Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования «Российский химико-технологический университет имени Д.И. Менделеева»

Факультет цифровых технологий и химического инжиниринга

Кафедра информационных компьютерных технологий

**ОТЧЕТ ПО ПРАКТИЧЕСКОЙ РАБОТЕ № 2**

**ПО КУРСУ**

**«ЦИФРОВОЕ МОДЕЛИРОВАНИЕ ФИЗИКО-ХИМИЧЕСКИХ СИСТЕМ»:**

**«Предсказание свойств материалов с помощью ML»**

Ведущий преподаватель

Ст. преподаватель Скичко Е. А.

**СТУДЕНТ группы КС-26** Лупинос А. В.

**Москва**

**2024**

# **Задание**

Разработать ML-модель для предсказания физико-химического свойства. Число параметров в модели определить из правила. Построить графики зависимости Свойства от Температуры [K] для различных значений P=const [бар], на графиках отметить точками экспериментальные значения из БД, линиями – результаты предсказания модели. Предсказать свойство для указанных значений давления и температуры и отметить на тех же графиках цветными точками. Модель сдается в виде кода и сохраняется в файле .keras.

Вариант №21.

Вещество - этан (жидкость); свойство - плотность, ; - 150 K; - 1 бар.

**Теоретическое обоснование решения**

**Машинное обучение** (англ. machine learning, ML) — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение за счет применения решений множества сходных задач.

Различают **три основных вида**:

1. **Обучение с учителем (supervised learning):**

На вход модели поступают размеченные данные (датасет) – входы и желаемые выходы. Модель сравнивает свой выход с желаемым и меняет свои параметры для улучшения рассогласования (метрика ошибки).

Основная решаемая задача – предсказание (регрессия) и классификация (в заданное число классов).

1. **Обучение без учителя (unsupervised learning):**

Классы не заданы. Желаемого выхода нет.

Основная решаемая задача – кластеризация (группировка данных по заранее не заданным кластерам).

1. **Обучение с подкреплением (reinforcement learning):**

Есть критерий оценки работы в целом (без отдельных примеров). Алгоритм изменяется так, чтобы улучшить критерий.

Пример: автоматическое управление транспортом.

Для решения задач ML часто используют **ИНС** (искусственная нейронная сеть) – набор искусственных нейронов, соединенных синаптическими связями; математическая модель, а также ее 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.

Существуют множество типов ИНС. Самый простой для решения задачи предсказания – **многослойный перцептрон[[1]](#footnote-0)** (класс искусственных нейронных сетей прямого распространения, состоящих как минимум из трех слоев: входного, скрытого и выходного). Нейроны обозначают кругами, которые группируются в слои в направлении прохождения сигналов, а связи – линиями. Стандартный многослойный перцептрон – полносвязанный между соседними слоями.

Обучение нейросети – подбор коэффициентов синаптических связей[[2]](#footnote-1).

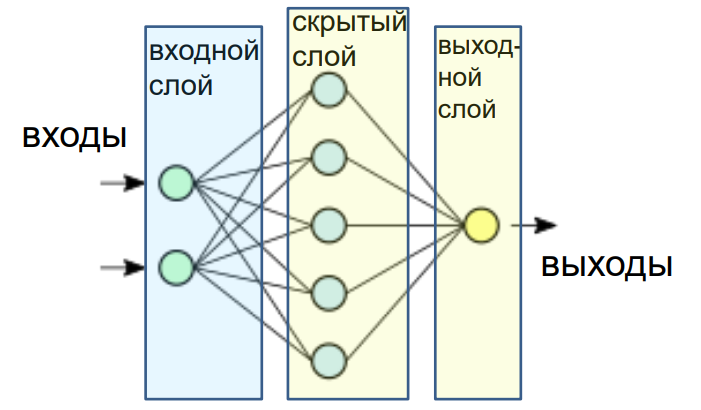


Рис. 1. Логическая схема многослойного перцептрона1.

Этапы решения задачи с помощью ИНС:

1. Сбор датасета;

2. Очистка и нормализация данных;

3. Выбор типа сети;

4. Обучение сети (на обучающей и тестовой выборке);

5. Валидация ИНС (на валидационной выборке).

Правила:

1. Число связей ИНС должно быть меньше числа примеров в обучающей выборке из датасета. Традиционно, в 10 раз.
2. Число нейронов скрытого слоя вычисляется по формуле:

,

иногда работает формула: ,

где i - входной слой, o - выходной слой, = 2 для начала.

