

Библиотека машинного обучения для PascalABC.NET: архитектура и использование в образовании

Михалкович С.С.

*Южный федеральный университет,
Институт математики, механики и компьютерных наук
E-mail: miks@sfedu.ru*

Система программирования PascalABC.NET на протяжении многих лет развивается как образовательная платформа, ориентированная на обучение программированию, алгоритмам и современным технологиям [1]–[2]. В последние годы машинное обучение стало важной частью образовательных программ. Однако большинство существующих инструментов (таких как scikit-learn и другие библиотеки анализа данных на языке Python) ориентированы на экосистему с большим числом внешних зависимостей, требующих установки, согласования версий и настройки среды выполнения. В образовательном процессе это создаёт дополнительные сложности, не связанные непосредственно с изучением алгоритмов.

В рамках PascalABC.NET была поставлена задача создания библиотеки машинного обучения, полностью интегрированной в среду PascalABC.NET, не требующей дополнительной настройки и обладающей понятным интерфейсом для студентов и школьников. При этом библиотека разрабатывается без упрощений, характерных для учебных решений, и строится на принципах промышленного программного обеспечения, включая слоистую архитектуру, типобезопасность и эффективность алгоритмов. Библиотека включает встроенные учебные датасеты и систему русскоязычных подсказок.

Таким образом, целью разработки является создание системы, сочетающей доступность для обучения с требованиями к промышленным программным решениям и обеспечивающей изучение методов машинного обучения.

Особенностью проекта является использование методов взаимодействия с большими языковыми моделями. Значительная часть кодовой базы (более 20 тыс. строк) была разработана в процессе итеративного диалога с языковой моделью ChatGPT с последовательным уточнением архитектуры, генерацией и доработкой кода с переносом контекста между сессиями [3]. Такой подход позволил ускорить разработку, сохранить целостность архитектуры и обеспечить быструю реализацию алгоритмов, при этом ключевая роль разработчика сохраняется: он определяет архитектуру и принимает окончательные инженерные решения.

Архитектура библиотеки машинного обучения

Общая архитектура библиотеки представлена на рис. 1.

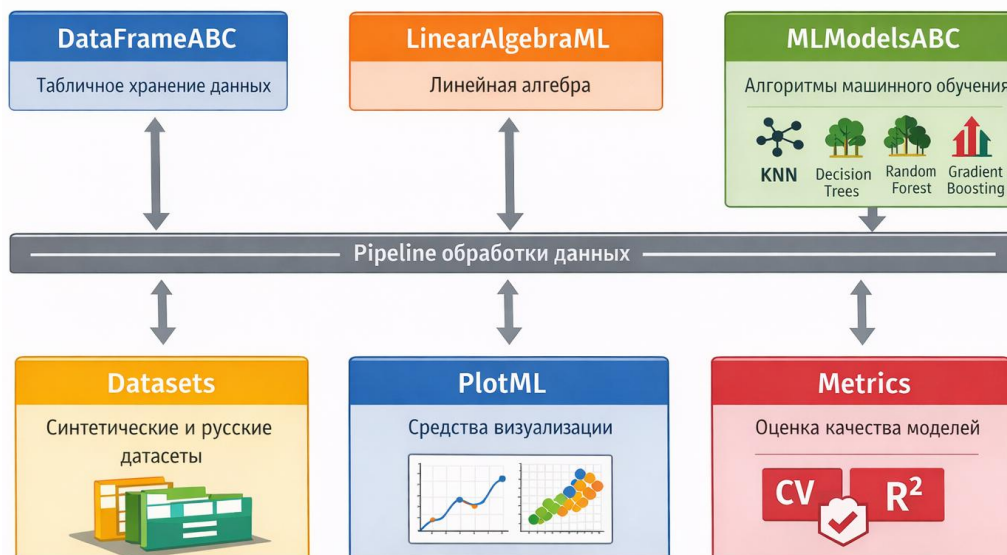


Рис. 1. Архитектура библиотеки ML PascalABC.NET

Система построена как набор модулей с чётким разделением уровней. В её основе лежит DataFrameABC — аналог pandas DataFrame, реализующий табличное хранение данных. В данной реализации используется колоночная (column-store) организация, при которой значения каждого столбца хранятся отдельно, что упрощает обработку и хорошо соответствует задачам машинного обучения. DataFrame используется как универсальный уровень представления данных на всех этапах обработки и обучения моделей.

