Анализ данных

Направление подготовки : 09.04.01 «Информатика и вычислительная техника»
Квалификация, присваиваемая выпускникам : магистр.
Форма обучения : очная.
Нормативный срок освоения : 2 года.

Трудоемкость освоения за весь период обучения составляет 120 зачетных единиц и включает все виды аудиторной и самостоятельной работы обучающегося, практики, время, отводимое на контроль качества освоения обучающимся образовательной программы.

Сведения об образовательной программе

Магистратура по направлению «Анализ данных» готовит востребованных в IT-индустрии специалистов новой профессии Data Scientist. В процессе обучения студенты осваивают современные методы хранения, обработки и анализа данных и получают опыт работы над реальными задачами в самых различных приложениях: от диалоговых систем до компьютерного зрения. Сочетание глубоких теоретических знаний и большого количества практики позволяет им не только эффективно использовать самые современные методы анализа данных, но и создавать новые.

В основе обучения лежит углубленное изучение математики и современных приёмов программирования, которые необходимы для успешного освоения актуальных методов работы с информацией на всех этапах: поиск, обработка, хранение и передача. В процессе обучения студенты учатся решать прикладные задачи, такие как распознавание образов, машинный перевод, работают с большими данными, обучают нейросети.

Преподаватели кафедры, совмещающие научную деятельность с работой в Яндексе, стремятся воспитать специалистов с широкими компетенциями в области работы с информацией, востребованных на рынке труда и в компании Яндекс.

Обучение в магистратуре на кафедре «Анализ данных» включает в себя обязательные курсы в МФТИ, учебу в ШАД, научно-исследовательскую работу по тематике кафедры, а также участие в одном из научных семинаров Школы:

- Компьютерные науки
- Machine Intellegence
- Методы анализа текстов
- Reinforcement learning
- Байесовские методы в машинном обучении

Основные направления образовательной и научной деятельности магистратуры:

- Машинное обучение и информационный поиск
- Алгоритмы и большие данные
- Компьютерное зрение
- Дискретная математика и теория оптимизации
- Компьютерная лингвистика

Сведения о реализации образовательной программы

Образовательный процесс осуществляется на кафедре анализа данных, заведующий кафедрой д.ф.-м.н., генеральный директор ООО «Яндекс» Елена Игоревна Бунина. Основной партнёр магистерской программы – компания «Яндекс».

Магистранты кафедры «Анализ данных» имеют возможность во время обучения или после него пройти стажировку или практику в компании «Яндекс», а также продолжить своё обучение в рамках аспирантуры.

Дисциплины учебного плана

Алгоритмы и структуры данных поиска

Курс дает базовые знания в области алгоритмов и структур данных, которые важны для понимания работы библиотек, алгоритмов и языков программирования. На лекциях студенты получают необходимую теоретическую базу, семинары содержат советы по написанию кода и реализации конкретных алгоритмов, а также разборы задач, которые показывают применения и скрытые возможности пройденных структур данных.

Домашние задания по курсу закрепляют полученные знания и воспитывают хороший стиль написания кода, который позволяет избежать стандартных, но от этого ничуть не менее распространенных даже у опытных разработчиков, ошибок.

Восстановление зависимостей с использованием эмпирических данных

В курсе изучаются вопросы восстановления функциональных закономерностей по данным наблюдений. Примерами задач из этой области являются обнаружение брачных аферистов на сайтах знакомств, прогноз риска развития заболевания на основании генетических признаков, предсказание землетрясений и т.д. Изучение курса концентрируется на трёх фундаментальных вопросах:

1. Какую зависимость восстанавливать
2. Как это сделать

3. Насколько хорошо получилось

Рассматриваются различные постановки задачи восстановления зависимостей: задача классификации (распознавания образов), задача регрессии, обратные задачи интерпретации косвенных наблюдений. При решении этих задач будут обосновываться и использоваться современные методы машинного обучения: "kernel trick", регуляризация, SVM и SVR, методы выбора моделей.

Основы стохастики. Стохастические модели

Курс позволяет студентам ознакомиться с основными типами стохастических процессов и овладеть всеми необходимыми инструментами для того, чтобы разрабатывать методы скорейшего обнаружения разладки и детектирования аномалий и использовать эти методы в задачах мониторинга состояния различных систем.

Обучение машин: доп. главы

Курс является продолжением курса «Восстановление зависимостей по эмпирическим данным», но в нём рассматривается другой класс восстановления зависимостей - решение обратных задач интерпретации косвенного эксперимента. Проблемы, возникающие при решении задач построения регрессии, существенно превосходят проблемы построения регрессии, а методы их решения (регуляризация) существенно обогащают и МНК, и методы классификации. Кроме методов решения обратных задач в рамках курса проходит знакомство с байесовскими методами восстановления зависимостей, проблемой выбора моделей, методом восстановления случайного поля - методом кригинга, методом конформных предикторов, применение ядерного подхода во всех этих задачах.

