Министерство образования Республики Беларусь

Учреждение образования
БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИНФОРМАТИКИ И РАДИОЭЛЕКТРОНИКИ

Факультет Компьютерных сетей и систем

Кафедра Информатики

МАШИННОЕ ОБУЧЕНИЕ

ЛАБОРАТОРНАЯ РАБОТА №1
«Логистическая регрессия в качестве нейронной сети»

БГУИР 1-40 81 04

Магистрант:
гр. 858641
Кукареко А.В.

Проверил:
Стержанов М. В.

Минск, 2020

ХОД РАБОТЫ

Данные.

В работе предлагается использовать набор данных notMNIST, который состоит из изображений размерностью 28×28 первых 10 букв латинского алфавита (A … J, соответственно). Обучающая выборка содержит порядка 500 тыс. изображений, а тестовая – около 19 тыс.

Описание данных от автора:

Набор данных состоит из небольшой части, очищенной вручную, около 19 000 экземпляров, и большого неочищенного набора данных, 500 000 экземпляров. Две части имеют примерно 0,5% и 6,5% ошибок в метках.

Задание.

1. Загрузите данные и отобразите на экране несколько из изображений с помощью языка Python;
2. Проверьте, что классы являются сбалансированными, т.е. количество изображений, принадлежащих каждому из классов, примерно одинаково (В данной задаче 10 классов).
3. Разделите данные на три подвыборки: обучающую (200 тыс. изображений), валидационную (10 тыс. изображений) и контрольную (тестовую) (19 тыс. изображений);
4. Проверьте, что данные из обучающей выборки не пересекаются с данными из валидационной и контрольной выборок. Другими словами, избавьтесь от дубликатов в обучающей выборке.
5. Постройте простейший классификатор (например, с помощью логистической регрессии). Постройте график зависимости точности классификатора от размера обучающей выборки (50, 100, 1000, 50000). Для построения классификатора можете использовать библиотеку SkLearn (http://scikit-learn.org).

Результат выполнения:

1. Загрузите данные и отобразите на экране несколько из изображений с помощью языка Python.

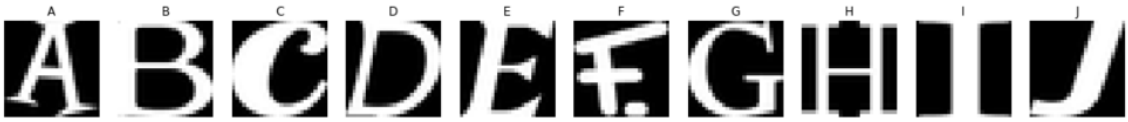

Рисунок 1 – пример данных из набора notMNIST.

2. Проверьте, что классы являются сбалансированными, т.е. количество изображений, принадлежащих каждому из классов, примерно одинаково (В данной задаче 10 классов).

Таблица 1 – Кол-во изображений в каждом классе.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	52909	52911	52912	52911	52912	52912	52912	52912	52912	52911

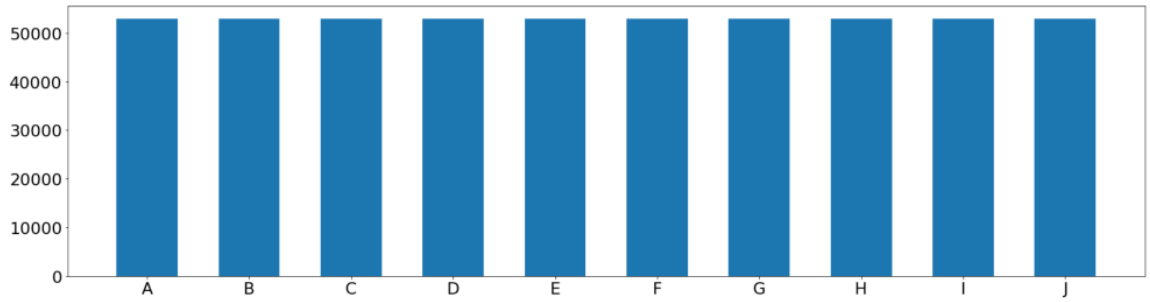

Рисунок 2 - кол-во изображений в каждом классе.

