Научная конференция НИИЯФ МГУ по итогам 2023 года
26.02.2024

Машинное обучение и обработка данных:

Основные результаты ЛАМОД за 2023 г.

С.А. Доленко

Основные направления работы ЛАМОД

- Решение многопараметрических обратных задач (ОЗ)
  - ОЗ разведочной геофизики (подход «от модели»)
  - ОЗ спектроскопии (подходы «от эксперимента» и «квазимодельный»)
- Решение задач прогнозирования космической погоды
  - Геомагнитные индексы, потоки заряженных частиц на ГСО
  - Прогнозирование уровня (класса) геомагнитного возмущения
- Обработка сигналов полупроводниковых газовых сенсоров в динамическом температурном режиме («электронный нос»)
- Развитие алгоритмов отбора входных признаков задачи
  - Итеративный подход на основе поочередного удаления
  - Алгоритм отбора в условиях мультиколлинеарности признаков
- Разработка гендерного варианта генетических алгоритмов
- Разработка алгоритмов и методик решения различных типов задач
  - Новые типы архитектур, генерация данных, доменная адаптация, перенос обучения, дрейф динамики и др.

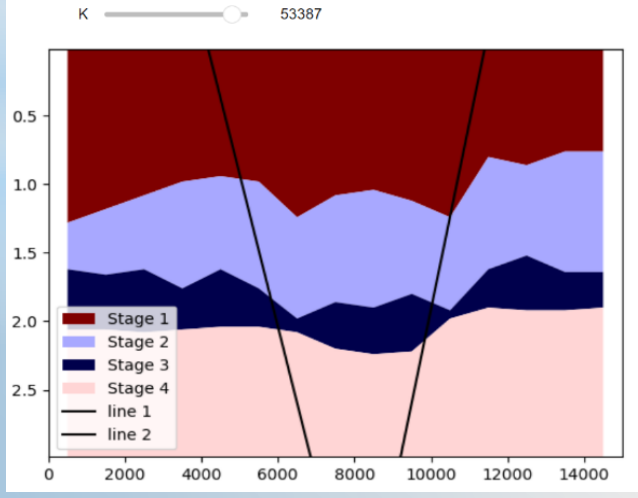
Основные направления работы ЛАМОД (2)

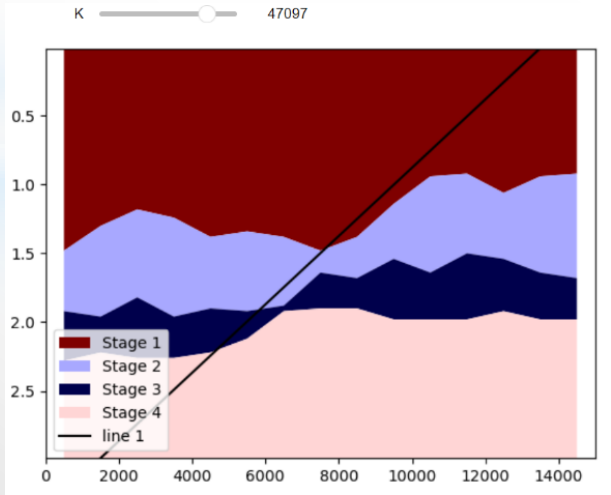
- Экспериментальные исследования в области спектроскопии
- Прогнозирование условий/параметров гидротерамального синтеза углеродных точек (задача «параметры-свойства»)
- Обработка данных в когнитивных исследованиях
  - Применение методов машинного обучения для анализа данных фМРТ. Задача «Когновизор».
  - Построение взаимного отображения параметров мимики и параметров эмоционального состояния (валентность, возбуждение, доминантность)
  - «Эволюция глубокого обучения» (перспективное направление): от содержательных моделей к нейронным сетям и их эволюционной оптимизации
- Преподавание в области машинного обучения
  - Факультативные курсы для студентов и курсы повышения квалификации

