**Постановка задачи**

**Источник: https://opendata.mkrf.ru/opendata/7705851331-stat\_museum**

Данные о музеях России: размер коллекции, график работы и т.п.

Структура:

**name** - Название музея

**base\_fond** - Число предметов основного фонда на конец года

**science\_fond** - Число предметов научно-вспомогательного фонда на конец года

**ams** - Автоматизированная музейная система: число внесенных музейных предметов

**n\_buildings** - Число строений

**n\_days** - Число дней в году, открытых для посещения

**guests** - Число посещений выставок, экспозиций + экскурсионных посещений, тысяч человек

**employees** - Численность работников, человек

Задача: вычисление значения параметра guests на основании остальных параметров.

**Подход к обработке данных.**

Ко всем столбцам кроме base\_fond , science\_fond , guests было применено MinMaxScaler().fit\_transform(data).flatten().

Выбросы были в столбцах science\_fond, ams, n\_days, n\_buildings, Employees, Guests к этим столбцам применилnp.clip(data, 0, 0).

В столбцах science\_fond, n\_buildings, guest, employees было большое количество нулей и отсутствующие значения. Я заменил их на среднее значение.

Квадратный корень применил к столбцам ams, n\_buildings, emloyees.

Name – не стал его обрабатывать так как в нем уникальные и независимые значение.

**Базовые параметры нейронной сети.**

функция активации - ReLU

Количество нейронов входного слоя – 6

Первый скрытый слой – 6

Второй скрытый слой – 6

Выходной слой – 1

Количество эпох – 50

Батчей – 16

Выборка 20 %

**Описание проведенных экспериментов.**

Для всех моделей были одинаковые параметры для объективности экспериментов.

Первая модель работает правильно, графики сошлись.

Во второй модели взял выборку 10 на 90. Графики почти сошлись.

В Третий модели взял выборку 60 на 40, графики сошлись.

Все графики корректны.

**Итоговый вывод.**

Выборка данных нужна для обучения модели на части данных. Выборку обычно делают 20 тестовых на 80 обучающих или 30 тестовых на 70 обучающих. Обычно выборка происходит случайным образом из этого следует что, если изначальных данных слишком много, то нейронная сеть будет переобучена, то есть она будет слишком узконаправленна, а если данных будет слишком мало, то конечная сеть будет слишком свободна и может не соответствовать поставленным требованиям.