Если посмотреть на таблицу 1 и рисунок 2, то можно сделать вывод, что классы являются сбалансированными.

3. Разделите данные на три подвыборки: обучающую (200 тыс. изображений), валидационную (10 тыс. изображений) и контрольную (тестовую) (19 тыс. изображений).

В качестве тестовой выборки будет выступать малый датасет «notMNIST_small.tar.gz». Для обучающей и валидационной выборки будет использован большой датасет «notMNIST_large.tar.gz».

Так же дальше по заданию будет проведен эксперимент со следующими параметрами:

```
experiments = {
    '50': {'train': 50, 'valid': 20},
    '100': {'train': 100, 'valid': 40},
   '1000': {'train': 1000, 'valid': 400},
    '5000': {'train': 5000, 'valid': 1000},
    '20000': {'train': 20000, 'valid': 1800},
    '50000': {'train': 50000, 'valid': 2500},
    '100000': {'train': 100000, 'valid': 5000},
    '200000': {'train': 200000, 'valid': 10000},
}
```

4. Проверьте, что данные из обучающей выборки не пересекаются с данными из валидационной и контрольной выборок. Другими словами, избавьтесь от дубликатов в обучающей выборке.

Первым делом от дубликатов будет очищен малый датасет «notMNIST_small.tar.gz», который будет использован в качестве “test” выборки.

Таблица 2 – кол-во изображений в каждом классе до очистки.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	1872	1873	1873	1873	1873	1872	1872	1872	1872	1872

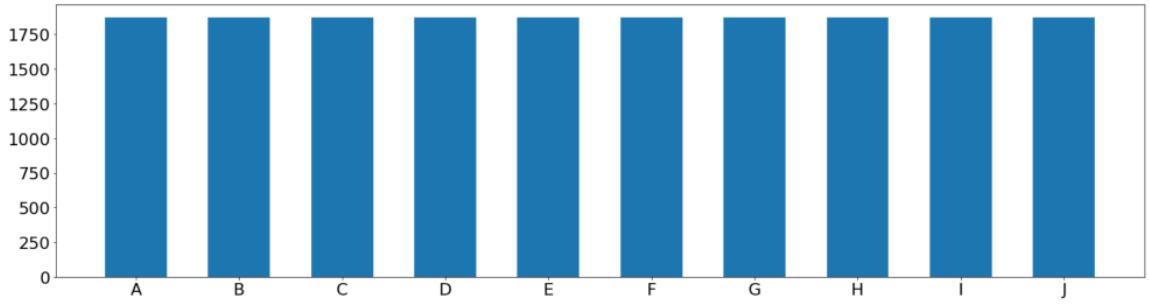

Рисунок 3 - кол-во изображений в каждом классе до очистки.

После очистки:

Таблица 3 – Кол-во изображений в каждом классе после очистки.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	1848	1853	1849	1847	1847	1850	1851	1846	1597	1850

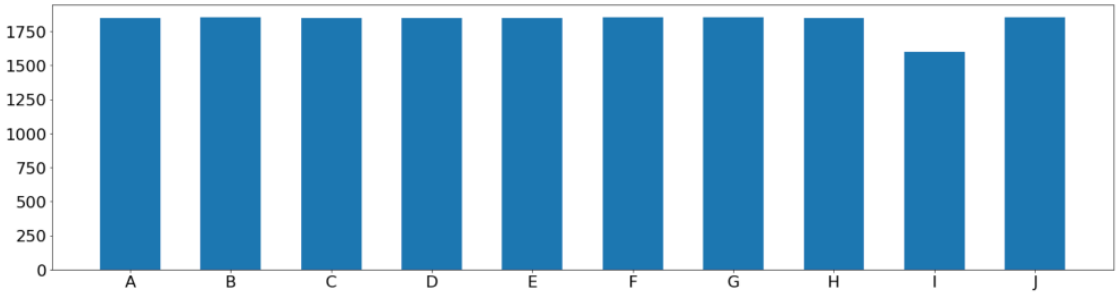

Рисунок 4 - кол-во изображений в каждом классе после очистки.

