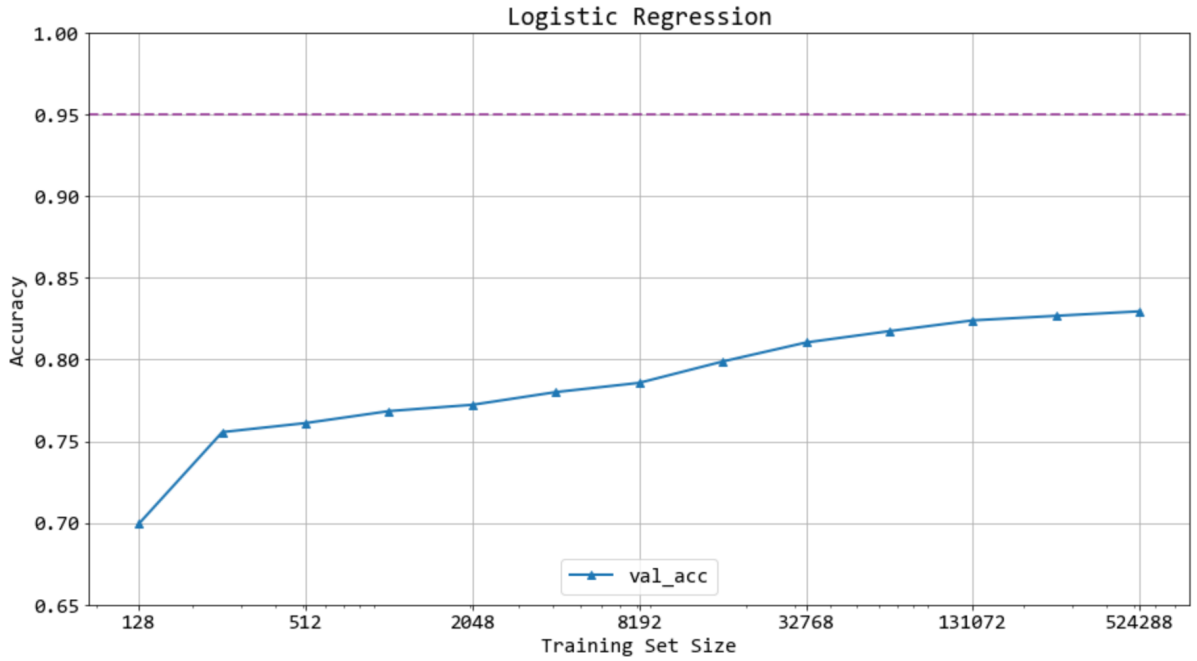

Рис. 5. Зависимость точности классификации от размера обучающей выборки.

Как мы видим из графика, точность классификации растет при росте размера обучающей выборки и достигает 0.83 при обучении на всей выборке.

Точность обученной модели на всей обучающей выборке при проверке на тестовой выборке составила 0.84.

6. Вывод

В результате работы был проанализирован архив notMNIST. Была построена базовая модель классификации изображений букв с использованием логистической регрессии.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] – notMNIST dataset [Электронный ресурс]. – Электронные данные. – Режим доступа: http://yaroslavvb.blogspot.com/2011/09/notmnist-dataset.html.

[2] – Sklearn Logistic Regression Classifier [Электронный ресурс]. – Электронные данные. – Режим доступа: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.