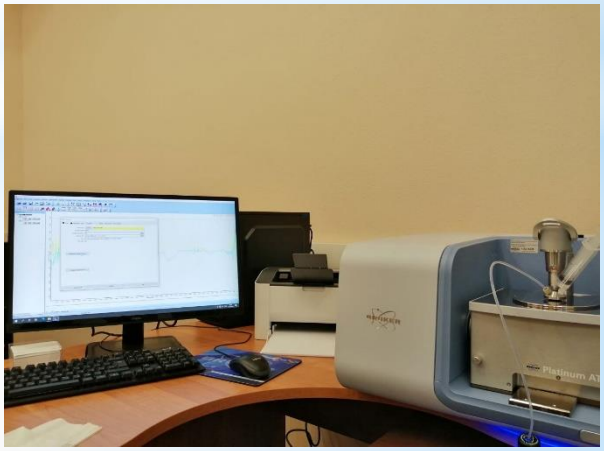
Сотрудничество с ЛКФИ ОКН НИИЯФ, физическим, химическим факультетами МГУ, РГГРУ, ИИКС НИЯУ МИФИ, обсуждение сотрудничества с ИПИМ МГУ, НИЦ «КИ». Участие в научно-образовательной школе МГУ «Космос».

Получение данных для решения многопараметрических обратных задач (ОЗ)

- ОЗ разведочной геофизики (подход «от модели»)
  - Путем расчета прямой задачи получены уникальные 2D и 3D массивы данных для различных геологических структур, для трёх видов разведочной геофизики (по 26 000 – 30 000 примеров) : гравиразведка, магниторазведка, электроразведка (магнитотеллурика)
- ОЗ оптической спектроскопии (подход «от эксперимента»)
  - Путем физического эксперимента получены уникальные массивы данных спектров оптического и ИК поглощения многокомпонентных модельных растворов (2442 пробы) для определения концентраций 5 основных компонентов в винах (этанол, сахар, глицерин, кислоты, диоксид серы)

