Машинное обучение, как средство повышения эффективности разработки экспериментальных систем

профессор Ососков Г.А.
gososkov@gmail.com
Home page http://gososkov.ru

Машинное обучение

Машинное обучение (Machine Learning-ML), это когда компьютер не просто использует заранее написанный алгоритм, а сам обучается решению поставленной задачи на большой выборке данных.

Три причины синхронного взрыва популярности ML в последние годы:

1. **Большие Данные**. Данных стало так много, что новые подходы были вызваны к жизни тем, что растущее разнообразие, как данных, так и возможных решений стало слишком велико для традиционных заранее запрограммированных систем.

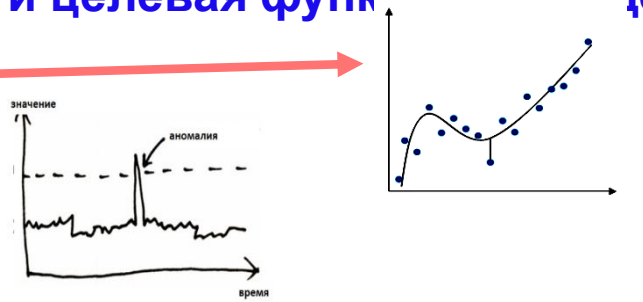
2. Снижение стоимости **параллельных вычислений и памяти компьютеров**.

3. Новые алгоритмы **глубокого машинного обучения**.

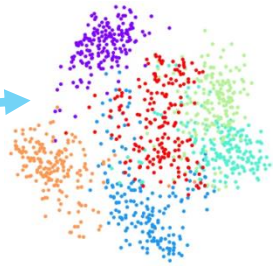
Способы машинного обучения

-обучение с учителем (supervised learning), когда у нас есть данные, на основании которых нужно что-то предсказать, конечный набор известных решений и целевая функция, определяющая качество решения. Примеры обучения с учителем –

- задачи регрессии,
- классификации,
- выявления аномалий,
- распознавание образов.



- **обучение без учителя (unsupervised learning), когда у нас есть только данные, свойства которых мы и хотим найти.** Примеры обучения без учителя –
  - задачи кластеризации
  - уменьшения размерности данных
  - Распознавание траекторий частиц (трекинг) нейросетями Хопфилда





- **Обучение с подкреплением (reinforcement learning),** когда нейросетевой агент взаимодействует с окружающей средой, а она его за эти действия поощряет. Агент не знает, как ему действовать дальше, но зато узнает, какое действие принесет максимальную награду. Примеры
  - все игровые программы,
  - робототехника,
  - самоуправляемые автомобили.

Наиболее популярные методы решения этих задач:
деревья решений, метод опорных векторов, роевые алгоритмы, генетические алгоритмы, **нейронные сети.**

