Применимость ИИ для прогноза популярности книги по начальным логам чтения

исследовательский отчёт

Исполнитель: Боровинский Арсен Исаевич

ИТ-университет

Пермь, 2019 г.

Мотивация

В системе образования распространено мнение об применении методов искуственного интеллекта (ИИ) в образовательном процессе для индивидуализации образовательной траектории. Предполагается, что по трекам (логам: событиям, генерируемым образовательными ресурсами) можно будет определить оптимальную сложность последующих ресурсов (которые необходимо предложить учащемуся) или может быть произведен возврат на изученные и забытые темы, по которым совершаются ошибки. В качестве одной из подзадач возникает необходимость ранжирования ресурсов (какие лучше, а какие хуже), причём в учебную среду постоянно будут попадать новые ресурсы, которые также должны быть отранжированы. Усложняет задачу существование большого количества учебных приложений, по которым детальная информация о прохождении не известна (например встраивая в учебную среду ресурсы LearningApps, можно узнать только время взаимодействия пользователя с ресурсом).

Цель проекта: протестировать перспективность применения ИИ для задачи предсказания популярности учебного ресурса в условиях ограниченных знаниях об взаимодействии пользователя с ресурсом.

Методы: обучение нейронной сети для задачи предсказания популярности книг в формате PDF, размещённых в электронной библиотеке вуза по логам чтения книги.

Результаты: получить удовлетворительные предсказания популярности документа с помощью ИИ не удалось.

Методика исследования

Для анализа использованы логи воспроизведения PDF-документов в электронной библиотеке **https://elis.psu.ru**. Логи считаются в плеере документа и содержат:

- h - час открытия документа;
- p - число прочитанных (пролистанных) страниц;
- t - общее время в течении которого документ открыт в браузере (в секундах);
- nid - идентификатор документа;
- s - статус доступа (доступ открыт — 1, доступ ограничен — 0).

Одна сессия работы с документом (одно открытие плеера в браузере) порождает одну запись в логах. Всего в исследовании использован набор из 500 000 логов.

Кроме логов известен рейтинг документа R, считающийся для фиксированного документа за всё время использования документа по числу просмотренных страниц, числу открытий, числу скачиваний.

Для исследования построены три модели нейронной сети.

Тестовое множество во всех моделях составляет 1/10 от обучающего.

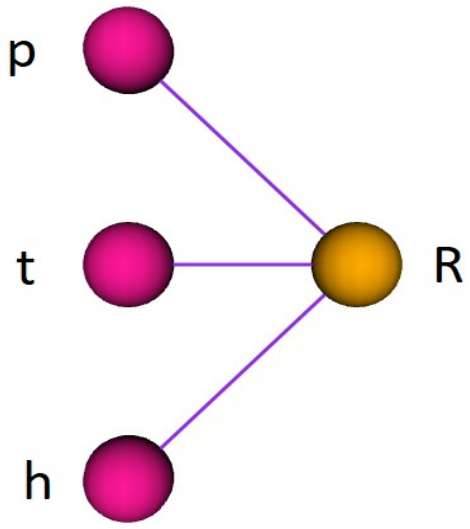
Исследование проводилось с использованием TensorFlow и для первой модели допольнительно на Нейросимуляторе5.

Модель 1

На вход нейронной сети подаются h, p, t (3 нейрона), выход содержит рейтинг R с сигмодной функцией активации. Рейтинг подсчитанный за всё время с загрузки книги.

На вход подавались как все 500 тыс. логов, так и отфильтрованные только при s=1 (только для разрешённых доступов).

Количество скрытых нейронов в одном скрытом слое варьировалось от 0 до 3 с функцией активации ReLU.



