УДК 004.021:004.75

**Нейронные сети в BIG DATa**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
| ***В.Д. Владымцев***  *Бакалавр инженер системный-программист, ассистент кафедры информатики* | ***М.В. Ганусевич***  *Cтудент 4 курса, кафедра информатики,*  *БГУИР* | ***М.В. Демещенко***  *Cтудент 1 курса,кафедра информатики,*  *БГУИР* | ***А. Н. Марков***  *Старший преподаватель,*  *магистр технических*  *наук, заместитель*  *начальника Центра*  *информатизации и*  *инновационных*  *разработок БГУИР* | ***Р.С. Марковец***  *Cтудент 1 курса, кафедра информатики, БГУИР* | ***Т.А. Сугако***  *Cтудент 1 курса, кафедра информатики, БГУИР* |

***В.Д. Владымцев***

*Бакалавр инженер системный-программист, ассистент кафедры информатики.*

***М.В. Ганусевич***

*Студент 4 курса специальности “Информатика и технологии программирования” БГУИР.*

***А.Н. Марков***

*Магистр технических наук, старший преподаватель кафедры ПИКС, заместитель начальника*

*Центра информатизации и инновационных разработок Белорусского государственного университета*

*информатики и радиоэлектроники*

***М.В. Демещенко***

*Студент 1 курса специальности “Информатика и технологии программирования” БГУИР.*

***Р.С. Марковец***

*Студент 1 курса специальности “Информатика и технологии программирования” БГУИР.*

***Т.А. Сугако***

*Студент 1 курса специальности “Информатика и технологии программирования” БГУИР.*

**Аннотация.**

**Ключевые слова:** Big data,нейронные сети, функция активации, сигмоидная функция активации, ReLU. ELU

**Введение.**

Под термином Big data понимается объем, скорость и разнообразие данных, с которыми необходимо работать. Отличительной чертой Big Data от обычных данных является невозможность результативного использования традиционных методов обработки данных, таких как реляционные базы данных и обычные статистические инструменты.

Для эффективной работы с большими данными необходимо использовать специальные технологии и инструменты, которые обеспечивают возможность обработки, хранения и анализа данных. Одним из таких инструментов являются нейронные сети.

Нейронная сеть -это математическая модель, которая использует структуру, аналогичную нейронам в человеческом мозге, для решения задач обработки информации. Нейронная сеть состоит из множества связанных между собой нейронов, которые обрабатывают информацию и передают ее дальше по сети.