Применение машинного обучения в прикладных задачах ОИЯИ

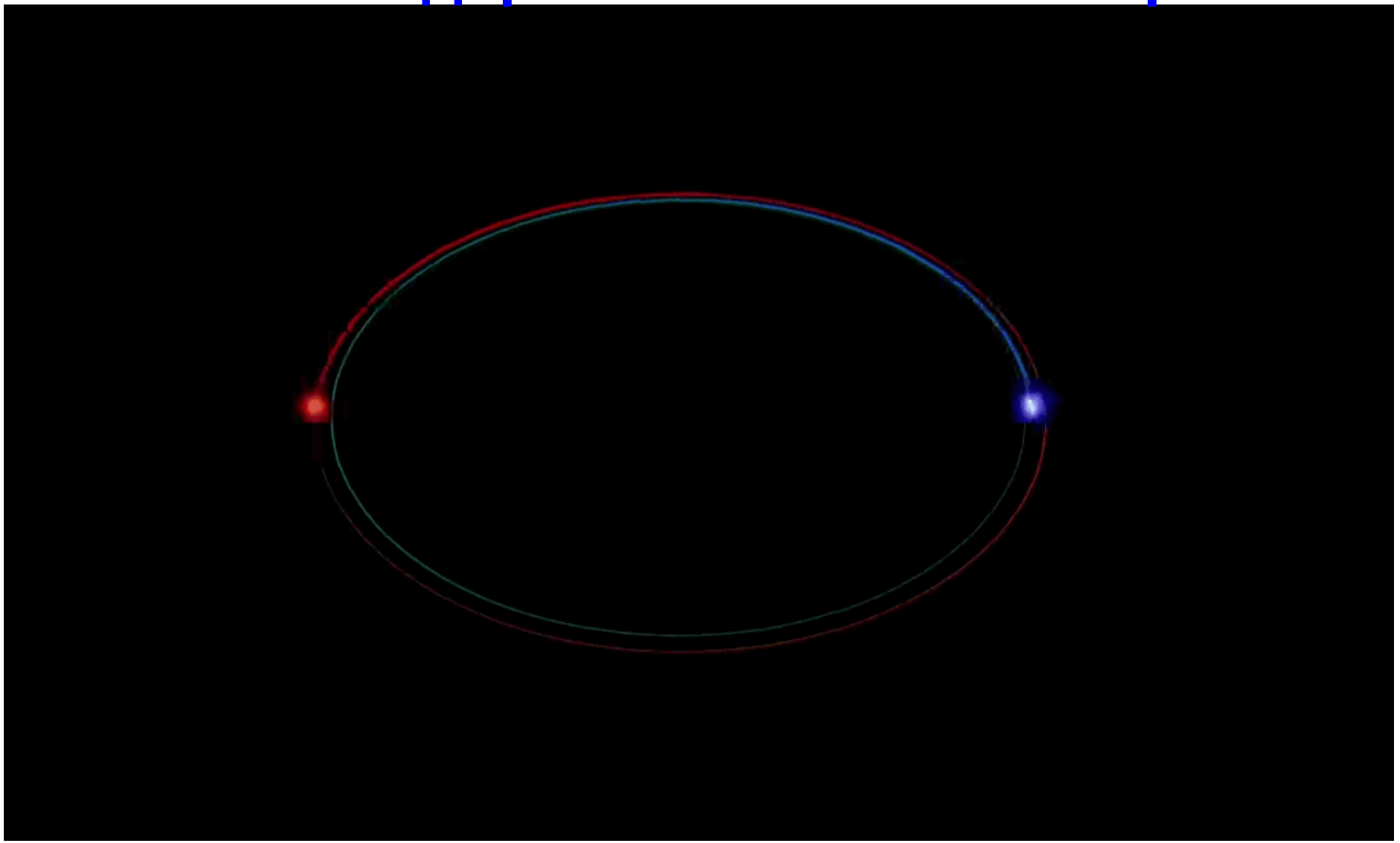
1. **Реконструкции событий по данным детекторов – ключевая проблема экспериментальной физики**
   - **Распознавание траекторий взаимодействующих частиц в экспериментах физики высоких энергий**
   - **Идентификация частиц по данным ддетекторов**
   - **Прогнозирование работы аппаратуры по данным временных измерений**

3. Прогнозирование загрязнения атмосферы тяжелыми металлами с использованием космических снимков

4. Разработка алгоритмов для выявления и профилактики заболеваний сельскохозяйственных растений

5. Моделирование центров хранения и обработки данных с учетом экономических составляющих

Взаимодействие частиц в коллайдере NICA. События и треки



Задачи анализа данных в ФВЭ и ЯФ

Распознавание образов для реконструкции событий: трекинг, нахождение вершин событий, распознавание колец черенковского излучения, а также выявление и удаление ложно распознанных объектов. Применяемые методы

- преобразования Хафа,
- клеточные автоматы,
- фильтр Калмана,
- искусственные нейронные сети,
- вейвлет-анализ и др.