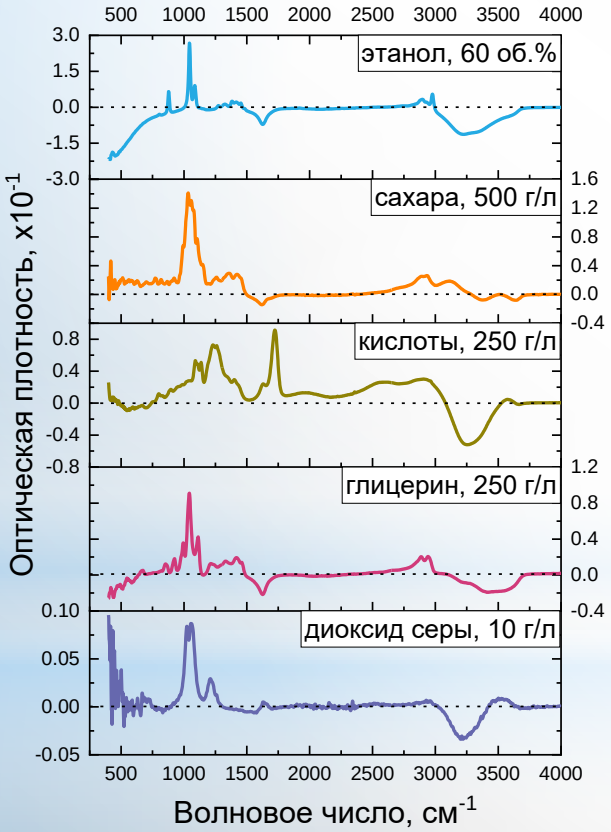




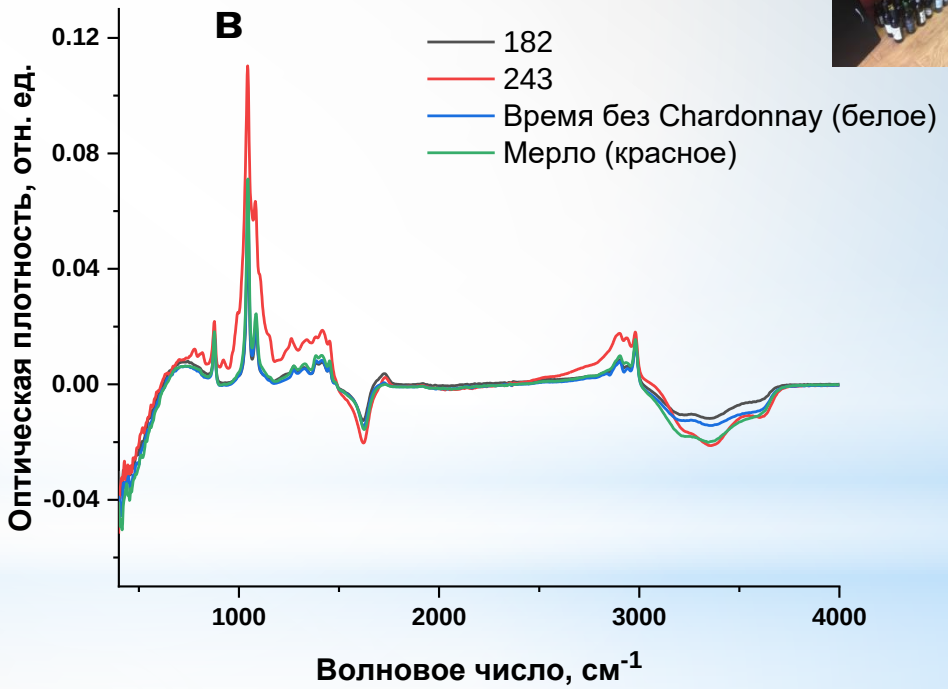


Диагностика вин по оптическим спектрам

- Наиболее эффективным оказалось использование спектров ИК поглощения. Можно работать и с красными винами!!!
- Спектры модельных растворов близки к спектрам реальных вин



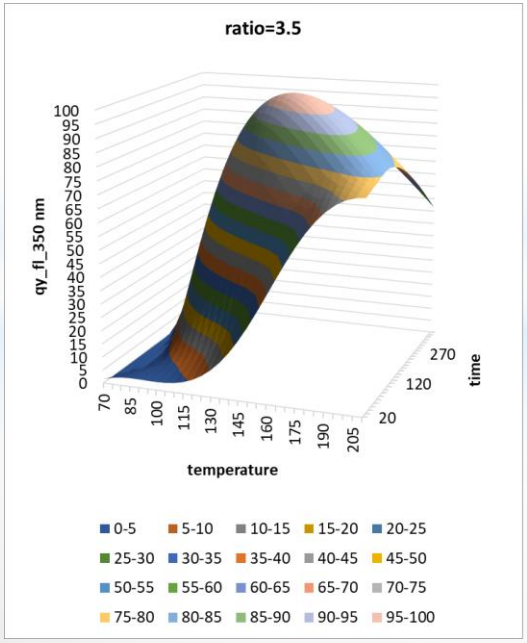
Спектры ИК поглощения базовых растворов компонентов

