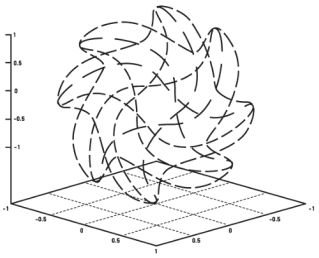

Олимпиада
Национальной
технолøгической инициативы

Карточка профиля

2018/19 учебный год

БОЛЬШИЕ ДАННЫЕ И МАШИННОЕ ОБУЧЕНИЕ

1. Общая информация

Название профиля:

Большие данные и машинное обучение

Научный руководитель - Райгородский Андрей Михайлович, МФТИ, mraigor@yandex.ru
Ведущий методист - Благодарный Евгений Владимирович, МФТИ, blagodarny@phystech.edu

2. Связь профиля с технологическими вызовами и заказами

Область технологий и индустрия, к которой относится данный профиль:

Машинное обучение, анализ больших данных, распознавание образов

Рынки или сквозные технологии НТИ, к которым относится данный профиль:

Сквозная технология, присутствует во всех рынках
В рынке NeuroNet представлена наиболее полно

Технологические барьеры НТИ, которые легли в основу задания данного профиля:

Технологии обработки больших данных, повышающие эффективность по сравнению с существующими подходами на 50% и более по направлениям: качественная оценка эмоционального состояния пользователя или группы пользователей, качественные системы поддержки принятия решений с применением технологий нейромаркетинга.
Обучающиеся алгоритмы на нейронных сетях для анализа больших данных с целью оптимизации процессов на 40% и более (требуется уточнение параметров барьера).
Технологии определения и улучшения потенциала учащихся.

Пример технологического заказа по теме, к которой относится задание профиля:

Реализация рекомендательных систем (рекомендации товаров на Amazon.com, рекомендация музыки на Яндекс.Музыка)
Повышение эффективности системы прогнозирования (ритейл, прогнозирование оттока клиентов)
Анализ и прогнозирование влияния окружающей среды на человека
Обработка естественного языка (синтез и распознавание речи, анализ эмоций по тексту, тематическое моделирование)

3. Практика будущего

Сейчас информационные технологии интегрированы практически во все области жизни, что позволяет собирать и анализировать большие объемы данных практически о всех явлениях общества и окружающей среды.

Пример задания №1 Анализ влияния окружающей среды на здоровье и самочувствие человека

В современном мире очень популярны такие устройства, как пульсометры и фитнес-браслеты. Благодаря доступным ценам и надежности, фитнес-браслеты используются не только спортсменами при тренировках и людьми со слабым здоровьем для постоянного мониторинга состояния, но и обычными людьми в повседневной жизни. Большое количество записей RR-интервалов людей, различающихся возрастом, состоянием здоровья и образом жизни, позволяет анализировать не только особенности индивидуального состояния, но и общие тенденции. В частности, грамотное применение современных методов анализа данных может позволить ответить на такой важных и актуальный вопрос, как влияние окружающей среды.

В качестве задачи будет предложено провести анализ влияния колебаний в ионосфере, длины светового дня, температуры, а также атмосферного давления на самочувствие людей по вариабельности сердечного ритма. Выделение и изучение других важных факторов. Изучить возможности прогнозирования и предупреждения обострения заболеваний.

Пример задания №2 Анализ текстовых потоков новостных лент

В современном мире все большее значение имеет информация. Ключевую роль играет возможность быстро ориентироваться в информационных потоках, отделять достоверную информацию от недостоверной, выделять основные тенденции. Информационное поле влияет, в частности, на ценообразование.

В качестве задачи будет предлагается анализ текстовых статей новостной ленты определенной тематики с целью прогнозирования цен на соответствующие товары. Предполагается не только выделять соответствующие ключевые слова, но и строить модели влияния косвенных факторов на ценообразование.

4. Задача заключительного этапа

Формулировка задания для заключительного этапа:

Задача будет представлять из себя анализ набора данных с целью предсказания некоторых характеристик, примеры ниже:

Пример задания №1 “Анализ новостных потоков”
Анализ текстовых статей новостной ленты с точки зрения сюжетов и тем.
Выделение из новостного потока ключевых событий.

Пример задания №2 “Диагностика заболеваний по ВСР”
Задача определения болезни по записям RR-интервалов (снятых с различных неточных датчиков, в том числе с фитнес-браслетов).

Пример задания №3 “Прогнозирование оттока клиентов”
По статистическим данным сотового оператора определить, какие из клиентов скорее всего скоро уйдут.