Численное ядро библиотеки реализовано в модуле LinearAlgebraML. Все модели основаны на операциях над матрицами и векторами, а модуль предоставляет специализированные реализации линейной алгебры, ориентированные на задачи машинного обучения. Это обеспечивает чёткое разделение уровня хранения данных и уровня вычислений и упрощает реализацию алгоритмов.

Алгоритмы машинного обучения реализованы в модуле MLModelsABC и используют единый матричный подход. В библиотеке реализованы линейные модели (LinearRegression, RidgeRegression), логистическая регрессия (LogisticRegression), ElasticNet, методы ближайших соседей (KNN), кластеризационные алгоритмы (KMeans, DBSCAN), деревья решений для классификации и регрессии (DecisionTree), ансамблевые методы RandomForest и GradientBoosting, а также методы отбора признаков (SelectKBest). Поддерживаются задачи классификации, регрессии и кластеризации, что позволяет использовать библиотеку в широком спектре учебных и прикладных задач.

Подготовка данных реализуется с использованием механизма пайплайнов, позволяющего последовательно применять преобразования признаков и объединять их с моделями в единую цепочку обработки. Пайплайн поддерживает два уровня работы: на уровне таблицы (DataFrame) выполняются препроцессоры, такие как заполнение пропущенных значений и кодирование категориальных признаков, после чего данные автоматически преобразуются в матричное представление, где применяются численные трансформации (например, StandardScaler, MinMaxScaler) и модель. Такое разделение позволяет явно разграничить этапы обработки данных и обучения, обеспечить корректный порядок преобразований и использовать единый интерфейс для построения и повторного применения решений.

В библиотеку также включён модуль Datasets, содержащий учебные и синтетические наборы данных с русскоязычными метаданными, предварительно подготовленные для образовательного использования. Для анализа данных реализован модуль PlotML, обеспечивающий базовые средства визуализации и интегрированный непосредственно в среду PascalABC.NET. Отдельный модуль отвечает за вычисление метрик и оценку моделей, включая кросс-валидацию, что позволяет проводить полноценные эксперименты.

Таким образом, библиотека представляет собой целостную систему, в которой данные, численные методы, модели и средства анализа образуют согласованную архитектуру, сохраняя свойства промышленного программного обеспечения, включая модульность, расширяемость и чёткое разделение ответственности.

Примеры использования библиотеки

Библиотека поддерживает полный цикл работы с данными и моделями машинного обучения, включая подготовку данных, обучение и оценку качества. Ниже приведён пример решения задачи регрессии на реальном датасете:

```
var ds := Datasets.MoscowHousing;
var df := ds.Data;
var features := ['rooms', 'area', 'kitchen_area',
'floor', 'floors_total', 'metro_minutes', renovation'];
var target := 'price';

var (trainDf, testDf) := df.TrainTestSplit(0.2, 42);

var pipe := DataPipeline.Build(target, features,
    new OneHotEncoder('renovation'),
    new LinearRegression);
```

```
pipe.Fit(trainDf);

var pred := pipe.Predict(testDf);
var y := testDf.ToVector(target);
Println('R²:', Metrics.R2(y, pred):0:3);
```

В данном примере используется встроенный датасет стоимости жилья. Исходная таблица разбивается на обучающую и тестовую выборки, после чего строится пайплайн, объединяющий преобразование признаков и модель линейной регрессии. Пайплайн представляет собой последовательность компонентов с интерфейсом Fit/Transform/Predict, в которой этапы подготовки данных (например, кодирование категориальных признаков) и обучение модели выполняются согласованно. Такой подход позволяет избежать дублирования кода, обеспечивает воспроизводимость вычислений и делает структуру решения явной. Качество модели оценивается с помощью коэффициента детерминации R^2 .

Для предварительного анализа данных библиотека предоставляет средства визуализации, реализованные в модуле PlotML, в частности построение матриц попарных диаграмм (PairPlot). Пример ниже демонстрирует анализ датасета Iris:

```
var ds := Datasets.Iris;
var df := ds.Data;
var X := df.ToMatrix(ds.Features);
var labels := df.EncodeLabels(ds.Target);
Plot.PairPlot(X, labels, ds.Features);
```

Данные преобразуются в матричное представление, после чего строится матрица попарных диаграмм, позволяющая визуально оценить взаимосвязи признаков и разделимость классов (см. рис. 2).