Оценивание физических параметров

- методы математической статистики;
- робастное оценивание

Проверка гипотез

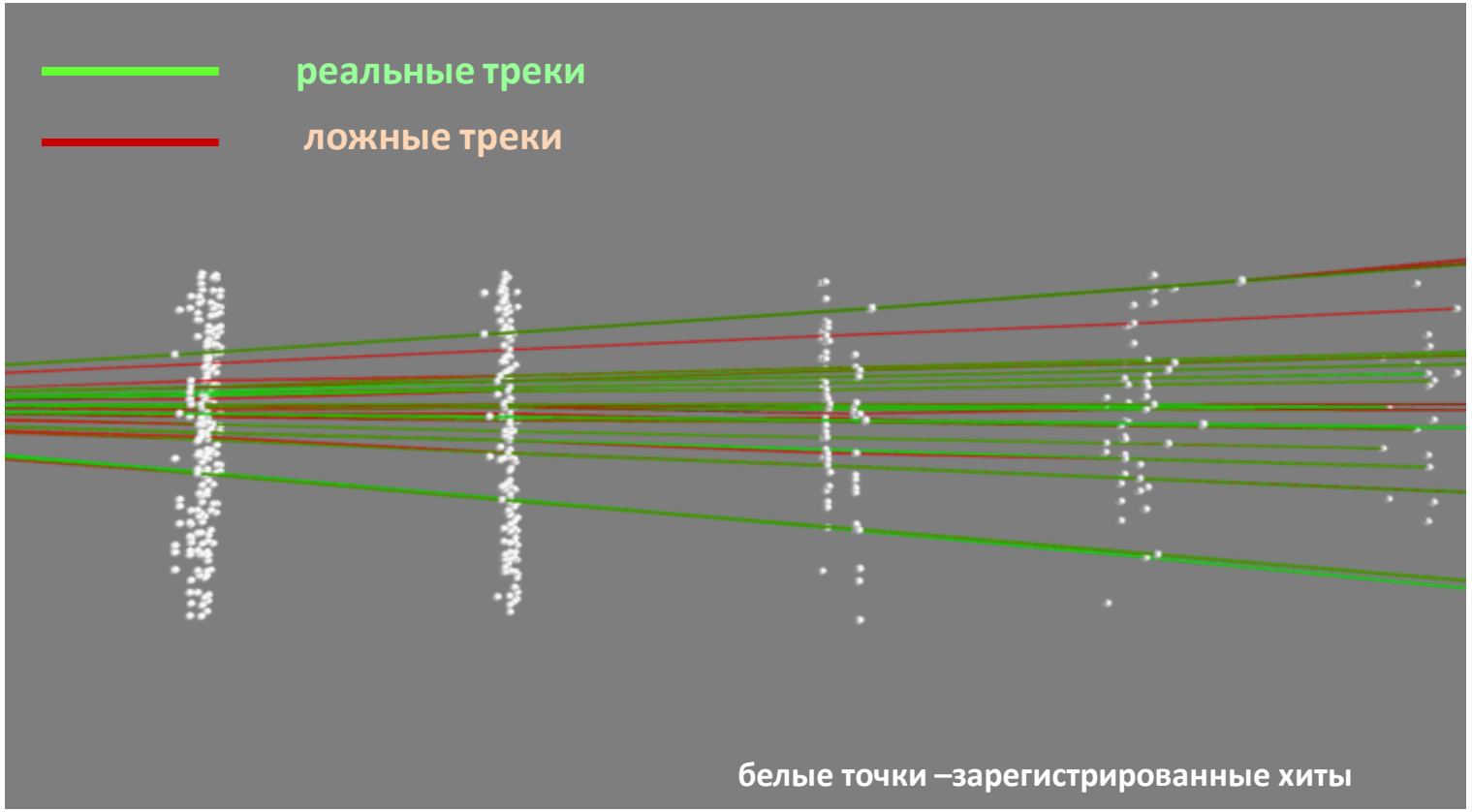
- отношения правдоподобия,
- Искусственные нейронные сети
- усиленные алгоритмы машинного обучения (boosted decision trees - BDT).

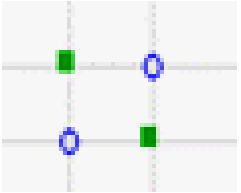
Моделирование выполняется на всех этапах анализа данных, более того, есть насущная необходимость **моделировать структуру и работу самих компьютерных систем обработки и хранения** данных, включая и их грид и облачную инфраструктуры

По каждому их этих методов может быть сформулирована тема магистерской диссертации!