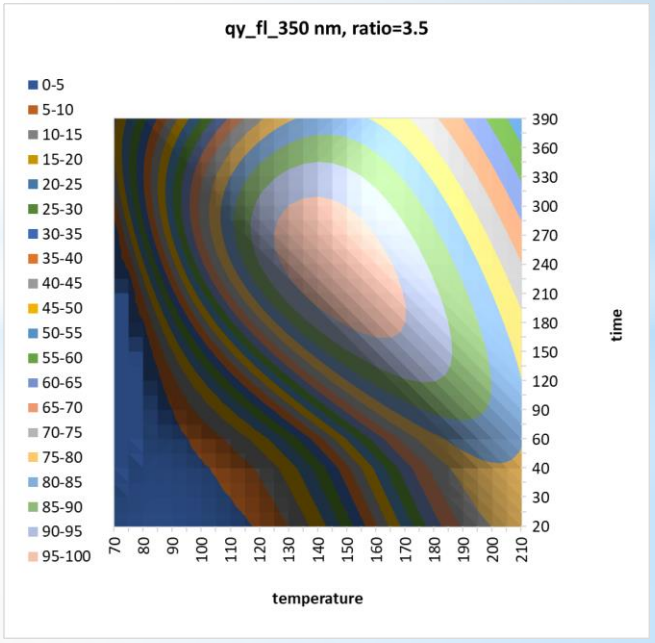


Спектры поглощения реальных и модельных вин

Прогнозирование оптимальных параметров гидротермального синтеза углеродных точек

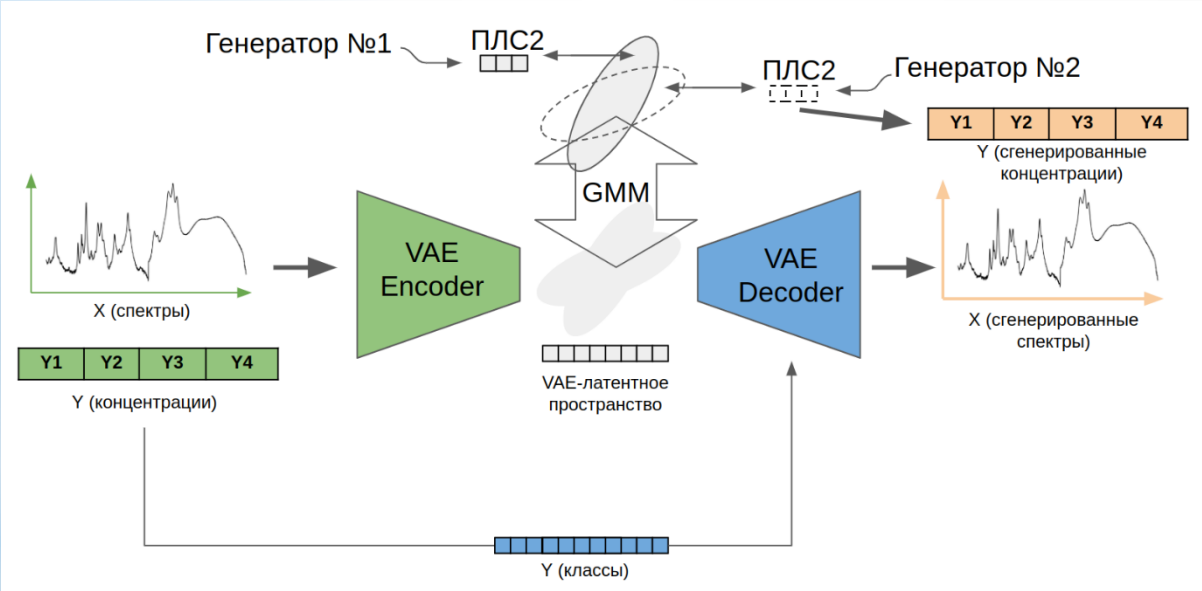
- Оптимизация параметров гидротермального синтеза углеродных точек с целью максимизации квантового выхода флуоресценции
- Оптимизируются значения следующих параметров:
  - Соотношение концентраций прекурсоров – от 0.1 до 20
  - Температура синтеза – от 80 °C до 200 °C
  - Время синтеза – от 30 до 360 сек
- Аппроксимация зависимости квантового выхода от параметров
- Оптимальные параметры синтеза: 3.5, 145 °C, 240 сек
- Прогноз квантового выхода 99.15%
- Экспериментальное значение 98.9%

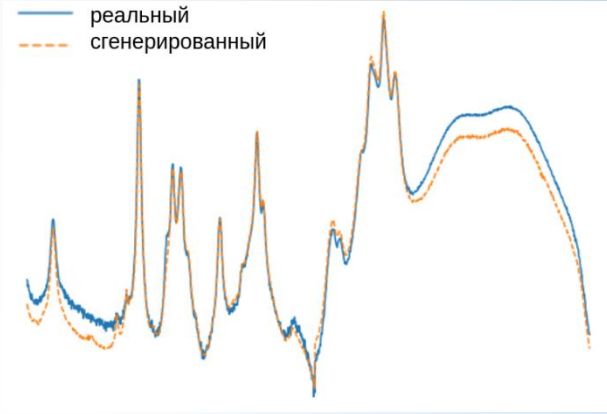




Нейросетевая генерация модельных спектров

- Для генерации модельных спектров использовались обусловленные вариационные автоэнкодеры– вид генеративных состязательных сетей, с заданием условий в латентном пространстве ВАЭ с использованием кластеризации (GMM) и моделей проекций на латентные структуры (ПЛС)