Такая визуализация особенно полезна в образовательных задачах, поскольку позволяет студентам наглядно понять структуру данных до применения алгоритмов.

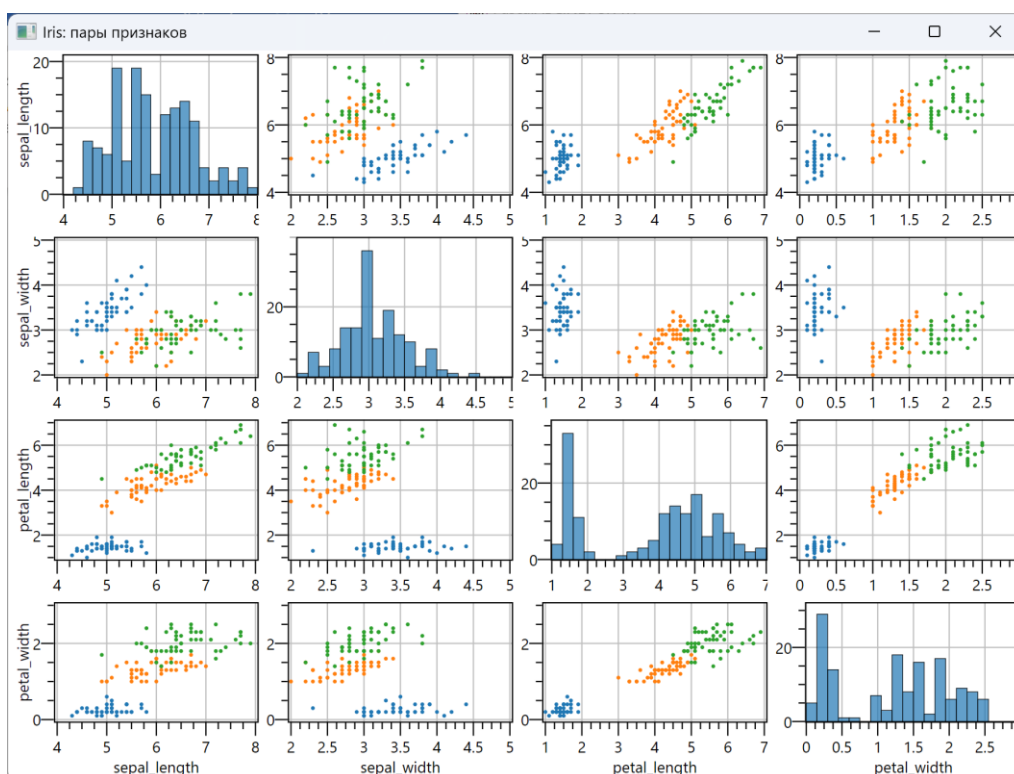


Рис. 2. Диаграмма PairPlot для датасета Iris.

Библиотека ориентирована на использование в образовательном процессе и может применяться в курсах по машинному обучению для студентов, обеспечивая простой интерфейс и прозрачность алгоритмов. Отсутствие внешних зависимостей и полная интеграция в PascalABC.NET упрощают проведение занятий. Она также может использоваться в ИИ-олимпиадах для школьников без дополнительной настройки среды. Наличие встроенных датасетов и средств визуализации создаёт основу для новых типов задач, сочетающих алгоритмическое мышление и анализ данных, а также формирования современной коллекции учебных датасетов, пополняемой в рамках проектной деятельности студентов.

Литература

1. Михалкович С.С. Учебная система программирования Pascal ABC / Современные информационные технологии в образовании: Южный Федеральный округ: Тезисы докладов. Ростов-на-Дону, 2004. С. 156–158.
2. Михалкович С.С. Система программирования PascalABC.NET - 20 лет развития / XXX Научная конференция «Современные информационные технологии: тенденции и перспективы развития». Материалы конференции. Ростов-на-Дону, 2023. С. 287–290.
3. Михалкович С.С. Использование генеративного искусственного интеллекта для системы программирования PascalABC.NET / XXXII Научная конференция «Современные информационные технологии: тенденции и перспективы развития». Материалы конференции. Ростов-на-Дону, 2025. С. 302–306.