ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ

ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ

«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

**Факультет информатики, математики и компьютерных наук**

**Программа подготовки бакалавров по направлению   
38.03.05 Бизнес-информатика**

*Сомов Александр Сергеевич*

**КУРСОВАЯ РАБОТА**

Исследование возможности применения нейронных сетей

для прогнозирования погоды

|  |  |
| --- | --- |
| Рецензент  д-р …. наук, проф.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  И.О. Фамилия | Научный руководитель.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  Масленникова Елизавета Андреевна  Консультант (при наличии)  д-р …. наук, проф.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  И.О. Фамилия |

Нижний Новгород, 2019

Содержание

**Введение 1**

**Глава 1: Постановка задачи прогнозирования погоды 4**

Обзор литературы 5

Постановка задачи подробно 5

**Глава 2: Нейронные сети для задач прогнозирования погоды 4**

**Глава 3: Результаты исследования 4**

3.1. Описание датасета 5

3.2. Эксперименты 5

**Заключение 4**

**Введение:**

Погода – это совокупность значений метеорологических элементов и атмосферных явлений, наблюдаемых в определённый момент времени в той или иной точке пространства. Исследованием возможности предсказания погоды занимались многие умы во все времена. Еще греки в 7-м веке до нашей эры начинали изучение науки под названием метеорология, а в 340 году до нашей эры Аристотель обнародовал свой труд под названием “Метеорологика”, где описывал паттерны по которым можно было предсказать погоду. Он же считается основателем такой науки как метеорология.

Данная область активно развивается с начала 20-го века, когда ученым пришла в голову идея прогнозирования погоды с помощью численных методов. Первой работой в этой области считается труд американского метеоролога Кливленда Эббе, который в своей основной работе «**The physical basis of long-range weather forecasting**» впервые привлек внимание математиков к данной проблеме. Первым же, кто действительно применил численные методы считается Льюис Фрай Ричардсон в работе под названием «**Weather Prediction by Numerical Process**». Однако, в те времена людям не хватало вычислительных мощностей, чтобы точно составлять прогнозы, так как все вычисления производились вручную.

Еще со времен прошлого века люди используют сложные системы дифференциальных уравнений для составления прогнозов. Стоит отметить, что для их решения используются огромные вычислительные мощности, но даже при таких условиях точность не достигает 100%.

В своей курсовой работе я исследую возможность применения нейронных сетей для задач прогнозирования погоды, так как нейронные сети хорошо решают многие задачи связанные с различными областями науки. Нейронные сети считаются гораздо более быстрым алгоритмом по сравнению с существующими в данной области. Хоть и для достижения точности нынешних алгоритмов понадобится большая работа, интересно узнать возможно ли применить эту технологию для решения данной проблемы.

Задача предсказания погоды крайне актуальна на данный момент. Есть множество компаний таких как Yandex, Accuweather, Gismeteo которые занимаются прогнозированием погоды. Так же для этих нужд созданы различные организации, например, World Meteorological Organization (WMO) и European Centre for Medium-Range Weather Forecasts.

Научная новизна моего исследования в том, что хоть многие компании пытаются прогнозировать погоду, однако мало кто использует для этого только нейронные сети. Я же решил попробовать делать прогнозы только с помощью них.

Предметом моего исследования является возможность использования нейронных сетей для этой задачи. Объектом же исследования является задача прогнозирования погоды.

Суммируя всё вышесказанное, целью моей курсовой работы является исследование возможности применения нейронных сетей для задачи предсказания погоды. Для этого необходимо исследовать предыдущие работы в данной области, сделать выводы и на основе анализа полученных данных попытаться построить модель с использованием нейронных сетей.

**Глава 1: Постановка задачи прогнозирования погоды**

В данной главе рассмотрены различные модели, которые применялись к этой задаче; описана модель, которая используется в курсовой работе для решения задачи; описаны различные подходы к её решению,

* 1. **Обзор методов и алгоритмов для решения задачи**

Идея предсказания погоды с помощью численных методов берет свое начало в 20-м веке. Первая успешная модель предсказания погоды была открыта Норманом Филлипсом и описана в работе "The general circulation of the atmosphere: a numerical experiment" (<https://rmets.onlinelibrary.wiley.com/doi/pdf/10.1002/qj.49708235202>).