**Нормализация -** приведение данных к единому масштабу. Часто может улучшить качество модели. В случае данных, меняющихся на порядки, возможен переход к логарифмической шкале (логарифмирование).

**Обучение нейросети.**

Датасет (строки) делится на две части:

* Обучающая выборка (80%);
* Тестовая выборка (20%).

Обучение состоит из фиксированного числа эпох (например, 1000). На каждой эпохе все данные обучающей выборки по одному образцу (строке) поступают на вход ИНС. Корректируются весовые коэффициенты. Они меняются так, чтобы уменьшить функцию потерь (loss function). Эта функция задается обычно равной некой **метрике качества предсказания** нейросети.

По окончании подачи на вход всех строк повторяется процесс с самого начала (новая эпоха обучения). После окончания обучения рассчитываются метрики качества предсказания ИНС на обучающей и тестовой выборке.

Метрика качества предсказания рассчитывается по каждому выходу модели отдельно. Основные две метрики:

1. **MAPE** (средняя процентная ошибка = средняя относительная ошибка):

,

где - значение выхода из i-ой строки данных; - значение выхода, рассчитанное ИНС, при поступлении на вход i-ой строки данных; 𝑛 – число строк данных в выборке.

1. **MAE** (средняя абсолютная ошибка):

.

Пятикратная **кроссвалидация[[3]](#footnote-2)** – ИНС 5 раз обучается полностью с нуля, причем 20% данных для тестовой выборки берутся все эти 5 раз по-разному.

Пусть задано 1000 эпох обучения. При 5-кратной кроссвалидации всего будет 5000 эпох с разным разбиением данных и 5 разных итоговых вариантов модели. Выбирается наилучшая модель.

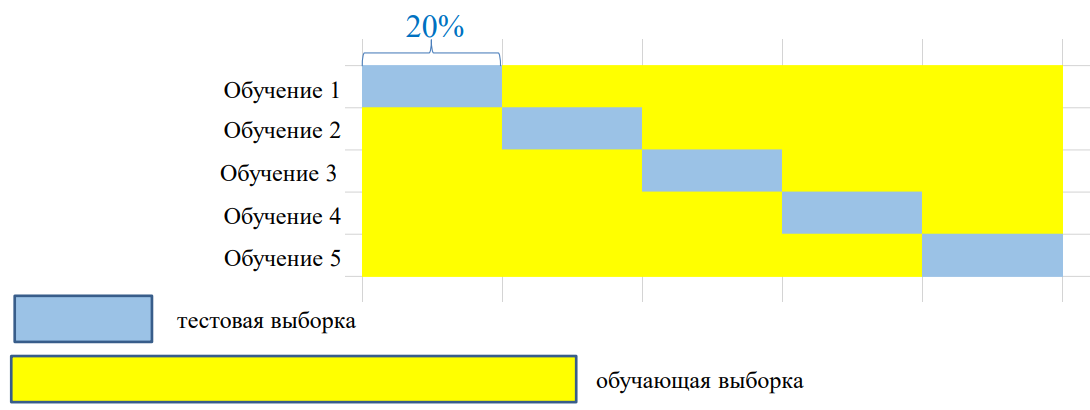


Рис. 2. Схема пятикратной кроссвалидации3.

**Код**

Для данной работы понадобился датасет для обучения модели, датасет для предсказания свойства для указанных значений давления и температуры, а также данные для построения графиков при разных постоянных давлениях (10 бар, 100 бар), которые нужно было перенести в файл с расширением .csv:

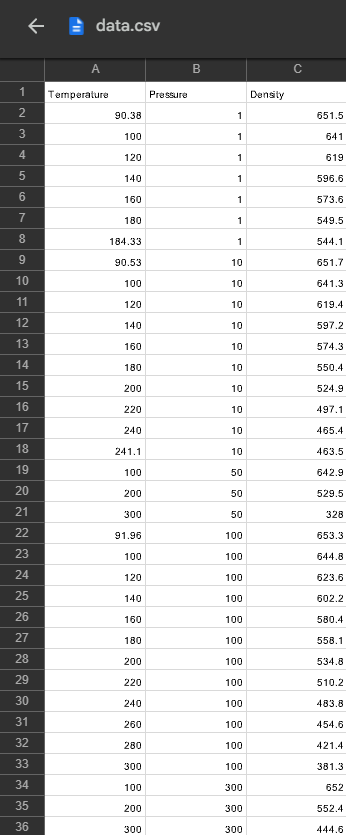


Рис. 3. Датасет для обучения модели.

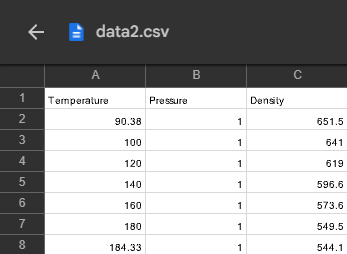


Рис. 4. Датасет для предсказания свойства.

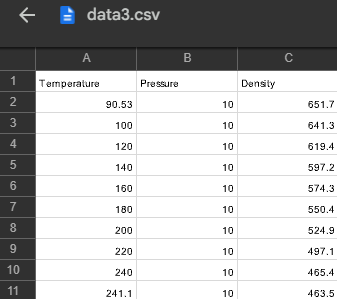


Рис. 5. Данные для построения графика при P = 10 бар.

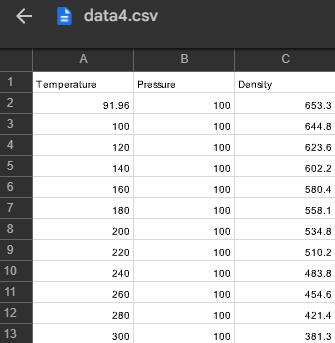
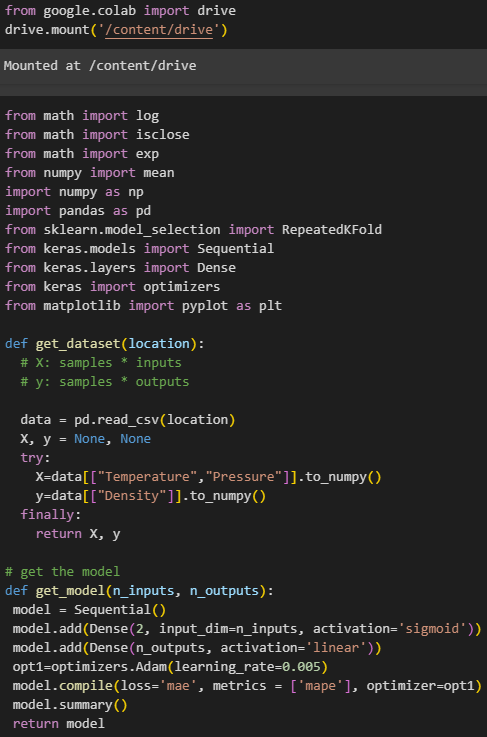
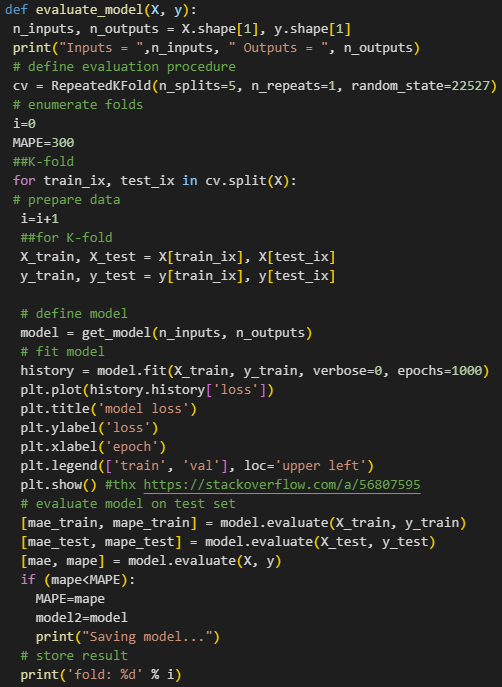
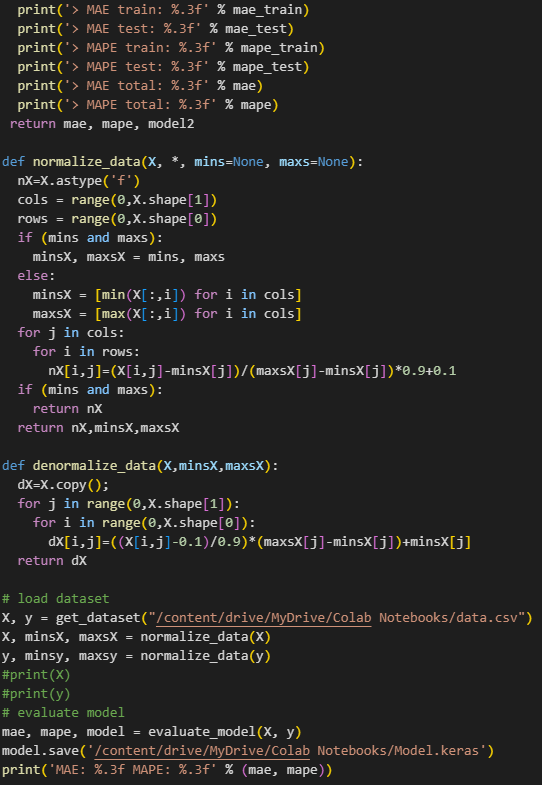


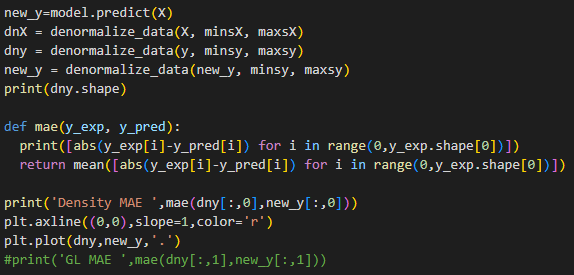
Рис. 6. Данные для построения графика при P = 100 бар.

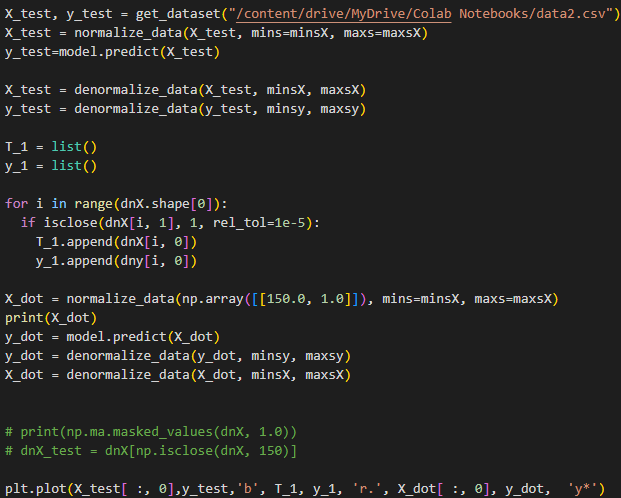
Код для данной практической работы представлен ниже:

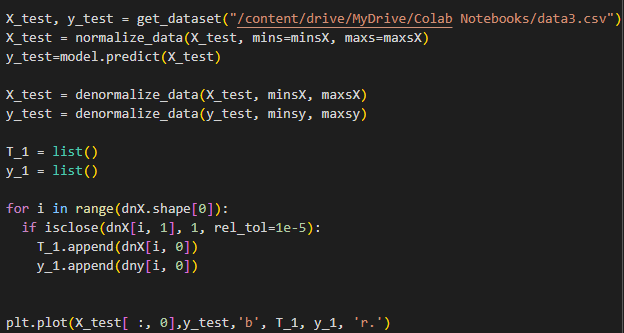


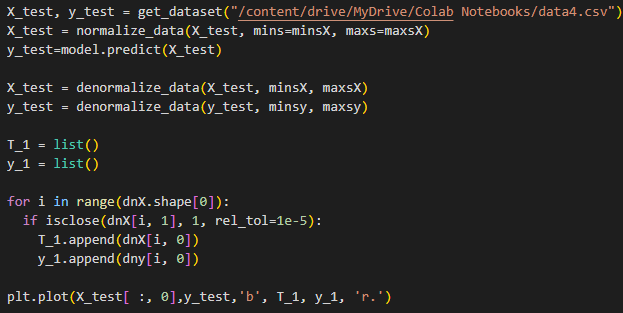






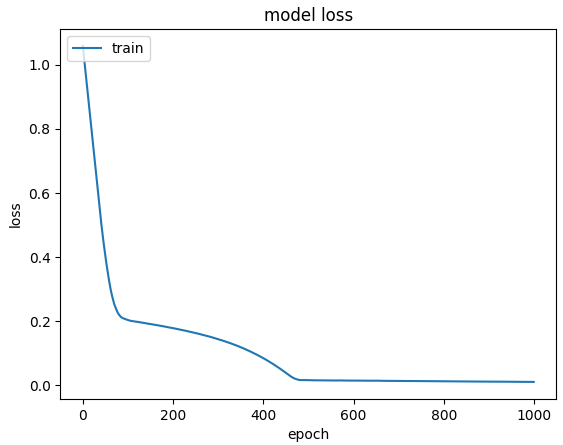


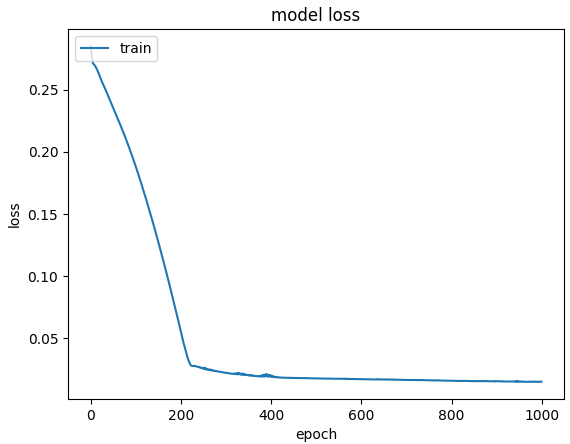


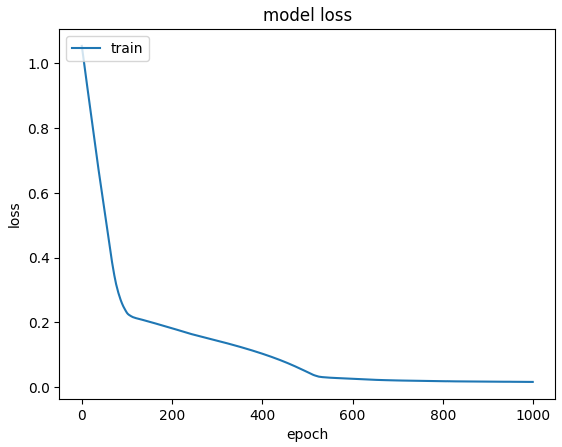
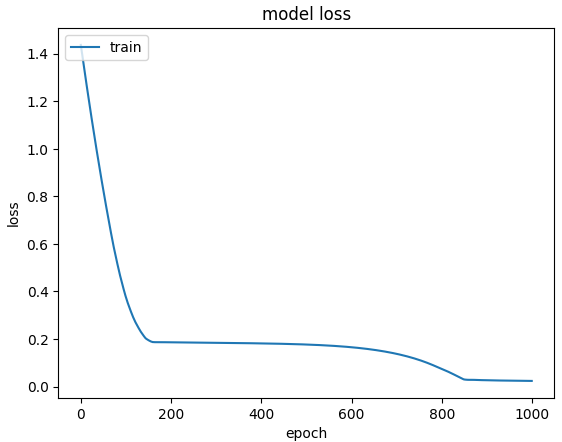
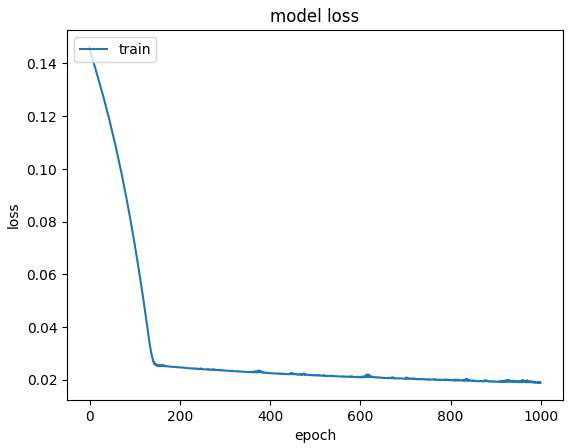


**Результаты расчетов**

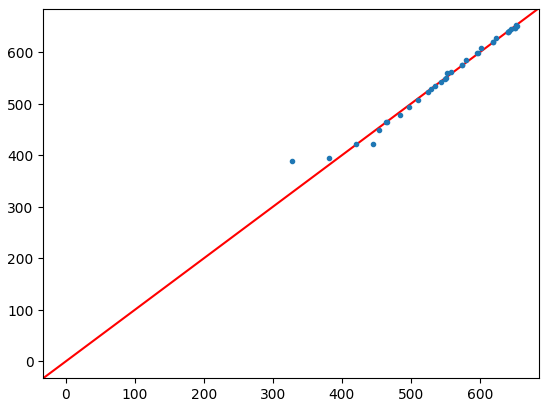
Данные графики отображают функцию потерь, зависимость потерей от эпох обучения ML. Так как была использована пятикратная кроссвалидация, всего таких графиков 5.



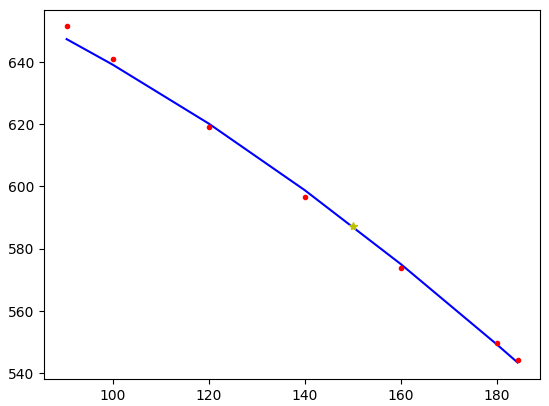




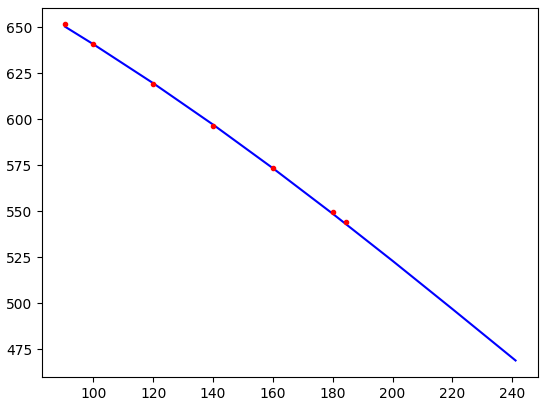
Далее был получен график зависимости плотности жидкого этана от температуры для различных постоянных значений давления с отмеченными экспериментальными значениями из БД и линией, показывающей результаты предсказания модели.

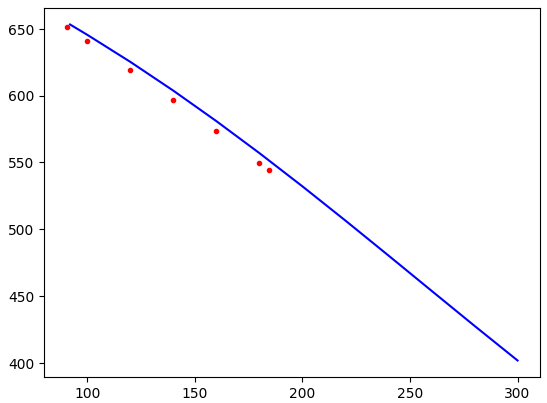


Далее была предсказана плотность жидкого этана для давления, равного 1 бар, и температуры 150 К и отмечена на графике цветной точкой (зеленой).



Также были построены графики зависимости плотности от температуры при P = 10 бар и P = 100 бар соответственно:





1. Перцептрон - математическая или компьютерная модель восприятия информации мозгом (кибернетическая модель мозга). [↑](#footnote-ref-0)
2. Синапс (греч. σύναψις, от συνάπτειν — соединение, связь) — место контакта между двумя нейронами или между нейроном и получающей сигнал эффекторной клеткой. Служит для передачи нервного импульса между двумя клетками, причём в ходе синаптической передачи амплитуда и частота сигнала могут регулироваться. [↑](#footnote-ref-1)
3. Кроссвалидация - метод оценки аналитической модели и ее поведения на независимых данных. При оценке модели имеющиеся в наличии данные разбиваются на k частей. Затем на k−1 частях данных производится обучение модели, а оставшаяся часть данных используется для тестирования. Процедура повторяется k раз; в итоге каждая из k частей данных используется для тестирования. В результате получается оценка эффективности выбранной модели с наиболее равномерным использованием имеющихся данных. [↑](#footnote-ref-2)