Трекинг - одна из самых важных задач

Трекинг или распознавание треков - это процесс **восстановления траекторий частиц в детекторе** ФВЭ путем прослеживания и соединения точек- хитов (*хит* – это реконструированный отклик детектора), которые каждая частица оставляет, проходя через плоскости детектора. К сожалению, из-за того, что большинство детекторов являются стриповыми, **хиты бывают и ложные, причем ложных на порядок больше правильных**





Процедура трекинга включает в себя фазы:
(1) получения хитов (hit clustering),
(2) построения треков-кандидатов - наборов хитов с вычисленными параметрами,
(3) прослеживания треков и
(4) их подгонки уравнением движения частицы в магнитном поле. Подгонка нужна для определения импульса, заряда и пространственных углов вылета частицы из точки взаимодействия необходимых для идентификации частицы, породившей данный трек .

Данные, измеренных в эксперименте и постановки задач

Задача: определять болезни растений по фото их листьев



Решение:

- **использовать облачную платформу ЛИТ (PDDP) для обнаружения болезней растений с помощью современных технологий организации и глубокого обучения ;**
- **В качестве конечного продукта разработать мобильное приложение, позволяющее пользователям отправлять фотографии и текстовое описание больных растений, чтобы узнавать причину болезни и получать рекомендации по спасению урожая.**

Нейтронный активационный анализ мхов

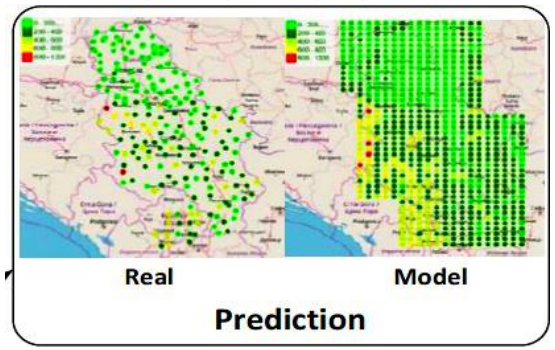
Международная программа мониторинга и оценки воздействия загрязнителей воздуха на растительность (ICP Vegetation).



Мхи - идеальные поглотители загрязненного воздуха. Волонтеры собирают мхи в определенных местах, отправляют их в ОИЯИ для проведения нейтронно-активационного анализа, чтобы выяснить, какие вредные вещества загрязняют воздух в данной местности.

Задача: экстраполировать эти данные из дискретных точек сбора мхов на окружающую местность для каждого из загрязнителей с учетом особенностей окружающей среды, таких как, направления ветра и водостоков

Решение: Система управления данными (DMS) разработана на облачной платформе ОИЯИ и предоставляет участникам современную единую систему сбора, анализа и обработки данных биологического мониторинга. DMS позволяет организовать гибкий сбор данных с учетом неоднородности исходных данных, их автоматическую проверку , вычисляет его статистические параметры, геоиндексы и так далее. Сейчас в DMS представлено более 6000 мест отбора проб из 47 регионов разных стран мира



Моделирование компьютерных систем

Разработка сложнейших компьютерных систем сбора, передачи и распределённой обработки сверхбольших объемов информации требует больших предварительных исследований по выбору оптимальной их структуры с учетом стоимости и предполагаемых ресурсов и загрузки.
Такие исследования должны основываться на тщательном моделировании как потока заданий с учетом их типов и статистических данных о распределении времени их поступления и требуемых компьютерных ресурсов для их выполнения, так и состава моделируемой компьютерной системы с учетом ее облачной и грид инфраструктуры

Актуальная тема по глубокому обучению:

Разработка интеллектуальных алгоритмов для обнаружения аномалий в показаниях систем мониторинга ЦОД

Список тем предлагаемыхмагистерских дипломов можно скачать на моем сайте
http://gososkov.ru/UNI-DUBNA/, кликнув
Темы магистерских диссертаций на 2026-27.docx
Там же лежит и эта лекция
Г.А. Ососков_приглашение магистрам 2026.pptx

Спасибо за внимание!