Далее за эту проблему брались многие ученые и организации. Прорывным стало изобретение модели компанией “The United States' National Center for Atmospheric Research”. Они разработали модель “Community Atmosphere Model” (<http://www.cesm.ucar.edu/models/atm-cam/docs/description/description.pdf>) , которая предсказывала погоду по всему земному шару. Данная модель до сих пор улучшается и применяется для предсказания погоды. Однако, минус данной модели в том, что она требует огромных вычислительных ресурсов.

В начале 21-го века началось бурное развитие нейронных сетей. Иску́сственная нейро́нная се́ть — математическая модель, а также её программное или аппаратное воплощение, построенная по принципу организации и функционирования биологических нейронных сетей — сетей нервных клеток живого организма. Это понятие возникло при изучении процессов, протекающих в мозге, и при попытке смоделировать эти процессы. Первой такой попыткой были нейронные сети У. Маккалока и У. Питтса. После разработки алгоритмов обучения получаемые модели стали использовать в практических целях: в задачах прогнозирования, для распознавания образов, в задачах управления и др. У искусственных нейронных сетей есть несколько преимуществ перед классическими алгоритмами. Первое преимущество состоит в скорости. Нейронная сеть с правильно выстроенной архитектурой работает в десятки раз быстрее алгоритма. Второе преимущество состоит в том, что есть задачи, которые алгоритмы решить не в состоянии. Например, задача распознавания образов. Третье и самое важное преимущество перед алгоритмами в том, что искусственные нейронные сети могут сами находить паттерны в признаках и на основании этих паттернов делать выводы. Таким образом они позволяют решать задачи без определенного алгоритма.

* 1. **Постановка задачи предсказания погоды с помощью нейронных сетей.**

Представленные модели являются классическими для задачи предсказания погоды. Однако, рассматривая задачу с точки зрения вычислительных мощностей, которые требуются для ее решения, можно сделать вывод, что прогнозирование погоды – это очень затратная область. В связи с этим в данной курсовой работе будет представлена новая модель, не требующая столь больших мощностей для данной задачи.

Рассмотрим данные, которые будут использоваться в дальнейшем для предсказания. Данные взяты с сайта <https://www.ncdc.noaa.gov/cdo-web/search>. Датасет основан исключительно на реальных данных о погоде в США. Мной было отобрано 12 различных городов по всей территории Соединенных Штатов. Города выбирались по следующему критерию: в городе должен быть аэропорт. Такой критерий выбран не случайно. К сожалению, из всего многообразия метеорологических станций на территории США для исследования подходят данные далеко не со всех. Дело в том, что в данных имеются пропуски и(или) в них нет необходимых нам признаков. Так вот, почему аэропорты? Потому что в данных из аэропортов практически нет пропусков и все необходимые признаки присутствуют.

Данные из себя представляют упорядоченный по дате набор признаков для каждого дня. То есть для i-й даты данные представляют собой вектор числовых значений признаков для каждого из 12-ти городов. Набор признаков включает в себя максимальную и минимальную суточную температуру, среднесуточную скорость ветра, направление самого сильного ветра, который дул в одну сторону в течение пяти минут, а также высоту над уровнем моря на которой располагается метеорологическая станция. Предлагаю подробнее на них остановиться. Максимальная и минимальная среднесуточная температура выражена в градусах Цельсия и для i-й даты измеряется в период с 00:00 по 23:59 текущего дня. Среднесуточная скорость ветра измеряется в метрах в секунду. Направление самого сильного ветра измеряется в градусах. Высота над уровнем моря является так же значимым критерием и измеряется в метрах. Данные в датасете представлены за период с первого января 2011 года по тридцать первое декабря 2018 года.

Идея модели довольна проста для понимания. Суть модели в том, чтобы взять данные по всем городам из выборки за i-й день и предсказать минимальную и максимальную температуру на i+1 день для всех городов из выборки. Таким образом, мы сможем узнать существуют ли зависимости между значениями температуры среди городов и зависит ли температура воздуха от таких факторов как скорость ветра, направление ветра и высота метеорологической станции над уровнем моря. Мы выясним возможно ли предсказать погоду по этим данным, и если возможно, то с какой точностью.

Таким образом в данной главе мы познакомились с существующими моделями для прогнозирования погоды, рассмотрели данные, которые мы будем использовать, а также описали идею модели, которую мы используем.

**Глава 2: Нейронные сети для задач прогнозирования погоды**

В данной главе будет введено понятие нейронной сети; понятие полносвязной нейронной сети и ее описание; понятие функции активации и несколько необходимых в дальнейшем функций; понятие нормализации. А также будут описаны метрики, которые будут использованы.

Искусственная нейронная сеть — это математическая модель, а также ее программные или аппаратные реализации, построенная в некотором смысле по образу и подобию сетей нервных клеток живого организма. Нейронные сети применяются для задач прогнозирования, распознавания образов, классификации объектов, сжатия информации, восстановления и зашумления данных. Пример нейронной сети можно увидеть ниже (рис. 1).

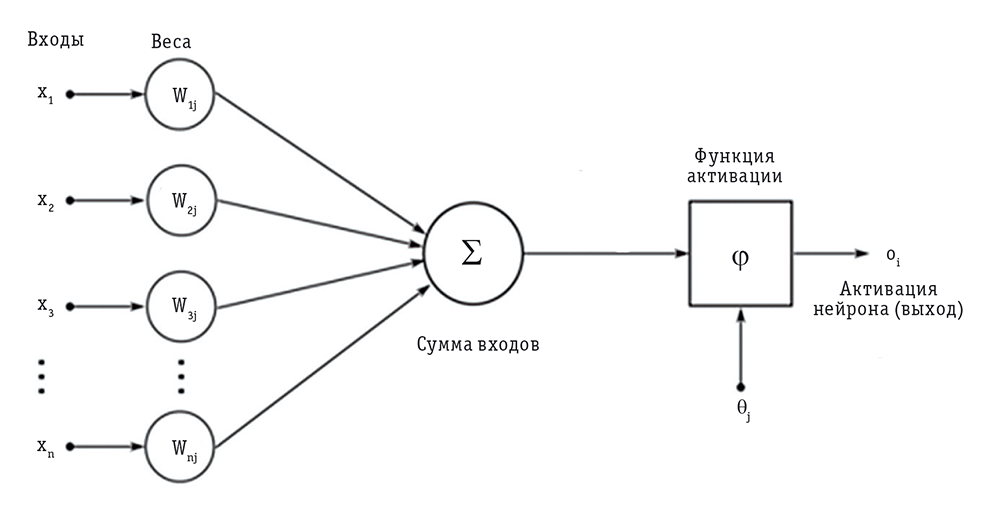


Рис. 1. Пример архитектур нейронной сети.

Способность находить зависимости между входными и выходными данными является характерной особенностью нейронной сети.

Искусственная нейронная сеть состоит из слоев. Каждый слой состоит из нейронов и функции активации. Искусственный нейрон — это узел [искусственной нейронной сети](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C), являющийся упрощённой моделью [естественного нейрона](https://ru.wikipedia.org/wiki/%D0%9D%D0%B5%D0%B9%D1%80%D0%BE%D0%BD). Математически, искусственный нейрон обычно представляют как некоторую нелинейную функцию от единственного аргумента — [линейной комбинации](https://ru.wikipedia.org/wiki/%D0%9B%D0%B8%D0%BD%D0%B5%D0%B9%D0%BD%D0%B0%D1%8F_%D0%BA%D0%BE%D0%BC%D0%B1%D0%B8%D0%BD%D0%B0%D1%86%D0%B8%D1%8F) всех входных сигналов, функцией активации. Входным сигналом же является результат функции активации, умноженный на вес. Интуитивно можно представить, что вес – это уровень значимости нейрона на предыдущем слое с нейроном в текущем слое. Чем больше вес связи между двумя нейронами, тем они больше влияют на конечный результат. Полученный результат посылается на единственный выход. Такие искусственные нейроны объединяют в сети — соединяют выходы одних нейронов с входами других. Таким образом и получается нейронная сеть. Если посмотреть на рис. 2, будет понятно, что слоев в такой сети может быть крайне много, а количество нейронов может исчисляться десятками тысяч.

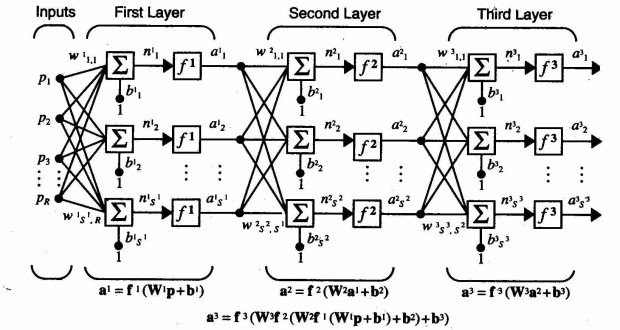


Рис 2. Полносвязная нейронная сеть.

Однако, говоря о нейронных сетях стоит упомянуть о полносвязных нейронных сетях (рис. 2). Это сеть в которой все нейроны предыдущего слоя соединены связью с нейронами текущего слоя. Полносвязные нейронные сети используются сейчас наиболее часто. Полносвязная нейронная сеть может моделировать функцию практически любой степени сложности, причем число слоев и число элементов в каждом слое определяют сложность функции. Определение числа промежуточных слоев и числа элементов в них является важным вопросом при конструировании. Можно описать эту зависимость так: чем больше нейронов в слое, тем большее количество признаков рассматривает сеть в данном слое; чем больше слоев, тем глубже она рассматривает эти признаки.

Остановимся поподробнее на функции активации. В [искусственных нейронных сетях](https://ru.wikipedia.org/wiki/%D0%98%D1%81%D0%BA%D1%83%D1%81%D1%81%D1%82%D0%B2%D0%B5%D0%BD%D0%BD%D0%B0%D1%8F_%D0%BD%D0%B5%D0%B9%D1%80%D0%BE%D0%BD%D0%BD%D0%B0%D1%8F_%D1%81%D0%B5%D1%82%D1%8C) функция активации нейрона определяет его выходной сигнал. В качестве аргумента ей передается сумма выходных значений нейронов предыдущего слоя. Есть несколько требований для функции активации: нелинейность и непрерывная дифференцируемость. Функция активации нужна чтобы ввести нелинейность в математическую модель. Приведем пример нескольких функций активации, которые были использованы в работе.

1. ReLU (рис. 3)

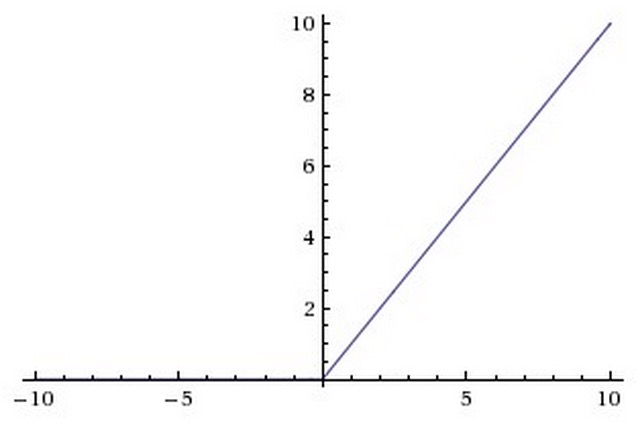


Рис. 3. График функции активации ReLU

Формула для данной функции: f(x) = max(0, x).

1. Sigmoid (рис. 4)

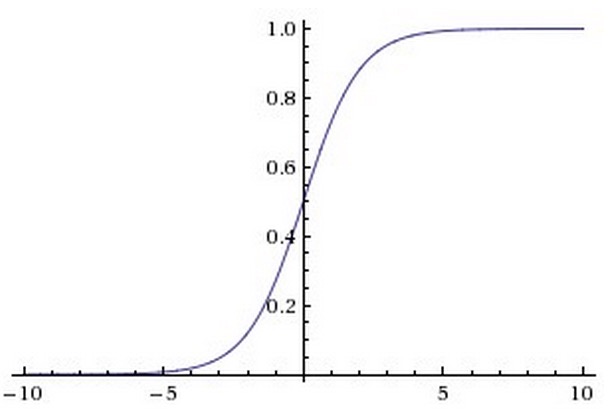


Рис. 4. Функция активации sigmoid.

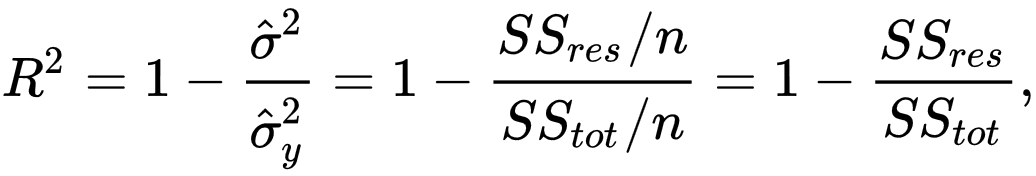
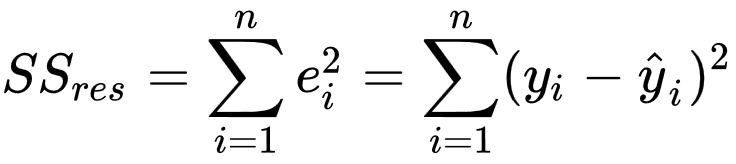
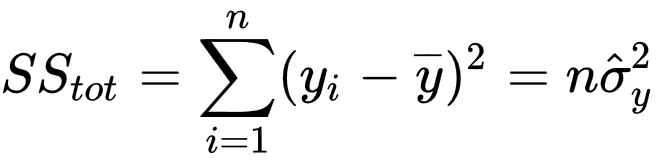
Формула: σ(x) = 1 / (1 + e-x). Область значения функции: (0;1).

1. Softmax (рис. 5)

Формула: . Область значения функции: (0;1).

Для обучения нейронной сети используется метод градиентного спуска. Метод призван уменьшить функцию потерь, изменяя веса в нейронной сети. В нашем случае в качестве функции потерь мы используем Mean Absolute Error: f(x,y) =, где y– вектор реальных значений, а x – вектор предсказанных значений. Я выбрал именно эту функцию, потому что в отличие от Mean Squared Error она не возводит разность (y-x) в квадрат. Возведение в квадрат хорошо подходит если функция потерь может принимать значение больше единицы. В таком случае при возведении в квадрат число становится еще больше, и нейронная сеть старается уменьшить это значение еще интенсивнее. В нашем же случае из-за того, что данные имеют разные единицы измерения, нам пришлось использовать нормализацию. Поэтому значение разности не может превосходить единицу, а значит при возведении в квадрат давало бы уменьшение функции потерь. Именно по этой причине Mean Absolute Error подходит лучше.

Как мы уже сказали, нейронная сеть чувствительна к входным данным. Поэтому нужно данные привести в нормализованную форму. То есть к стандартному отклонению 1 и математическому ожиданию 0. Это сделано для того, чтобы обучение нейронной сети шло качественнее. В качестве функции нормализации возьмем следующую: . ­­

Для оценки качества прогнозирования был использован коэффициент R2: где - сумма квадратов разностей реального и предсказанного значений;  - сумма квадратов разностей реальных значений и среднего. Таким образом, мы можем понять на сколько наш прогноз лучше, чем если бы мы просто брали среднее значение по выборке. Коэффициент принимает значения от [-∞;1]. Чем меньше квадрат разности между реальным и предсказанным значением, тем коэффициент ближе к 1. Если коэффициент = 0, то вместо нашего предсказания можно брать среднее значение. Если R2 < 0, то прогноз хуже, чем если бы мы брали среднее значение.

Таким образом, в данной главе было введено понятие нейронной сети, и ее описание; понятие функции активации и несколько примеров функций активации; понятие нормализации и приведен пример функции нормализации; понятие нормализации. А также были описаны метрики, которые будут использованы для оценки качества сети.

**Глава 3: Результаты исследований**

**3.1: Описание датасета**

Датасет был получен с сайта <https://www.ncdc.noaa.gov>. В нем содержатся данные о 12-ти городах, отмеченных на карте (рис.5).

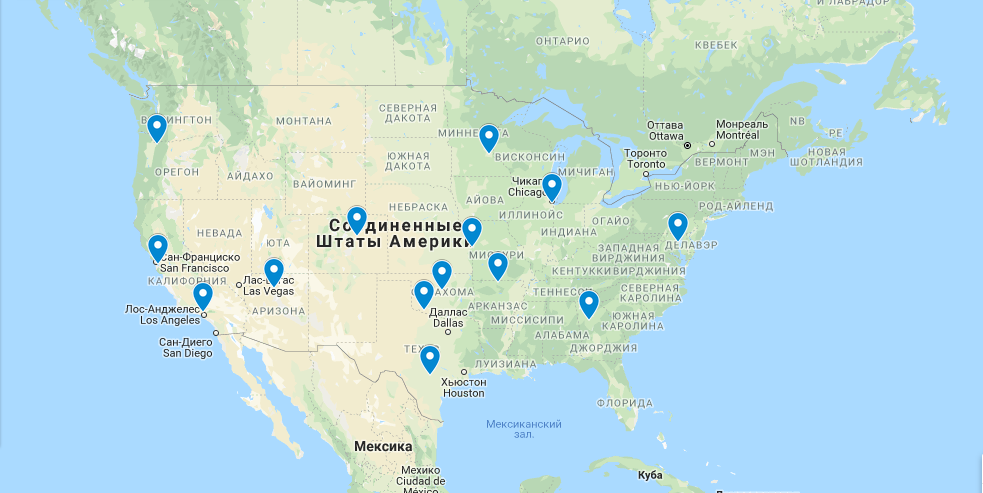


Рис. 5. Карта выбранных городов.

Города были выбраны не случайно. В каждом городе находится крупный аэропорт, на котором располагается метеорологическая станция. Так как для воздушных перевозок крайне важны показатели погоды, мы можем воспользоваться этим, чтобы получить хорошие данные. Таким образом удалось почти полностью избежать пропусков в данных. По каждому из аэропортов мы имеем 5 метрик: средняя скорость ветра, направление ветра, высота станции над уровнем моря, максимальная и минимальная дневная температура. Всего было собрано таких данных за 2922 дня. Как уже говорилось, в данных крайне мало пропусков. Однако, они все же присутствуют. Дабы исключить пропуски, мы заменяем их средним значением по признаку. Так как пропусков < 1%, то на обучение модели это никак не повлияет.

**3.2: Эксперименты**

В этой главе будет выбрана оптимальная структура нейронной сети; проведено сравнение нескольких архитектур; проведены эксперименты с использованием различных данных.

Для начала выберем архитектуру нейронной сети. Количество нейронов и слоев крайне сильно влияет на результат. Если сделать количество слоев и нейронов в слоях слишком большим, то, во-первых, сеть будет медленно обучаться, а во-вторых, она попросту запомнит правильные ответы вместо того, чтобы найти шаблоны. Проще говоря, переобучится. Однако, если же сделать малое количество слоев и нейронов в слоях, то сеть не сможет детектировать шаблоны. т. к. не сможет хранить достаточно информации для этого.