Рекомендуемая численность команды школьников и ее предполагаемый состав:

2-3 человека (каждый человек берет несколько обязанностей)

- координатор-администратор, организующий работу детей
- инженер-программист, организующий проверку заданий
- инженер-программист, организующий работу вычислительных комплексов
- математик-программист, формулирующий задания, отвечающий на вопросы детей
- специалист предметной области (экономист, биолог и проч.)

Требование к оборудованию:

Возможность восполнить для каждого участника наличие компьютера (пк, или ноутбук)
Быстрый доступ к удаленному серверу - по одному на команду. GPU (не меньше 8 GB видеопамяти), от 16GB RAM

Требование к программному обеспечению:

На серверах должна быть поставлена операционная система семейства Unix с настроенной CUDA и CUDNN, а также развернут пакет Anaconda

Требование к расходникам:

Отсутствуют

Требования к знаниям, способностям и компетенциям участников (например, темы по школьным предметам или компетенции WorldSkills):

Понимание основ теории вероятностей и математической статистики Знакомство с основными понятиями больших данных и машинного обучения Владение одним из основных языков программирования (Python, R, C#, C++, Java) и соответствующими библиотеками, релевантными для машинного обучения. Чаще всего — Python (numpy, pandas, sklearn, scipy…) Умение использовать и выбирать оптимальные модели в контексте задачи Знание методов оценки моделей (различных метрик) Умение устранять причины переобучения и недообучения Компетенции в предметной области по задаче заключительного этапа

5. Первый и второй отборочные этапы

6.1. Школьные предметы, по которым будет проводиться отбор на первом этапе:

Математика, информатика Информатика (8-11 кл). Предметные: Знание основных структур программирования (условные конструкции, циклы, типы данных, функции и процедуры) Надпредметные: Язык Python и библиотеки Математика (8-11 кл). Предметные: Основные знания из алгебры: векторы, понятие производной, владение операциями над функциями, основы математического анализа Надпредметные: Основы Теории Вероятностей и задачи на построение Алгоритмов Теория Вероятностей https://www.coursera.org/learn/probability-theory-basics, Основый машинного обучения https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie Основы языка Python https://stepik.org/course/512/

Содержание и формат проведения второго этапа (в т.ч. что нового дети узнают в рамках второго этапа в рамках подготовки к финалу):

Подготовка ко второму этапу, а также самии задачи предметных туров по математике и информатике будут подобраны так, чтобы подвести теоретическую и

практическую основу под решение задач анализа данных. Обсудить проблемы комбинаторной оптимизации, вероятности и вычислительной сложности на языке, понятном школьнику. Установить связь этих вопросов со школьной программой.

На втором этапе кроме предметных туров по математике и информатике предполагается провести соревнование по анализу данных.

В рамках подготовки к финалу предполагается провести подготовительный интенсив (3 дня). Его целью является заложить базовую грамотность и развить уже имеющиеся знания по анализу данных как с инженерной, так и с математической точки зрения. Обсудить такие понятия, как недообучение, переобучение и обобщающая способность алгоритмов. Обсудить baseline и state-of-the-art методы.

Перечень открытых соревнований и конкурсов, победители которых могут быть выбраны в качестве участников заключительного этапа без прохождения отборочных этапов:

Кружки МФТИ (Deep Learning School)
Победители и призеры финала олимпиады НТИ прошлых лет по профилю “Большие данные и машинное обучение”
Финалисты и призёры заключительных этапов олимпиад 2 и 3 уровня РСОШ по математике или информатике
Финалисты и призёры отборочного этапа олимпиад 1 уровня РСОШ по математике или информатике
Призёры интенсива по профилю “Большие данные и машинное обучение”
Призёры хакатонов по направлениям “Большие данные”, “Машинное обучение”, “Анализ данных”
Участники проектных смен по направлению “Большие данные” ОЦ “Сириус”
Победители конкурсов анализа данных на kaggle и аналогичных крупных конкурсах
Призёры олимпиад 1-3 уровня РСОШ относительно предметной области задачи текущего года (экономика, биология etc)
Призёры хакатонов относительно предметной области задачи текущего года (экономика, биология etc)

6. Подготовка к финалу

Примеры доступных онлайн-материалов, которые могут быть рекомендованы участникам и ссылки на них:

Материалы олимпиады НТИ 2017/18
Материалы олимпиады “Академия искусственного интеллекта”

Материалы кружков (общее) deepmipt.github.io/dlschl/
Материалы кружков (Deep Learning) https://github.com/deepmipt/dlschl
Материалы кружков (Machine Learning) https://miptmlschool.github.io/

Перечень тематик, по которым будут разработаны и проведены хакатоны по подготовке к заданию заключительного этапа (включая описания заданий и требования к оборудованию):

Большие данные
Машинное обучение
Особенности предметной области финальной задачи
Подготовка к предметному туру олимпиады