Фигура 1: Нейронная сеть для модели 1

Результат модели 1

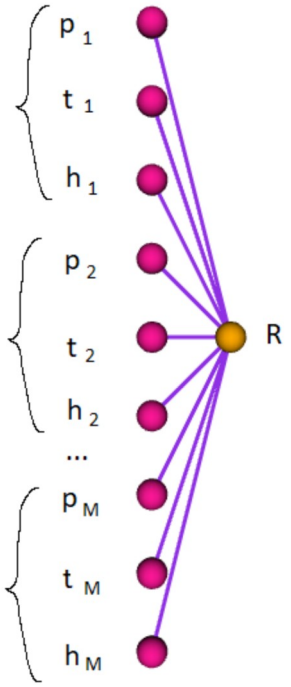
Ошибка обучения за время обучения падает только в три раза. Нейронная сеть запоминает 2 значения R и выдаёт одно из них.

Выводы: сеть не обучилась и использует значения R наиболее читаемых ресурсов.

Модель 2

Модель 1 была модернизирована, чтобы считать рейтинг R по каждому документу в отдельности.

Гипотеза: если взять первые M логов по конкретному документу, то по ним можно предсказать итоговый рейтинг (популярность) документа.



Фигура 2: Нейронная сеть для моделей 2 и 3

Все логи сгруппированы по документам. Для каждого документа с известным nid берётся M первых логов, которые составляют входные нейроны, где M варьируется от 10 до 100. Оставшиеся по данному документу логи отбрасываются. С учётом содержания в каждом логе i данных на 3 нейрона (h_i, p_i, t_i), всего получается 3M входных нейронов (от 30 до 300) и один выходной нейрон R.

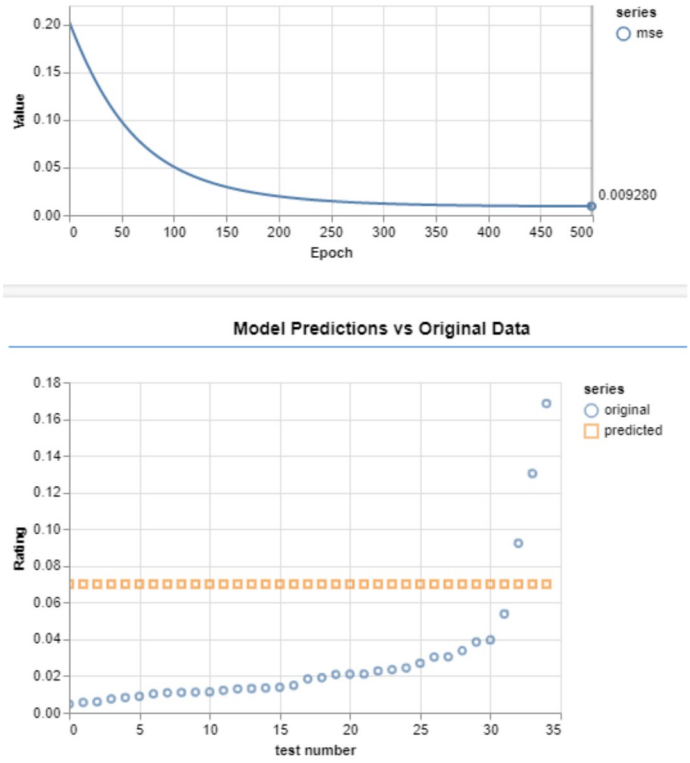
Число скрытых слоев варьировалось от 0 до 2 с числом скрытых нейронов от 1 до 20.

Результат модели 2

При малом числе скрытых нейронов (1-3) повторяется результат модели 1 — нейронная сеть запоминает одно или несколько значений и выдаёт в качестве R одно из них.

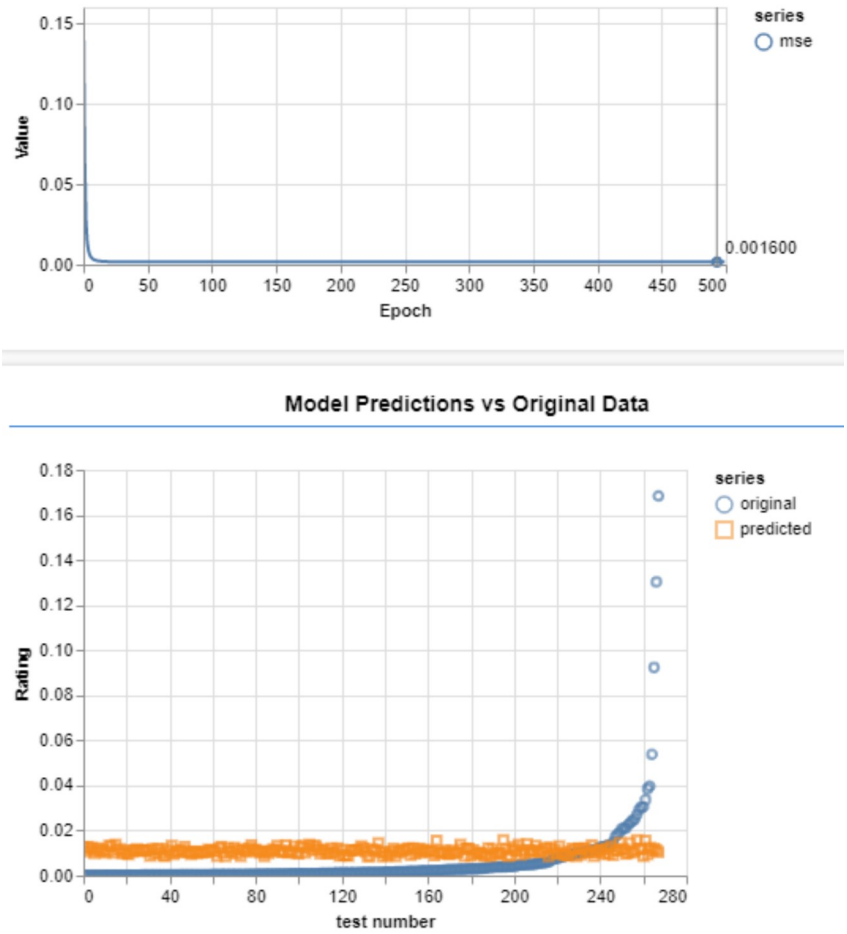
Оптимальная конфигурация сети — 1 скрытый слой с 6-10 скрытыми нейронами и функцией активации ReLU.

При одном скрытом нейроне повторяется результат модели 1: все предсказания имеют один рейтинг R.



Фигура 3: Предсказание модели на тестовом множестве при одном скрытом нейроне и M=100

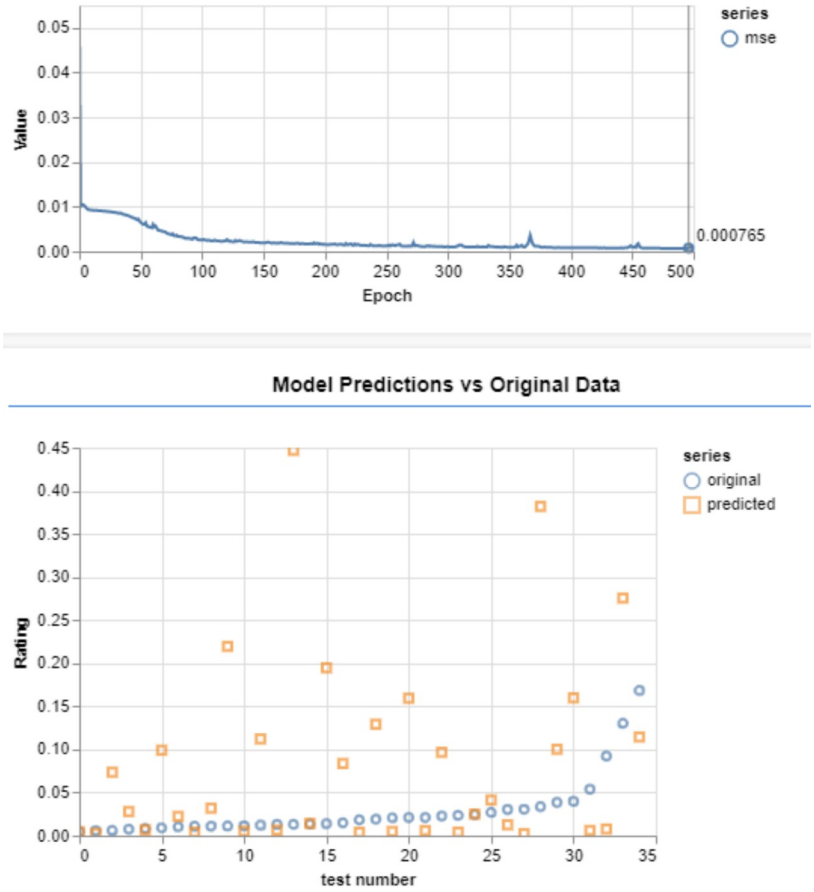
При уменьшении M до M=1 можно получить модификацию модели 1, в которой не происходит повторений обучающих выборок для разных нод. В этом случае, в отличие от модели 1 удаётся получить разнообразные предсказания, но они держатся в районе среднего значения.



Фигура 4: Предсказания для M=1 на модели 2 без скрытого слоя.

При увеличении числа скрытых нейронов в скрытом слое до 6-10 предсказания становятся разнообразны, но предсказательная способность хуже, чем просто всегда предсказывать медианное значение. Оптимальное M от 50 до 100.

Стоит отметить, что кривая рейтинга представляет собой «длинный хвост» продаж (см. https://www.wired.com/2004/10/tail/), характерный и для онлайн-продаж литературы.



Фигура 5: Предсказание рейтинга документа на тестовом множестве при 6 скрытых нейронах

Для улучшения предсказательной способности предпринимались следующие попытки:

- включение в обучающий набор данных логов с отказом в доступе (s=0);
- фильтрация логов по числу просмотренных страниц p>1 и p<21;
- включение в логи только логов за 2018 год с использованием только ресурсов, загруженных до 1 января 2018 года;
- рейтинг R пересчитывался на логах за 2018 год, чтобы уменьшить существующее влияние на рейтинг продолжительности с даты загрузки ресурса.

Хотя некоторые из изменений влияли на результат, существенного улучшения предсказательной способности добиться не удалось.

Модель 3

Модель 3 является модифицированной моделью 2, в обучающее множество для которой входили логи за 2018 год с пересчитанным рейтингом на основе логов только за 2018 год с исключением логов с отказом в доступе и числом просмотренных страниц от 2 до 20 в течении не менее 15 секунд. M=100.

Из набора исключены документы с рейтингом более 0.2 от максимального (т. е. взяты документы с низким рейтингом, чтобы уменьшить возможное влияние небольшого числа документов с высоким рейтингом.

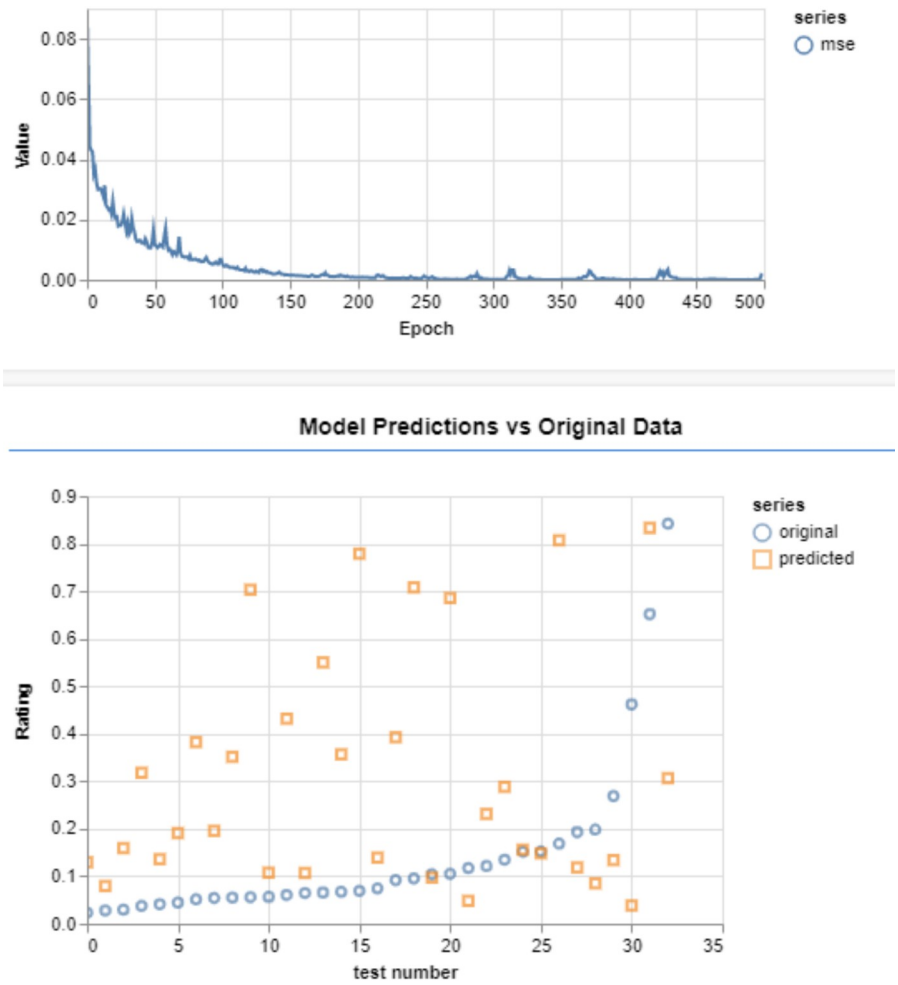
Гипотеза: если взять первых M логов для документа, в которых пользователи точно хоть что-то прочитали и исключить самые популярные документы, то можно будет предсказать итоговый рейтинг документа.

В результате фильтрации обучающая выборка составила 325 наборов входных параметров с отличием от модели 2, что использовались логи только когда было гарантированное взаимодействие с книгой.

Проверка проводилась как со скрытым слоем из 6-10 нейронов, так и без скрытого слоя.

Результаты модели 3

Предсказательная способность модели отсутствует как со скрытым слоем, так и без скрытого слоя. В отличие от модели 1 модели 2, в модели 3 удаление скрытого слоя не приводит к выводу единственного значения, а качество предсказания существенно не уменьшается.



Фигура 6: Презсказательная способность модели 3 для M=100, 1<p<21, t>15 сек.

Иные параметры также не позволили обучить сеть до корректных предсказаний.

Выводы

Гипотезу о возможности предсказания популярности (рейтинга) электронной книги по М актам взаимодействия с книгой подтвердить не удалось.

Для модели 1 решение недостижимо: нейронная сеть запоминает наибольшие значения рейтинга.

Возможные причины, из-за которых не удалось подтвердить гипотезу в моделях 2 и 3:

- Отсутствует взаимосвязь рейтинга с ограниченным набором логов, в которые входят числом просмотренных страниц, время открытия плеера

книги, часом в котором произошло открытие. Популярность книги слабо зависит от того, насколько хорошо книгу читают в рамках отдельных сессий.

- Анализируемые данные не кластеризованы (например смешаны книги для учебы и научные статьи, не разделяются книги на которые пришли пользователи вуза по учебным нуждам и внешние пользователи из интернета). Для кластеризации собираемых данных недостаточно и необходимо обновление системы сбора логов электронной библиотеки.
- Необходим более детальный трекинг, например времени отображения отдельных страниц на экране, число событий ввода при чтении.