В данной работе использовались различные архитектуры, подбиралось разное количество слоев, нейронов в каждом слое и функций активации. Такие параметры обучения как количество эпох, batch size и optimizer не изменялись. В таблице 1 ниже можно увидеть примеры лучших архитектур.

Таблица 1.

|  |  |  |
| --- | --- | --- |
| **Описание модели** | **R2 на тренировочной выборке** | **R2 на тестовой выборке** |
| dense\_1 (None,100)  ReLu    dense\_2 (None, 100)  Softmax  batchNormalization    dense\_3 (None, 50)  ReLu    dense\_4 (None, 30) | 0.9209 | 0.7606 |
| conv1d\_1 (None, 71, 20)  max\_pooling1d\_1  conv1d\_2 (None, 16, 10)  max\_pooling1d\_2  conv1d\_3(None, 7, 10)  max\_pooling1d\_3  flatten\_1  dense\_1 (None, 30) | 0.9057 | 0.7270 |
| conv1d\_1 (None, 71, 30)  average\_pooling1d\_1  conv1d\_2 (None, 16, 20)  max\_pooling1d\_2  conv1d\_3(None, 7, 10)  max\_pooling1d\_3  flatten\_1  dense\_1 (None, 30) | 0.9065 | 0.7270 |
| dense\_1 (None,100)  ReLu  dense\_2 (None, 80)  Softmax  dense\_3 (None,70)  ReLu  dense\_4 (None,60)  batchNormalization    dense\_5 (None, 40)  ReLu  dense\_6 (None, 40)  batchNormalization    dense\_7 (None, 30) | 0.9148 | 0.7508 |

Из таблицы видно, что все сети показывают примерно одинаковую точность как на тренировочной, так и на тестовой выборке. Возьмем за основу архитектуру с 3-мя полносвязными слоями, т. к. она показала наибольшее значение R2.

Теперь попробуем поменять количество нейронов в каждом слое. В таблице 2 приведены результаты экспериментов. И будем тренировать сеть на 200 эпохах.

Таблица 2

|  |  |  |
| --- | --- | --- |
| **Описание модели** | **R2 на тренировочной выборке** | **R2 на тестовой выборке** |
| dense\_1 (None,100)  ReLu    dense\_2 (None, 100)  Softmax  batchNormalization    dense\_3 (None, 50)  ReLu    dense\_4 (None, 30) | 0.9305 | 0.7664 |
| dense\_1 (None,60)  ReLu    dense\_2 (None, 120)  Softmax  batchNormalization    dense\_3 (None,80)  ReLu    dense\_4 (None, 30) | 0.9296 | 0.7588 |
| dense\_1 (None,60)  ReLu    dense\_2 (None, 120)  Softmax  batchNormalization    dense\_3 (None,80)  ReLu    dense\_4 (None, 30) | 0.9302 | 0.7743 |
| dense\_1 (None,70)  ReLu    dense\_2 (None, 80)  Softmax  batchNormalization    dense\_3 (None, 90)  ReLu    dense\_4 (None, 30) | 0.9335 | 0.7827 |

И так, из таблицы видно, что последний набор параметров дает наилучший показатель точности.

Далее пихаем в сетку еще и центральные города, но т.к. они далеко, они могут ухудшить наш скор.

Применяем лучшую сетку, смотрим скор, строим график score - эпоха.

Заключение:

Мы рассмотрели задачу исследование возможности прогнозирования погоды с помощью нейронных сетей, применили несколько архитектур нейронок, попробовали разные слои на реальном датасете.

Таким образом все задачи выполнены и цель курсовой работы достигнута.

Дальнейшая работа: улучшение модели, выявление новых признаков, увеличение точности, расширение зоны предсказания.

Список литературы: 8 ссылок

Приложение: код (лучшей нейронки)