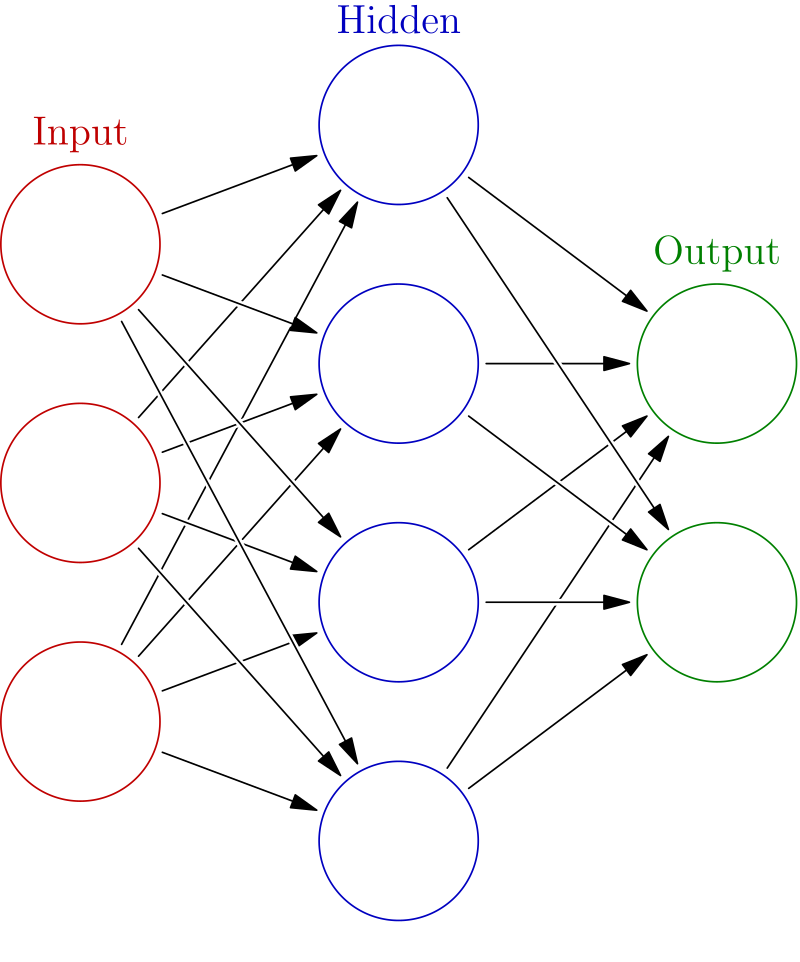


Рисунок 1. Схема полносвязной нейронной сети

Для обучения нейронной сети и получения точных результатов, ей требуется много входных данных. Чем больше данных используется для обучения сети, тем более точные и надежные результаты можно получить в процессе работы сети.

**Актуальность.**

В настоящее время, с увеличением объемов и разнообразия данных, возникает потребность в разработке новых методов и инструментов для обработки и анализа таких данных. В связи с этим, использование нейронных сетей является одним из самых перспективных подходов для работы с большими данными.

Нейронные сети позволяют автоматически обнаруживать скрытые зависимости и закономерности в данных, а также делать точные прогнозы и предсказания. Благодаря возможности обучения на больших объемах данных, нейронные сети могут достигать высокой точности и производительности в решении различных задач, таких как распознавание образов, анализ текстов, предсказание временных рядов и т.д.

**Подготовка данных.**

Изображение 2 иллюстрирует цикл работы модели нейронной сети, который включает в себя 5 шагов. Один из ключевых этапов цикла - подготовка данных. Он состоит из двух шагов, которые идут после определения и постановки задачи. Каждый из этих шагов имеет свои подэтапы, которые необходимо выполнить для того, чтобы гарантировать высокое качество данных и, следовательно, более высокую точность работы модели. Важно уделить должное внимание этапу подготовки данных, так как этот выбор существенно влияет на качество и эффективность работы нейронной сети.

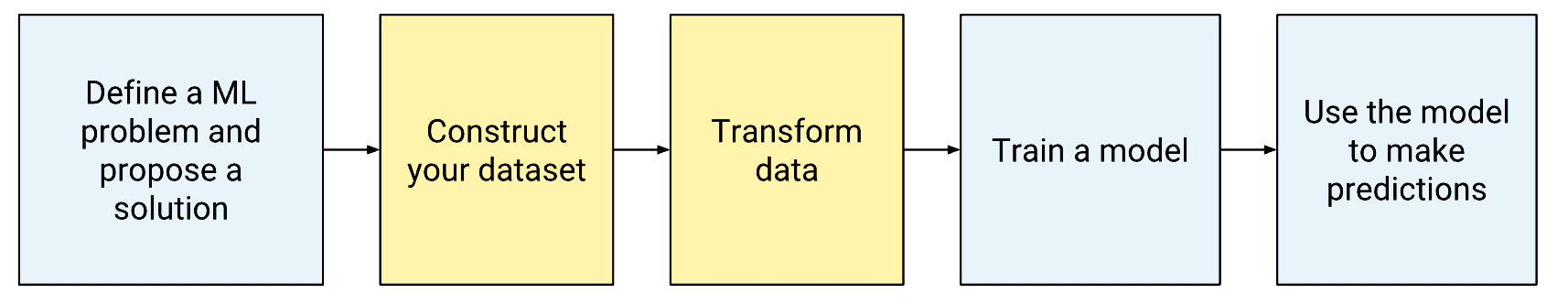


Рисунок 2. Цикл работы модели нейронной сети

**Конструирование датасета.**

Создание набора данных - важнейший шаг в процессе машинного обучения. Прежде чем приступить к преобразованию данных, важно тщательно собрать исходные данные, которые будут использоваться для обучения и тестирования модели. Следующий шаг - определение признаков и маркеров, которые будут использоваться для создания набора данных. В зависимости от проблемы, которую вы пытаетесь решить, вам может понадобиться извлечь признаки из различных источников, таких как изображения, текст или числовые данные.

