ИССЛЕДОВАНИЕ ЭФФЕКТИВНЫХ АЛГОРИТМОВ МАШИННОГО ОБУЧЕНИЯ В ЗАДАЧАХ КЛАССИФИКАЦИИ

Расюк С.А., *студентка группы ИУ6-42Б*
Тюсин Д.Е., *студент группы ИУ6-42Б*
Московский государственный университет им. Н.Э. Баумана

Научный руководитель: ***Егоров В.Г.,***
доктор исторических наук, доктор экономических наук,
профессор кафедры «Информационная аналитика и политические технологии»

Аннотация: в статье рассматриваются эффективные алгоритмы машинного обучения в контексте задач классификации. Машинное обучение представляет собой область искусственного интеллекта, где системы способны самостоятельно учиться на основе опыта и данных. Рассматриваются различные классы, анализ которых не только способствует улучшению точности и качества классификации данных, но и обеспечивают базу для разработки инновационных решений и выбора оптимального алгоритма машинного обучения с учетом специфики данных и требований проекта.

Ключевые слова: машинное обучение, классификация, технологии, алгоритмы, анализ больших данных, извлечение признаков, оценка моделей, нейронные сети, автоматизация процесса обучения.

В современном мире объем данных постоянно растет, поэтому важно иметь совершенные методы их анализа и интерпретации. Машинное обучение предоставляет невероятно мощный и современный инструментарий для разработки и классификации данных в различных областях. Люди активно исследуют новые алгоритмы с целью определения наиболее подходящих для конкретных типов задач.

В контексте исследования эффективных алгоритмов машинного обучения, значимость этой области подтверждается многочисленными работами. Недавние исследования, проведенные Yandex, MLCentre и Skillbox, подчеркивают важность разработки алгоритмов, способных эффективно обрабатывать и классифицировать данные в различных сценариях.

Сегодня информационные технологии окутывают целый мир, они словно проникают во все сферы общественной жизни. Постоянное развитие человечества приводит к увеличению общего числа знаний, применяемых во многих отраслях, таких как медицина, финансы, маркетинг и другие. Классификация объема данных – невероятно важный аспект, позволяющий упорядочить любую информацию. Именно поэтому исследование эффективных алгоритмов машинного обучения становится актуальным и востребованным, что можно проследить на рисунке 1.1.

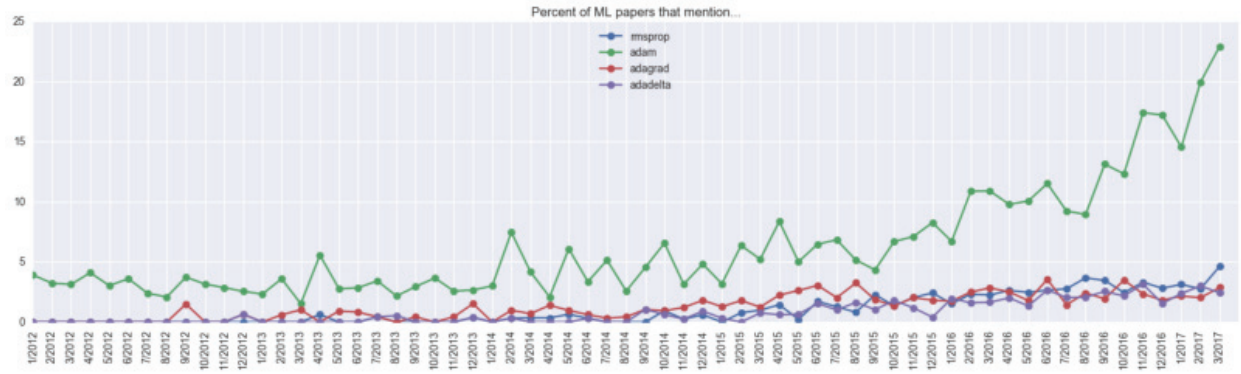

Рис. 1.1. Рост востребованности алгоритмов оптимизации

Машинное обучение (**ML – Machine Learning**) – класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение засчёт применения решений множества сходных задач. [7] Проще говоря, вместо того, чтобы явно задавать правила компьютеру для выполнения задачи, мы предоставляем ему данные и позволяем самому находить закономерности и шаблоны для этих данных. Исходя из этого, разработанная модель стремится разделить данные на заранее определенные «категории». Эти «категории» обычно разбивают на следующие понятия [1]:

1) **Двухклассовая классификация.** Это тип классификации, в котором модель должна разделить данные на два класса. Например, определение, является ли электронное письмо спамом или не спамом, предсказание наличия или отсутствия определенного заболевания у пациента и т. д.

2) **Многоклассовая классификация.** В этом случае модель должна классифицировать данные на более чем два класса. Например, классификация рукописных цифр от 0 до 9, изображений на разные типы животных или классификация распознавания иероглифов.

3) **Непересекающиеся классы.** В этом случае каждый объект данных принадлежит только к одному классу.

4) **Пересекающиеся классы.** Объекты данных могут принадлежать к нескольким классам одновременно. Например, при классификации изображений объект может быть как кошкой, так и собакой одновременно, если на изображении изображены оба животных.

5) **Нечеткие классы.** В некоторых сценариях классы могут быть нечетко определены, и объекты данных могут принадлежать к классам частично или с различной степенью принадлежности.

Для определенных задач выбираются соответствующие классы, подбор которых зависит от характера данных, требуемой точности классификации и конкретных целей анализа. Порой для решения проблемы объект данных может быть классифицирован одновременно в несколько

типов. Для обобщения изложенного вопроса можно использовать математическую запись.

Формальная постановка:

Пусть X – множество описаний объектов; Y – конечное множество номеров (имен, метод) классов. Существует неизвестная целевая зависимость – отображение $□^*: □ \to □$, значения которой известны только на объектах конечной обучающей выборки $□^□ = \{(□_1, □_1), \ldots, (□_□, □_□)\}$. Требуется построить алгоритм $\alpha: □ \to □$, способный классифицировать произвольный объект $□ \in □$.

Выбор наилучшего алгоритма:

Плавно переходя к вопросам алгоритмизации в машинном обучении, задача классификации решается с использованием готовых аналитических моделей, называющихся классификаторами, при помощи которых происходит непосредственная реализация деления на классы.

Некоторые из **наиболее эффективных** алгоритмов, занимающихся вопросами классификации [4, 6]:

- Логистическая регрессия (LogisticRegression)
- Дерево решений (DecisionTree)
- K-Ближайшие соседи (K-Nearest Neighbors)
- Наивный Байес (NaiveBayes)
- Машины опорных векторов (SupportVectorMachines)

Логистическая регрессия [7] – это алгоритм классификации, используемый для оценки дискретных значений, обычно двоичных, таких как 0 и 1 (False или True соответственно), «да» или «нет». Логистическая регрессия хорошо работает, когда характеристики и вероятность событий линейны, используется для задач двоичной (**двухклассовой**) классификации – результаты попадают в одну из двух категорий, что можно посмотреть на рисунке 1.2.

Деревья решений [7] – это универсальные и простые методы классификации и регрессии. Рекурсивное разделение набора данных на подгруппы ключевых критериев обеспечивает древовидную структуру. Моделируются решения и их возможные результаты в виде дерева, где ветви представляют выбор, а листья представляют результаты. Наиболее распространенный вид дерева представлен на рисунке 1.3.

K-ближайшие соседи (KNN) [7] – это алгоритм «ленивого» обучения, основанный на экземплярах, в котором функция аппроксимируется только локально, а все вычисления откладываются до вычисления функции. Непараметрический KNN не имеет никаких предположений о распределении данных. Классификация новых случаев происходит на основе меры сходства (например, функции расстояния).

KNN широко используется в рекомендательных системах, обнаружении аномалий и распознавании образов благодаря своей простоте и эффективности при обработке нелинейных данных. Обратимся к рисунку 1.4, являющимся примером классификации K-ближайших соседей.

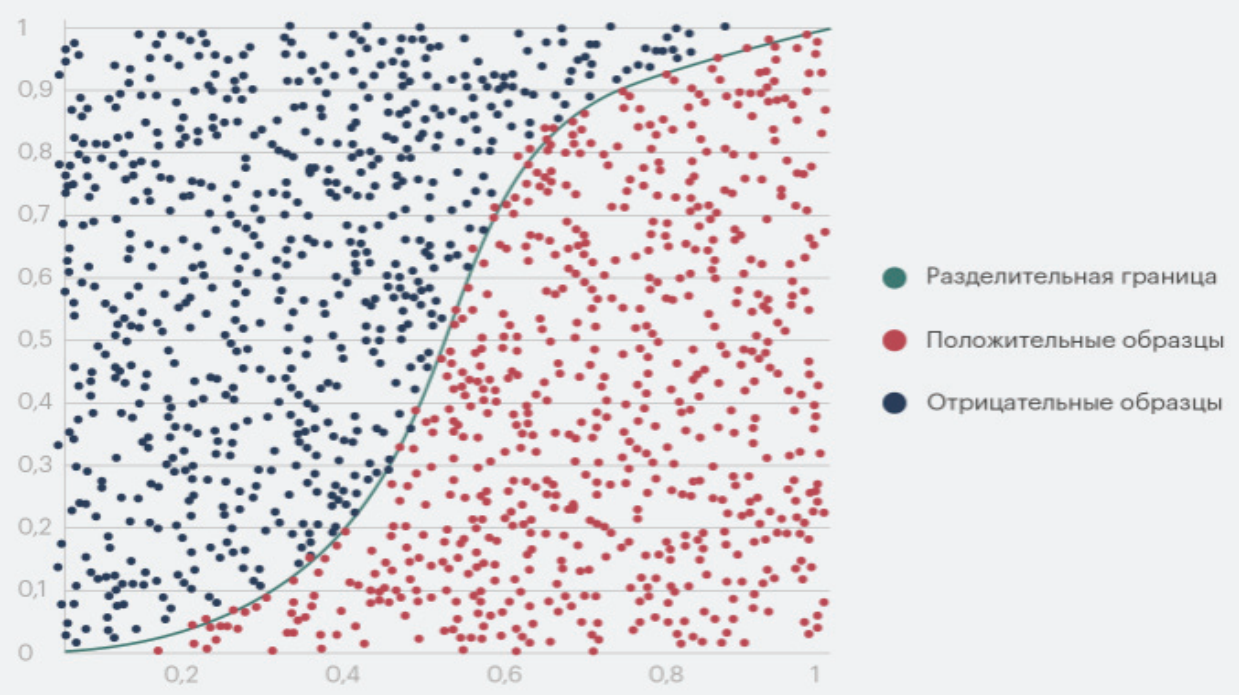

Рис. 1.2. Пример логической регрессии

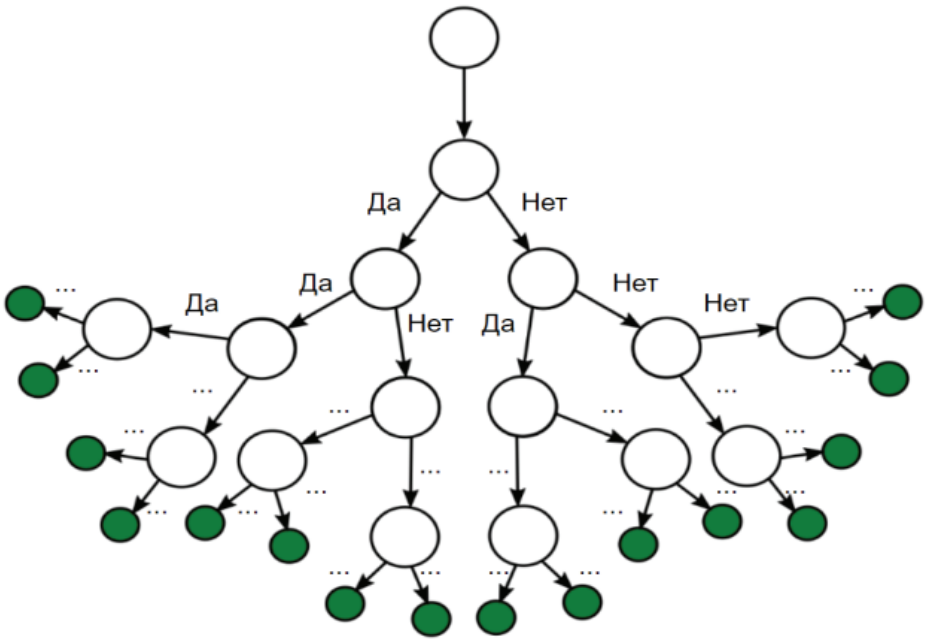

Рис. 1.3. Пример дерева решений

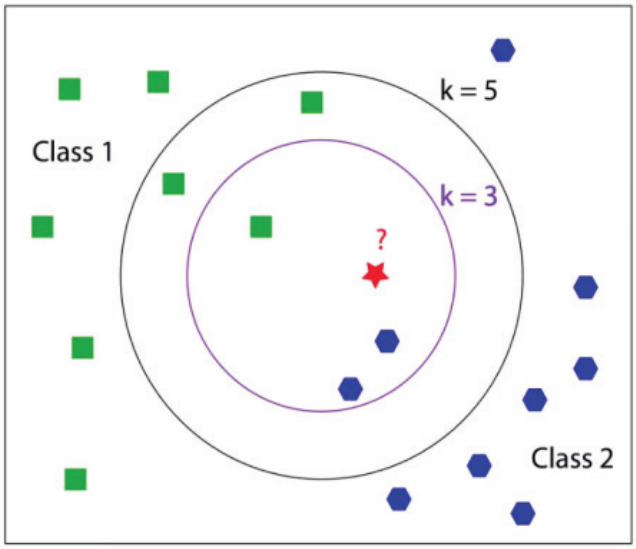

Рис. 1.4. Пример классификации K-ближайших соседей

Тестовый образец (красная звезда) должен быть классифицирован как зеленый квадрат (класс 1) или как синий восьмиугольник (класс 2).

Наивные классификаторы Байеса представляют собой набор алгоритмов классификации, основанных на теореме Байеса. Это не отдельный алгоритм, а семейство алгоритмов, все из которых имеют общий принцип, то есть каждая пара классифицируемых признаков независима друг от друга. Классификатор предполагает, что признаки, используемые для описания наблюдения, условно независимы, учитывая метку класса.

Наивный Байес быстро обрабатывает многомерные наборы данных. Он широко используется в классификации текста, а также для фильтрации спама и т. д. Преимуществом использования наивного Байеса является его скорость. На рисунке 1.5 можно посмотреть анализ задачи данным методом.

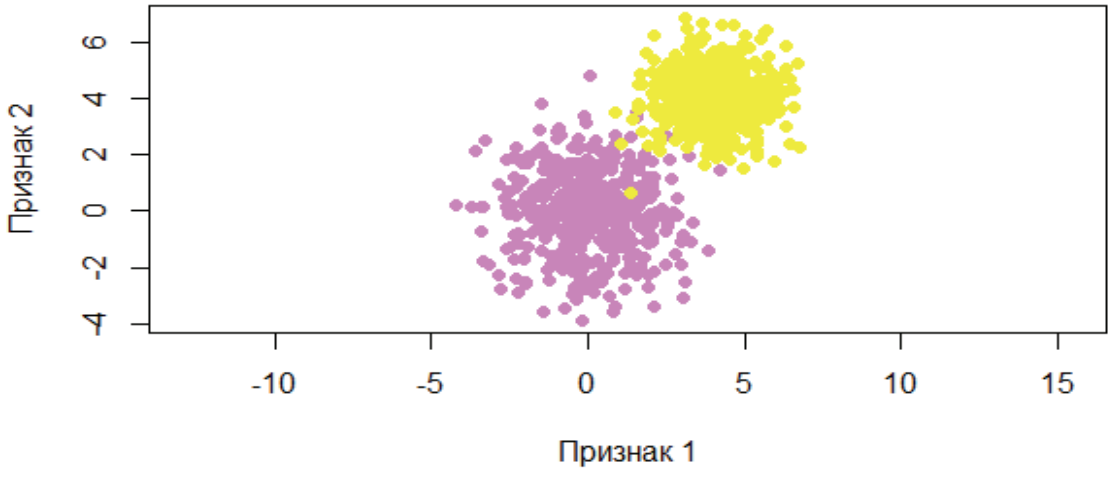

Рис. 1.5. Наивный Байесовский классификатор

Машина опорных векторов (SVM) [6] – еще один эффективный алгоритм классификации и регрессии. Он ищет гиперплоскость, которая лучше всего классифицирует данные, увеличивая при этом границу. SVM хорошо работает в многомерных областях и обрабатывает нелинейное взаимодействие функций с помощью своей техники ядра. Это мощный алгоритм классификации, известный своей точностью в многомерных пространствах.

SVM устойчив к переоснащению, а также находит применение в распознавании изображений, классификации текста и биоинформатике, где точность имеет первостепенное значение. Обратимся к рисунку 1.6, иллюстрирующему данный алгоритм.

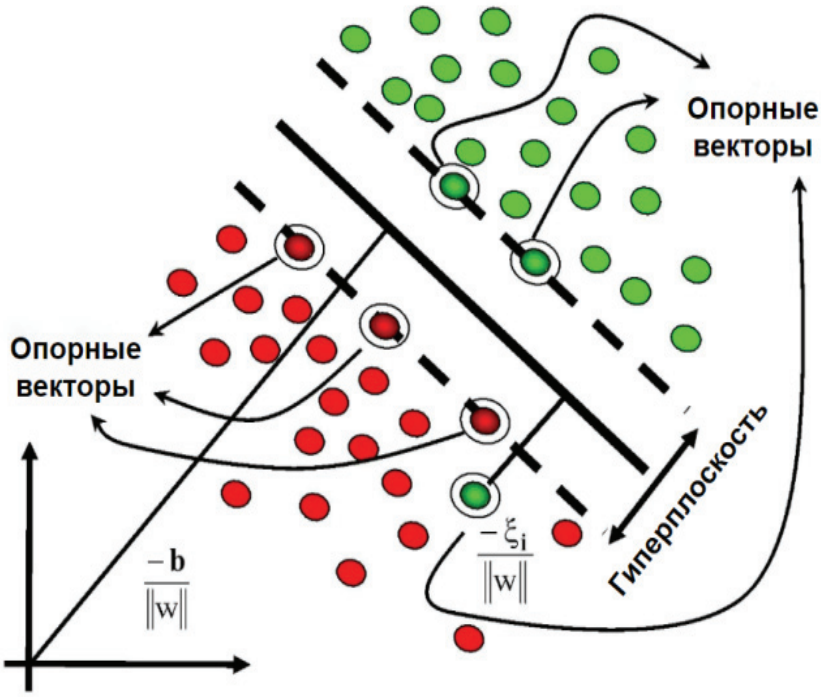

Рис. 1.6. Метод опорных векторов

Исследование эффективных алгоритмов машинного обучения в задачах классификации является важным шагом в развитии компьютерных технологий и автоматизации процессов анализа данных. Полученные результаты могут быть использованы для улучшения точности и надежности систем машинного обучения, а также для усовершенствования инновационных решений в различных областях применения.

Литература и источники:

1. Задачи машинного обучения в ML.NET, https://learn.microsoft.com/ru-ru/dotnet/machine-learning/resources/tasks.

2. 10 популярных алгоритмов машинного обучения для решения задач классификации. https://mlcentre.ru/articles/665241/.

3. Информационная аналитика и информационно-аналитические технологии в контексте социального управления. МГТУ имени Н.Э. Баумана. Москва, 2023.
4. Как устроено машинное обучение: задачи, алгоритмы и виды machinelearning. https://skillbox.ru/media/code/kak-ustroeno-mashinnoe-obuchenie-zadachi-algoritmy-i-vidy-machine-learning/.
5. Модели машинного обучения: что это и как выбрать подходящую. https://practicum.yandex.ru/blog/modeli-mashinnogo-obucheniya/.
6. Ericson G., Franks L., Rohrer B. Как выбирать алгоритмы для машинного обучения Microsoft Azure. https://habr.com/ru/companies/microsoft/articles/317512.
7. Machine Learning: Algorithms, Real-World Applications and Research Directions.
8. Kotsiantis S.B., Zaharakis I.D., Pintelas P.E. Machine learning: a review of classification and combining techniques. Published online: 10 November 2007. Springer Science+Business Media B.V. 2007.