Алгоритмы для работы с большими объёмами данных

В данном курсе изучаются алгоритмы для работы с данными, которые не помещаются в оперативную память компьютера. В таком случае необходимо учитывать не только затраты CPU алгоритмы, но и количество и характер обращений к данным.

В курсе изучаются следующие области:
1. Алгоритмы во внешней памяти
2. Cache-oblivious алгоритмы
3. Алгоритмы потоковой обработки данных

Анализ изображений и видео

Курс посвящен методам и алгоритмам компьютерного зрения, т.е. извлечения информации из изображений и видео. В курсе рассматриваются в основном методы анализа отдельных изображений, основы обработки изображений (шумоподавление, тональную коррекцию, выделение краёв), классификации изображений (основные признаки), поиск изображений по содержанию (сжатие дескрипторов, приближенные методы сравнения дескрипторов), распознаванию лиц, нейросетевые модели (deep learning) для решения всех перечисленных задач.

Вероятностное моделирование статистических данных и их анализ

За последние двадцать лет существенно возросла потребность в решении ряда практических задач, таких как автоматическое обнаружение исправностей (разладок, сбоев, и т.п.), обслуживание оборудования на основе автоматического контроля его состояния, мониторинг в биомедицине и финансовой сфере и др. Основная черта вышеперечисленных задач состоит в том, что, по сути, все они сводятся к выявлению момента резкого изменения (разладки) некоторых характеристик рассматриваемого объекта на основе статистических данных о других характеристиках этого объекта и/или оценивании неизвестных характеристик объекта по известным в режиме реального времени (фильтрация). С развитием информатики появилась возможность построения автоматизированных информационных систем для статистической обработки огромного объема реальных данных с целью вынесения тех или иных суждений об истинных характеристиках процесса. Для создания таких систем с привлечением программных средств требуется прежде всего разработка соответствующих фундаментальных математических методов обработки поступающей и поступившей информации исходя из естественных критериев оптимальности, именно это и изучается в рамках курса.

Выпуклый анализ и оптимизация Базовый курс по моделям

и методам выпуклой оптимизации.

Курс состоит из трех частей:

- - вводная теоретическая часть (выпуклые множества и функции, условия оптимальности в задачах оптимизации, теория двойственности);
- - методы оптимизации (быстрые градиентные методы, прямо-двойственные методы, методы внутренней точки, метод модифицированной функции Лагранжа и ADMM, метод условного градиента и проекции градиента);

- . - оптимизация и неопределенность: этом разделе будут обсуждаться как задачи, включающие неопределенность, и подходы для их решения (стохастическая оптимизация, робастная оптимизация, онлайн оптимизация), так и использование рандомизации для построения новых методов численного решения задач оптимизации без неопределенности (рандомизированные методы первого порядка).

Информационный поиск

В курсе рассматриваются общие вопросы построения информационно-поисковых систем: задачи информационного поиска и архитектура поисковых систем, машинное обучение в поиске и компьютерная лингвистика, построение поискового индекса и обнаружение дубликатов, поисковый робот и оценка качества. Практическая сторона курса связана со знакомством с широким спектром технологий и алгоритмов, применяемых на практике при построении компонентов поисковой системы.

Машинный перевод

Данный курс освещает некоторые из основных и наиболее актуальных алгоритмов и моделей, используемых для построения крупномасштабных систем машинного перевода данных. Курс охватывает ключевые компоненты перевода на основе фраз и нейронных сетей, с которыми студенты имеют возможность поэкспериментировать.

Параллельные и распределенные вычисления: доп. главы

Целями курса являются знакомство с параллельными и распределенными вычислениями, различными классами высокопроизводительных систем, принципами реализации параллельных алгоритмов и используемыми моделями программирования, а также получение навыков практического использования соответствующих технологий и систем при решении прикладных задач. В рамках курса предусмотрены домашние задания, включающие написание параллельных программ и работу на вычислительном кластере.

Теория информации

В науке не существует единого подхода к определению понятия информации. В разных областях это понятие трактуется по-разному. Имеются информация по Хартли, энтропия Шеннона, Колмогоровская сложность, коммуникационная сложность. Каждое из этих понятий отражает некоторую грань интуитивного

понятия информации. В курсе будет рассказано об этих понятиях и как они применяются в решении разных задач.