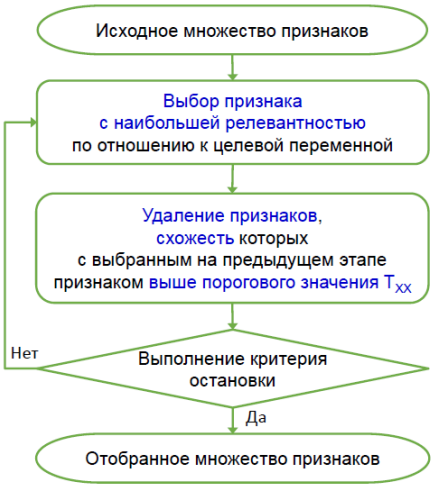


Высокая реалистичность сгенерированных спектров

- Подача на генератор вектора-условия позволяет генерировать примеры, соответствующие определённым концентрациям компонентов
- Разрабатывается альтернативный подход на основе обыкновенных ВАЭ с определением значений концентраций компонентов отдельной моделью в пространстве спектров

Алгоритм отбора входных признаков в условиях их мультиколлинеарности

Итеративный алгоритм отбора признаков

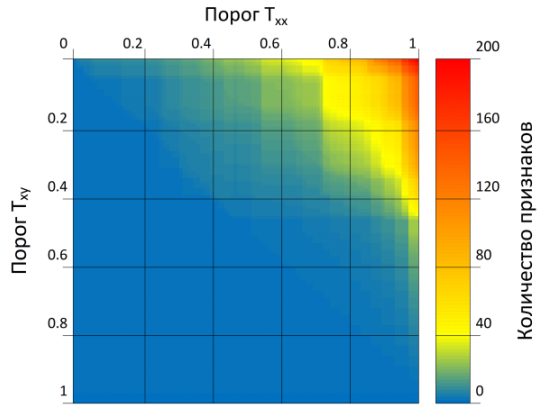


- Если входные признаки имеют сильную связь, то отбирать их одновременно неэффективно
- Отбор позволяет уменьшить число признаков без ухудшения или с улучшением модели
- Работа алгоритма показана на примере задачи прогнозирования индекса Dst на 1 час вперед
- Интересен анализ отбираемых признаков с физической точки зрения

Зависимость количества отбираемых признаков от параметров алгоритма

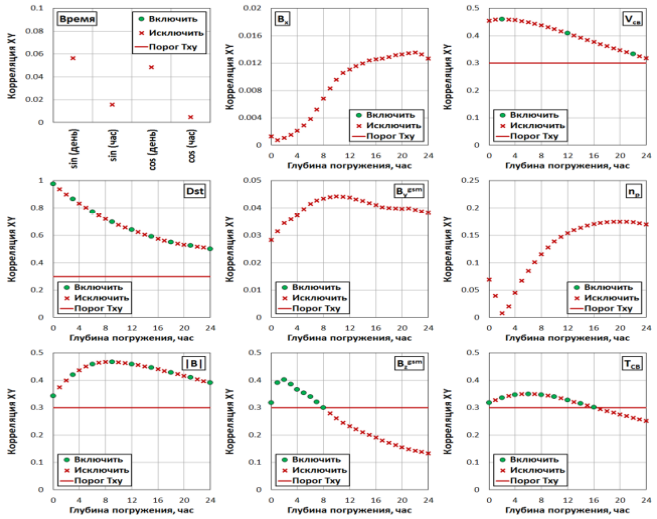
Порог Txx – максимальная кросс-корреляция

Порог Txy – минимальная корреляция с выходом



Отбираемые признаки

Приведены результаты для Txx=0.9, Txy=0.3



Результаты

- В таблице приведены результаты для различных комбинаций параметров Txx и Txy
- Выбраны комбинации с примерно 5-кратным сокращением количества входных признаков
- Тривиальная инерционная модель: прогноз равен последнему известному значению
- Полный набор признаков: Txx=1, Txy=0
- Корреляционный фильтр: Txx=1

Параметры отбора	Кол-во признаков	Тестовый набор 2018-2022		
		R^2	MAE	RMSE
Трив. модель	1	0.9268	0.1229	0.1803
Txx=1, Txy=0	204	**0.9578**	**0.0961**	**0.1351**
Txx=1, Txy=0.45	40	*0.9379*	*0.1120*	*0.1633*
Txx=0.9, Txy=0.3	39	**0.9509**	**0.1037**	**0.1456**
Txx=0.85, Txy=0.2	38	0.9499	0.1053	0.1472
Txx=0.8, Txy=0.15	41	0.9498	0.1050	0.1473
Txx=0.75, Txy=0.15	39	0.9491	0.1057	0.1480