После определения признаков и маркеров необходимо выбрать стратегию выборки, которая обеспечит репрезентативность набора данных для проблемной области. Это включает определение размера набора данных, распределение примеров по различным классам или категориям, а также баланс между положительными и отрицательными примерами.

Размерность набора данных, необходимая для достижения устойчивой и точной модели, может быть определена через нестрогое правило, основанное на отношении между числом обучаемых параметров модели и размерностью набора данных. В соответствии с этим правилом для достижения оптимальных результатов необходимо использовать объем данных, превышающий число обучаемых параметров как минимум на порядок. Таким образом, большой объем данных позволяет модели учесть большее количество возможных вариантов и сценариев, что в свою очередь повышает точность и устойчивость модели при решении задач машинного обучения.

Наконец, важно разделить набор данных на обучающий и тестовый. Это позволит вам оценить производительность модели на данных, которые она не видела во время обучения, и выявить проблемы с переподгонкой (overfitting) или недоподгонкой (underfitting). Размер обучающего и тестового наборов, а также соотношение между ними зависит от размера вашего набора данных и сложности вашей модели. Самые часто используемые разделения включают в себя отношения вида:

Таблица 1. Отношение обучающего набора к тестовому

|  |  |
| --- | --- |
| Обучающий набор | Тестовый набор |
| 80% | 20% |
| 67% | 33% |
| 50% | 50% |

**Качество данных.**

В машинном обучении качество данных, используемых для обучения, является критическим фактором, который может существенно повлиять на производительность полученной модели. Качественный набор данных - это тот, который позволяет успешно решить поставленную бизнес-задачу. Надежность, представление признаков и минимизация перекоса - это аспекты качества данных, которые обычно ассоциируются с более эффективными моделями.

Надежность относится к достоверности набора данных, что может оказать значительное влияние на точность прогнозов. Чтобы определить надежность набора данных, необходимо оценить ошибки маркеров, зашумленные характеристики (noisy features) и надлежащую фильтрацию для конкретной решаемой задачи. Наиболее распространенными источниками ненадежных данных являются пропущенные значения, повторяющиеся примеры, плохие маркеры и плохие значения признаков.

Представление признаков - это преобразование данных в полезные признаки, что имеет решающее значение для успеха моделей машинного обучения. Чтобы обеспечить оптимальное представление признаков, необходимо учитывать способ представления данных модели, нормализацию числовых значений и обработку выбросов.

Перекос в обучении является серьезной проблемой в машинном обучении, когда в автономных экспериментах могут быть получены разные результаты по сравнению с экспериментами в реальном времени. Эта проблема может возникнуть из-за несоответствия между данными, используемыми при обучении, и данными, доступными при обслуживании. Для уменьшения перекоса между обучением и обслуживанием при обучении следует использовать только те характеристики, которые доступны во время обслуживания, а набор данных для обучения должен быть репрезентативным для обслуживающего трафика.

**Модификация данных.**

**Заключение.**

**Список литературы**

[1]

**ACTIVATION FUNCTIONS**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| ***V. D. Vladymtsev*** *Bechalor engineer system-programmer, assistant of professor computer science* | ***M.V. Hanusevich***  *Pregraduate student of the BSUIR* | ***M.V. Demeschenko***  *Pregraduate student of the BSUIR* | ***A. N. Markov***  *Senior lecturer of the department, Deputy head of the Center for Informatization and Innovative Developments* | ***R.S. Markovets***  *Pregraduate student of the BSUIR* | ***T.A. Sugako***  *Pregraduate student of the BSUIR* |

*Department of Information and Computer Systems Design*

*Faculty of Computer Systems and Networks*

*Belarusian State University of computer science and Radio Electronics, Republic of Belarus*

*E-mail:*

**Abstract.**

**Keywords**: