Элементы машинного обучения в практикуме по вычислительной физике

Кирюхина Н.В.
Калужский государственный университет им. К.Э. Циолковского

Введение

Алгоритмы интеллектуального анализа данных приобретают все большее значение в физическом эксперименте, превращаясь в важнейший инструмент исследования. За последние годы наблюдается экспоненциальный рост объемов информации и усложнение ее характеристик («большие данные»), особенно в области физики высоких энергий и астрофизики. Имеются примеры привлечения к анализу данных широкого круга пользователей (краудсорсинговые проекты, «гражданская наука», соревнования на специальных площадках). Назрела потребность ознакомить студентов, обучающихся по образовательным программам, предполагающим специализацию в области физики и астрономии, с элементами машинного и глубинного обучения. Для подготовки студентов и аспирантов в этой области на профессиональном уровне целесообразно разработать и включить в учебный план специализированный курс, однако продемонстрировать возможности методов машинного обучения в физике и астрофизике можно в рамках курса вычислительной физики.

Цель

Разработка и практическая апробация системы учебно-исследовательских заданий для практикума по вычислительной физике на тему «Машинное обучение в современной физике и астрофизике».

Методология, методы и методики

Выбор темы и отбор содержания обусловлен актуальностью использования методов интеллектуального анализа данных в области физики элементарных частиц и астрофизики, а также мировоззренческим и методологическим значением тех проблем, на основе которых проектируются задания. В качестве объектов учебного исследования выступают наборы данных, размещенные в открытом доступе на платформе Kaggle (площадка для соревнований по машинному обучению). Там же имеется среда для написания кода, не требующая установки автономных приложений и дополнительных библиотек. Работа с данными включает разведочный анализ с использованием различных способов визуализации, разбор примеров постановки и решения задач для данного набора.

Результаты

В рамках практикума по вычислительной физике предполагается ознакомить студентов с возможностями разведочного анализа и визуализации данных, и дать представление о задачах МО для данного набора в демонстрационном режиме. Более глубокое изучение этих алгоритмов и библиотек, в которых они реализованы (Scikit-learn, LightGBM, XGBoost, CatBoost, Tensorflow) не предусматривается, это может быть задачей отдельного курса.

На платформе Kaggle можно найти довольно много датасетов для иллюстрации задач, возникающих в таких предметных областях как астрофизика и физика высоких энергий, в том числе, связанных с важнейшими проектами, достижениями и открытиями последних десятилетий. Но далеко не все из них подойдут для начинающих. Основное требование датасету для первичного ознакомления с методами МО: он должен содержать небольшое число признаков (в пределах десяти) с понятным для обучающихся предметным смыслом. Желательно также, чтобы предметное содержание, отраженное в данных, было им знакомо из ранее освоенных образовательных программ. Этим критериям удовлетворяют наборы [1] и [2] астрофизической тематики.

Набор [1] основан на данных, предметное знание о которых должно быть знакомо достаточно широкому кругу обучающихся: эти сведения входят даже в программу школьных курсов физики и астрономии. Цель создания набора, как она сформулирована его автором, состоит в

том, чтобы продемонстрировать, что звезды следуют определенному графику в небесном пространстве, называемому диаграммой Герцшпрунга - Рассела, на основе которого можно классифицировать звезды по типу. Для создания набора использовались несколько соотношений: закон Стефана-Больцмана (для определения светимости звезды), закон смещения Вина (для определения температуры поверхности звезды с использованием длины волны), связь абсолютной звездной величины со светимостью, радиус звезды через температуру и светимость. Данные о 240 звездах, собраны на основе открытых источников, размещенных в сети Интернет, недостающие данные добавлены в набор вручную с использованием перечисленных выше соотношений. Набор содержит всего семь признаков: абсолютная температура звезды (К), относительная светимость (по отношению к Солнцу), относительный радиус (по отношению к Солнцу), абсолютная звездная величина, цвет, спектральный класс и тип звезды (красный карлик, коричневый карлик, белый карлик, звезда главной последовательности, супергигант, гипергигант) как целевая переменная для классификации. Учебные задания для этого набора группируются вокруг разведочного анализа и визуализации его результатов и могут иметь следующие формулировки:

1. Построить гистограммы для температуры, относительной светимости, радиуса и других признаков для всех типов и для каждого типа в отдельности.
2. Построить диаграммы рассеяния.
3. Сделать выводы

В наборе [2] абсолютная звездная величина и показатель цвета B-V могут быть использованы для бинарной классификации «карлик-гигант». Он может использоваться для знакомства и сравнительного анализа алгоритмов классификации.

Для более подготовленных студентов (в том числе в качестве заданий для специализированного курса) можно составить задания на классификацию на примере распознавания объектов на снимках звездного неба, в том числе идентификации экзопланет на основе кривых блеска. В 2017 году использование нейросети, созданной инженерами из Google Brain, привело к открытию двух новых экзопланет на основе данных, полученных с космического телескопа «Кеплер». Эта тематика отражена в датасете [3].

Проект Sloan Digital Sky Survey (SDSS), «Слоуновский цифровой небесный обзор» [4], действующий еще с 2000 года, позволил составить трехмерную карту Вселенной с глубокими многодиапазонными изображениями и спектрами более чем 3000 объектов. С данными SDSS связан один из старейших проектов в области «гражданской науки» - «Galaxy Zoo» [5]. Наборы данных [6] и [7] основаны на данных, полученных в результате их реализации. Набор [6] позволяет рассмотреть задачу классификации «Звезда, галактика или квазар?», а [7] – классификацию галактик по их изображениям.

Таблица 1.

Эксперимент	Годы сбора статистики	Схема распада	Эффективность
CMS	2011-2012	$H \rightarrow \gamma\gamma$ (распад на два фотона)	51%
ATLAS	2011-2012	$H \rightarrow \tau^{+}\tau^{-}$ (распад на два тау-лептона)	85%
ATLAS	2011-2012	$VH \rightarrow b\,\bar{b}$ (распад на пару b-кварк и его антикварк)	73%
ATLAS	2015-2016		15%
CMS	2011-2012		125%

Без использования МО сегодня невозможно представить экспериментальные исследования на ускорителях в области физики высоких энергий. Таблица 1, составленная на основе сведений, приводимых в [8] иллюстрируют ту роль, которую сыграли методы МО в одном из самых значимых открытий последних десятилетий – обнаружении бозона Хиггса. В перечень задач, которые решаются с помощью алгоритмов МО в экспериментах на ускорителях входит группировка сигналов, в соответствии с тем, какая частица их создала, определение типов и свойств частиц по информации о связанных с ними событиях, определение процессов, в результате которых возникли эти частицы. Это, как правило, задачи регрессии или классификации.

Пример относительно простого по структуре датасета, который можно предложить для анализа студентам - Particle Identification from Detector Responses [9]. Это задача распознавания 4 видов частиц (позитрон, пион, каон, протон) по данным шести детекторов. Набор представляет собой данные моделирования с помощью пакета Geant4 электронно-протонного неупругого рассеяния, измеренного системой детекторов частиц. Это процесс, используется для исследования внутренней структуры адронов, в данном случае протонов. Падающая частица (электрон) сталкивается с протоном-мишенью. Во время неупругого рассеяния протон может распадаться на составляющие его кварки, которые затем образуют адронную струю. Углы отклонения дают информацию о характере процесса. Набор содержит значения геометрических, кинематических и динамических характеристик частиц, которые могут быть использованы для идентификации: импульс, угловые координаты, число электронов, энергию на входе и на выходе.

Задания для работы с данными:

- выполнить разведочный анализ данных (построить гистограммы для признаков, корреляционную матрицу, диаграммы рассеяния);
- добавить еще одну переменную на основе имеющихся признаков (например, длину волны де Бройля электрона);
- сравнить результаты работы различных алгоритмов классификации.

Особой разновидностью классификационной задачи стало «обучение обнаружению», реализованное в соревновании «Higgs Boson Machine Learning Challenge», организованном ЦЕРНом в 2014 году на платформе Kaggle [10]. Своеобразие заключалось в том, что требовалось найти в пространстве признаков области значительного превышения сигнальных событий по сравнению с фоном и определить значимость превышения. Если вероятность того, что событие обусловлено фоновыми процессами, падает ниже предела то новая частица считалась обнаруженной. Результат конкурса стал свидетельством значения краудсорсинговых проектов с привлечением широкого круга пользователей, не являющимися специалистами в предметной области (более 2000 участников, высокие результаты победителей [11]). Целевая функция, представляющая значимость открытия новой частицы, в работах победителей составила 3.80σ, в то время как альтернативные модели давали не более 3.50 σ. Вдохновленные этим успехом, ученые ЦЕРНа организовали еще одно соревнование – «Flavours of Physics: Finding $\tau \rightarrow \mu\mu\mu$. Identify a rare decay phenomenon» с аналогичной задачей [12, 13]. Для начинающих сложной является как сама постановка, так и предметное содержание, а также структура данных, поэтому в рамках занятий можно только продемонстрировать некоторые готовые решения и результаты.

Работа с датасетами [1] и [4] была апробирована в рамках курсового проектирования в 2020-2021 году. Задания выполнялись при написании курсовых работ по дисциплинам «Физика атомного ядра и элементарных частиц», «Проектирование в профессиональной деятельности» со студентами, обучающимися по направлению «Педагогическое образование» с предметной специализацией «Физика» и «Математика»). В 2021-22 годах

практические работы на основе этих данных включены в программу практикума по дисциплине «Вычислительная физика».

Заключение

Представлена система учебно-исследовательских заданий для занятий компьютерного практикума по вычислительной физике: разведочный анализ данных, добавление новых признаков на основе имеющихся, решение задач классификации. Показаны возможности использования наборов данных, размещенных в открытом доступе для ознакомления студентов с примерами решения фундаментальных проблем современной физики и астрофизики с помощью алгоритмов машинного обучения.

Список источников.

1. Star dataset to predict star types. A 6 class star dataset for star classification with Deep Learned approaches [Electronic resource]. – URL: https://www.kaggle.com/deepu1109/star-dataset (дата обращения: 7.11.2021)
2. Star Dataset: Stellar Classification [Beginner] Identify Giants and Dwarfs through Machine Learning [Electronic resource]. – URL: https://www.kaggle.com/vinesmsuic/star-categorization-giants-and-dwarfs (дата обращения: 7.11.2021)
3. Kepler Exoplanet Search Results [Electronic resource]. – URL: https://www.kaggle.com/nasa/kepler-exoplanet-search-results (дата обращения: 7.11.2021)
4. Dark Energy Survey completes six-year mission // Symmetry magazine. [Electronic resource]. – URL: https://www.symmetrymagazine.org/article/dark-energy-survey-completes-six-year-mission / (дата обращения: 7.11.2021)
5. Galaxy Zoo [Electronic resource]. – URL: http://zoo1.galaxyzoo.org (дата обращения: 7.11.2021)
6. Sloan Digital Sky Survey DR14. Classification of Stars, Galaxies and Quasars [Electronic resource] – URL: https://www.kaggle.com/lucidlenn/sloan-digital-sky-survey (дата обращения: 7.11.2021)
7. Galaxy Zoo 2: Images. Shape Galaxy Clustering [Electronic resource]. URL: https://www.kaggle.com/jaimetrickz/galaxy-zoo-2-images (дата обращения: 7.11.2021)
8. Machine learning at the energy and intensity frontiers of particle physics // Nature [Electronic resource]. – URL: https://www.nature.com/articles/s41586-018-0361-2 (дата обращения: 7.11.2021)
9. Particle Identification from Detector Responses [Electronic resource]. URL: https://www.kaggle.com/naharrison/particle-identification-from-detector-responses/version/2 (дата обращения: 7.11.2021)
10. Higgs Boson Machine Learning Challenge [Electronic resource]. – URL: https://www.kaggle.com/c/higgs-boson (date of treatment: 7.11.2021).
11. ML2014: Higgs Boson Machine Learning Challenge. University of California. – URL: https://www.math.uci.edu/icamp/summer/research/student_research/ml_hb_2014.pdf (дата обращения: 7.11.2021)
12. Flavours-of-physics [Electronic resource]. – URL: https://www.kaggle.com/duncandean/flavours-of-physics-baseline (date of treatment: 15.04.2020) / (дата обращения: 7.11.2021).
13. Flavours of Physics: the machine learning challenge for the search of $\tau - \rightarrow \mu - \mu - \mu +$ decays at LHCb [Electronic resource]. – URL: https://storage.googleapis.com/kaggle-competitions/kaggle/4488/media/lhcb_description_official.pdf