После очистки, можно заметить, что доля класса «I» уменьшилась по сравнению с остальными классами.

Размер датасета до очистки: 18724 изображений.

Размер датасета после очистки: 18238 изображений, из датасета удалили 486 дубликатов.

Теперь почистим от дубликатов большой датасет «notMNIST_large.tar.gz».

Таблица 4 – кол-во изображений в каждом классе до очистки.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	52909	52911	52912	52911	52912	52912	52912	52912	52912	52911

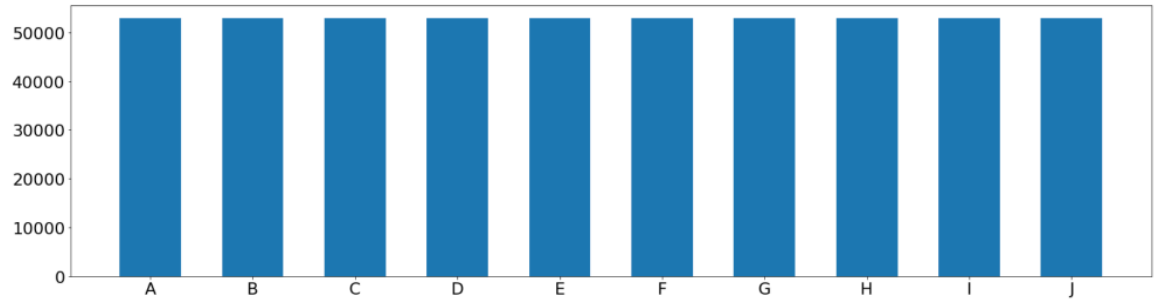

Рисунок 5 - кол-во изображений в каждом классе до очистки.

После очистки:

Таблица 5 – кол-во изображений в каждом классе после очистки.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	47102	47281	46654	46731	46954	46844	47090	46182	41173	46658

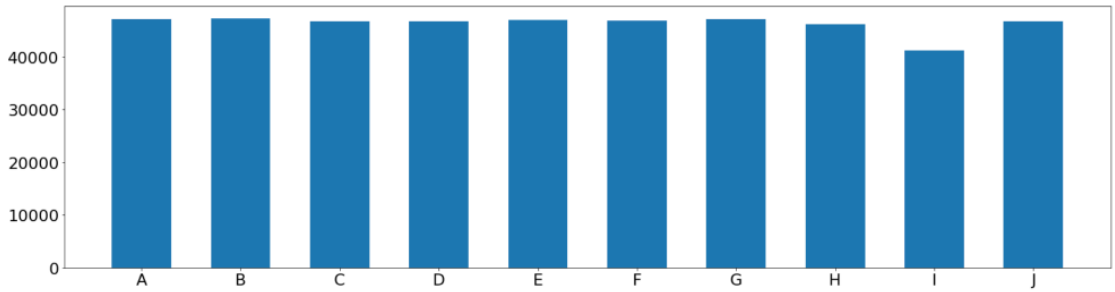

Рисунок 6 - кол-во изображений в каждом классе после очистки.

Теперь из большого датасета удалим пересечения из малого:

Таблица 6 – кол-во изображений в каждом классе после удаления пересечений.

Классы	A	B	C	D	E	F	G	H	I	J
Кол-во	46652	46852	46185	46264	46497	46383	46646	45677	40732	46193

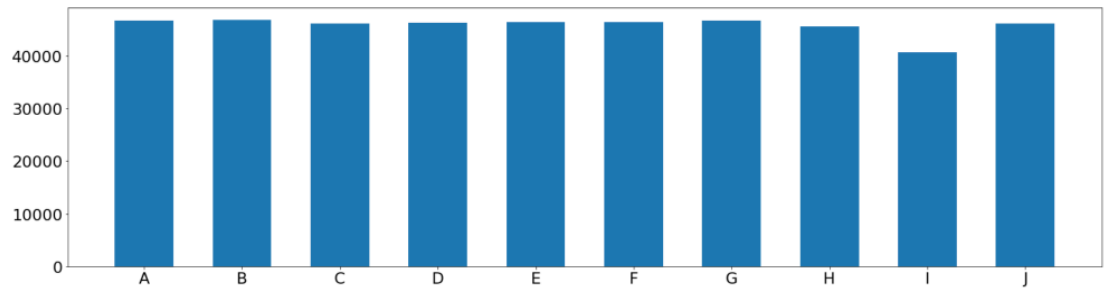

Рисунок 7 - кол-во изображений в каждом классе после удаления пересечений.

Таблица 7 – сравнение размеров датасета после преобразований.

notMNIST_large.tar.gz	**Кол-во изображений**
До очистки	529 114
После очистки	462 669
Удаление пересечений	458 081

В результате различных преобразований, из датасета «notMNIST_large.tar.gz» была удалена 71 тыс. (71 033) изображений.

Если посмотреть на таблицу 6 и рисунок 7, можно заметить, что класс «I» значительно меньше чем остальные классы, что говорит нам о дисбалансе, однако, так как мы не будем использовать весь датасет, а только его половину (200 тыс. изображений), наши классы останутся сбалансированными.

5. Постройте простейший классификатор (например, с помощью логистической регрессии). Постройте график зависимости точности классификатора от размера обучающей выборки (50, 100, 1000, 50000).

Для построения классификатора была использована библиотека «sklearn», модель «linear_model. LogisticRegression».

Так же был проведено 8 экспериментов:

Таблица 8 – результаты экспериментов.

Эксперимент	**Train acc**	**Valid acc**	**Test Acc**
train - 50, valid - 20, test - 19к	1.0	0.95	0.602
train - 100, valid - 40, test - 19к	1.0	0.8	0.722
train - 1000, valid - 400, test - 19к	1.0	0.9	0.827
train – 5к, valid – 1000, test - 19к	0.985	0.882	0.829
train – 20к, valid – 1800, test - 19к	0.876	0.819	0.855
train – 50к, valid – 2500, test - 19к	0.839	0.814	0.876
train – 100к, valid – 5к, test - 19к	0.829	0.831	0.880
train – 200к, valid – 10к, test - 19к	0.82	0.812	0.887

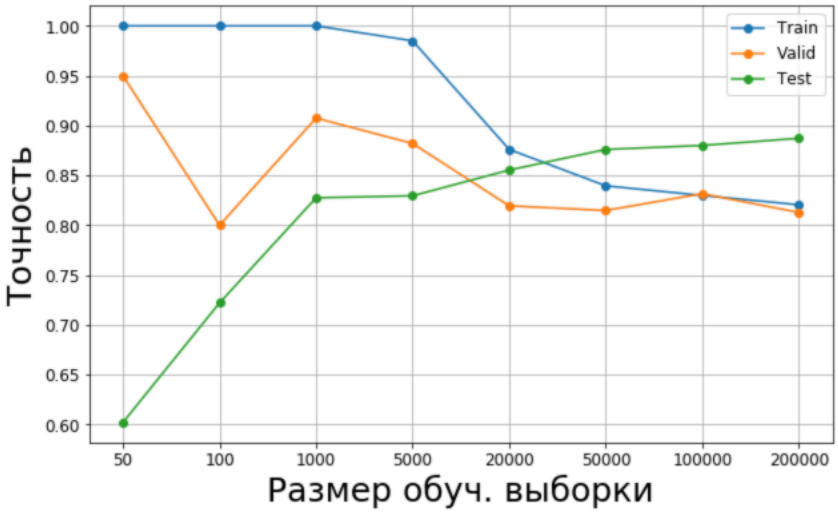

Рисунок 8 – график зависимости точности классификатора от размера обучающей выборки.

Если посмотреть на график, то можно заметить, что после отметки в 20к, график «тестовой» выборки начинает опережать графики «обучающей» и «вариационной» выборок. Это связано с тем, что обучающей датасет «notMNIST_large.tar.gz» содержит 6,5% ошибок в метках, а «тестовый» датасет «notMNIST_small.tar.gz» всего лишь 0,5%.

Вывод.

В ходе выполнения лабораторной работы мной был проанализирован и «очищен» датасет «notMNIST». Так же его основе был построен классификатор с использованием логистической регрессии.