Прогнозирование класса геомагнитного возмущения

- Вместо уровня геомагнитного возмущения (значения индекса) прогнозируем его класс (принадлежность к тому или иному диапазону)
- Рассматривались три класса по индексу Kp (классы не сбалансированы!!!):
  - Kp < 2 – отсутствие возмущений (спокойная магнитосфера)
  - 2 ≤ Kp ≤ 3+ – слабые геомагнитные возмущения
  - Kp ≥ 4- – средние и сильные геомагнитные возмущения
- Горизонт прогнозирования от 3 до 24 часов с шагом 3 часа
- Без погружения или с погружением временных рядов на 8 отсчетов (сутки)
- Спектр методов машинного обучения:
  гребневая и логистическая регрессии, случайный лес, градиентный бустинг, многослойные персептроны, рекуррентные сети (LSTM и GRU)
- Методика SMOTE для преодоления несбалансированности классов
- Кросс-валидация для выбора оптимальных гиперпараметров
- Для оценки качества моделей перешли от метрики F_1 к метрике F_2 (важнее не пропустить событие геомагнитного возмущения)
- Предсказания тривиальной модели удалось превзойти на всех горизонтах
- Адаптивный отбор входных признаков даёт физически объяснимые результаты

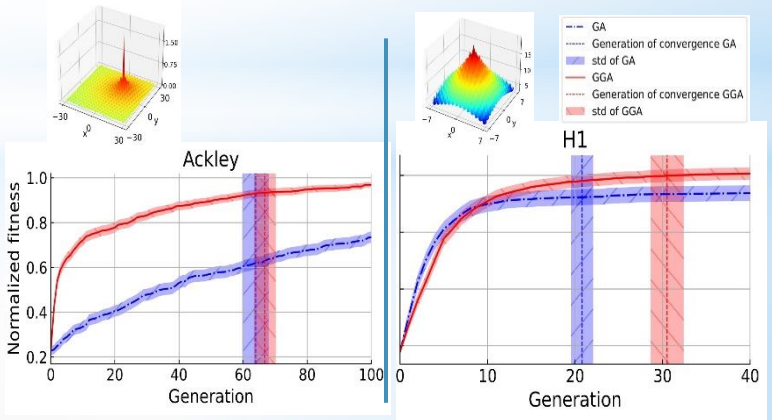
Обработка данных газовых сенсоров

Продемонстрирована высокая эффективность нейросетевых методов при обработке данных полупроводникового газового сенсора, работающего в динамическом температурном режиме

- Исследовано высокоселективное обнаружение различных отдельных газов (CO, H_2, CH_4, NH_3, NO, NO_2, H_2S, SO_2, HCOH) при низких концентрациях (0.01–667 ppm) в воздухе с помощью одного датчика из оксида металла (MOX-sensor) (12 типов сенсоров) – задачи бинарной и многоклассовой классификации, определение концентрации газа (задача регрессии).
- Датчик работает в динамическом температурном режиме (6 различных режимов модуляции температуры, линейный или импульсный нагрев).
- Задачи обработки данных решались методами машинного обучения.
- Были определены сенсоры, наиболее селективные по отношению к каждому из газов, и оптимальный режим модуляции температуры.
- Планируемое направление развития работ – одновременное определение нескольких газов с помощью набора датчиков.

Разработка гендерного варианта генетических алгоритмов

- Генетические алгоритмы (ГА) – многоагентный метод оптимизации
- Одновременно рассматривается набор (популяция) решений
- На основании отбора, скрещивания и мутации порождаются новые решения-индивидуумы, улучшающие значение целевой функции
- В стандартном ГА все индивидуумы имеют одинаковые свойства
- В гендерном ГА (ГГА) индивидуумы делятся на два пола (гендера)
- Отбор внутри одного гендера, скрещивание – между разными
- Мужские индивидуумы – более высокая вероятность мутации
- Модификация алгоритмов отбора, скрещивания, мутации
- ГГА сходятся дольше (популяция вырождается позже)
- ГГА справляется с многоэкстремальными функциями с явно выраженным экстремумом лучше, чем ГА
- Использовались для разложения спектров на гауссовы составляющие



Кросс-адаптация данных космических аппаратов

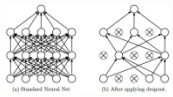
- Использование для прогнозирования методов машинного обучения требует наличия длинных временных рядов из одного источника
- При измерении одних и тех же физических величин на разных КА используются разные приборы с разными характеристиками
- Для совместного использования данных разных КА необходимо научиться адекватно отображать данные из одного домена в другой
- Актуальность связана с переходом с ACE на DSCOVR (параметры солнечного ветра и межпланетного магнитного поля в точке L1 и с GOES-13/15 на GOES 16+ (потоки заряженных частиц на геостационарной орбите).
- Перевод признаков 1 в 1 или все в 1, с погружением или без погружения временного ряда, с помощью линейной регрессии или многослойного персептрона
- Отображение данных ACE и DSCOVR в единый домен позволяет повысить качество прогнозирования

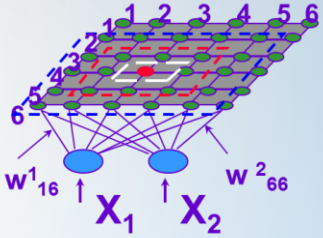
Преподавание машинного обучения

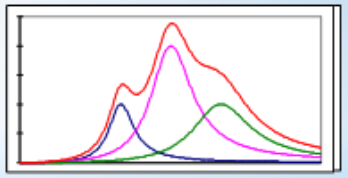
Курс «Машинное обучение в физике» (факультатив для студентов) (осень)

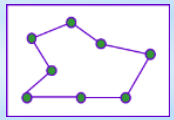
Курс «Машинное обучение. Искусственные нейронные сети и генетические алгоритмы» (курсы повышения квалификации и факультатив для студентов)

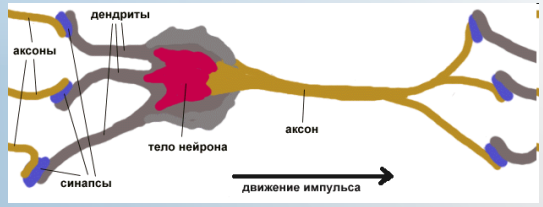
- Искусственные нейронные сети и другие алгоритмы машинного обучения
- Основы предобработки данных
- Генетические алгоритмы, генетическое программирование
- Нечёткая логика, метод группового учёта аргументов
- Современный инструментарий для практической работы по анализу данных
- Практические занятия с использованием языков R и Python
- Самостоятельная работа под руководством преподавателя
- Сайт курсов: http://kpk-nnga.sinp.msu.ru/
- Для студентов, аспирантов и сотрудников МГУ участие бесплатное
- В этом году весенний курс стартовал 20 февраля, прошло 1 занятие

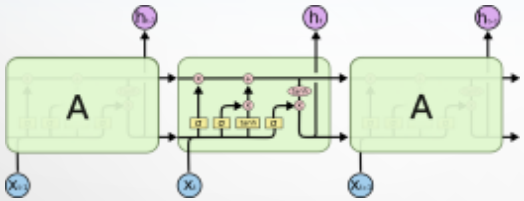












Спасибо за внимание!

Благодарю коллег и соратников:

- И.В.Исаев, В.Р.Широкий, К.А.Лаптинский, И.В.Пластинин, Р.Д.Владимиров – ЛАМОД
- И.М.Гаджиев – аспирант физического факультета МГУ
- А.А.Гуськов, Э.З.Каримов, Ю.М.Витюгова, А.Д.Першин, Н.О.Щуров, Г.А.Куприянов, А.С.Макаров – студенты физического факультета МГУ

- И.Н.Мягкова, О.Г.Баринов, В.В.Калегаев – ЛКФИ ОКН
- С.А.Буриков, Т.А.Доленко, О.С.Сарманова – физический факультет МГУ
- И.Е.Оборнев, Е.А.Родионов, Е.А.Оборнев, М.И.Шимелевич – РГГРУ
- В.В.Кривецкий, М.Н.Румянцева – химический факультет МГУ
- А.В.Самсонович, Д.В.Тихомирова – ИИКС НИЯУ МИФИ