ПРОГРАМА НАВЧАННЯ

2023

BIG DATA LAB

vodafone

Програма навчання може змінюватись протягом курсу

Зміст

Підготовчий курс

Основи математики для аналітиків
(Math Basics for Analytics) 3

Python для початківців
(Python for beginners) 5

Основний курс

Основи програмування та інженерії
(Programming and Engineering Basics) 7

Основи математичного аналізу для науки про дані
(Calculus fundamentals for Data Science) 8

Основи лінійної алгебри, методів оптимізації, теорії ймовірностей та математичної статистики для науки про дані
(Linear Algebra, Optimization Methods, Probability Theory and Mathematical Statistics fundamentals for Data Science) 10

Зберігання даних, їх аналіз та візуалізація
(Data warehousing, business intelligence and data visualization) 13

Машинне навчання та інтелектуальний аналіз даних
(Machine Learning and Data Mining) 15

Python для машинного навчання
(Python for Machine Learning) 18

Технології та інфраструктура Великих Даних для аналізу даних
(BDIT 4DA - Big Data Infrastructure and Technologies for Data Analytics 19

Нейронні мережі та Глибоке навчання
(Neural networks and Deep Learning) 23

Прогнозний аналіз і часові ряди
(Predictive Analytics and Time Series Analysis) 24

ПІДГОТОВЧИЙ КУРС

«Основи математики для аналітиків»
«Math Basics for Analytics»

Лектор: Ставицький Андрій Володимирович, доктор економічних наук, доцент.

Тривалість курсу: 24 аудиторні години
4 лекційні занять
11 практичних занять
1 екзаменаційна контрольна
Одне заняття - 1,5 години

Програма курсу:

Заняття 1 (лекція). Вступ.
- Задача дисципліни
- Організаційні задачі

Заняття 2 (лекція). Основні математичні перетворення.
- Розрахункові задачі
- Алгебраїчні перетворення
- Тригонометричні перетворення
- Логарифмічні перетворення
- Розв'язання рівнянь
- Розв'язання систем рівнянь

Заняття 3 (практичне заняття). Розв’язання задач.
- Розв’язання практичних завдань (Р1)
- Самостійна робота

Заняття 4 (практичне заняття). Розв’язання задач.
- Розв’язання практичних завдань (Р2)
- Самостійна робота

Заняття 5 (практичне заняття). Розв’язання задач.
- Розв’язання практичних завдань (Р3)
- Самостійна робота

Заняття 6 (практичне заняття). Розв’язання задач.
- Розв’язання практичних завдань (Р3)

Заняття 7 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р6)
- Самостійна робота

Заняття 8 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р6)
- Самостійна робота

Заняття 10 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р7)
- Самостійна робота

Заняття 11 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р8)
- Самостійна робота

Заняття 11 (лекція). Похідна.
- Визначення похідної
- Застосування похідної

Заняття 12 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р15)
- Самостійна робота

Заняття 13 (лекція). Інтеграли.
- Невизначені інтеграли
- Визначені інтеграли
- Способи інтегрування

Заняття 14 (практичне заняття). Розв'язання задач.
- Розв'язання практичних завдань (Р15)
- Самостійна робота

Заняття 15 (Підсумкова контрольна робота).
- Розв'язання задач

Заняття 16 (практичне заняття). Розв'язання задач.
- Розбір підсумкової роботи
- Самостійна робота

«Python для початківців»
«Python for beginners»

Лектор: Софія Швець. Data scientist, Vodafone Ukraine

Тривалість курсу: 24 аудиторні години
2 - лекційних заняття
6 - лекційно-практичних занять

Програма курсу:

Модуль 1.
- Введення в Python: переваги та недоліки мови, середовище розробки, встановлення та налаштування, бібліотеки та менеджер пакетів, синтаксис мови та основні типи даних, динамічна типізація, ключові слова та вбудовані функції, перша програма. Самостійна робота.

Модуль 2.
- Числові типи даних та операції над числами. Перетворення чисел. Логічний тип boolean та логічні операції. Самостійна робота.

Модуль 3.
- Умовні оператори та цикли. Інструкції break, continue, pass. Обробка виключень. Самостійна робота.

Модуль 4.
- Список (List), індекси та зрізи. Методи, функції та операції зі списками. Генератори та ітератори. Самостійна робота.

Модуль 5.
- Рядок (String), функції та методи. Екрановані послідовності та форматування. Множина (Set) та операції над множинами. Самостійна робота.

Модуль 6.
- Кортеж (Tuple), функції та методи. Словник (Dictionary), функції та методи. Самостійна робота.

Модуль 7.
- Функції. Область видимості змінних. Знайомство з бібліотеками візуалізації matplotlib, seaborn, plotly. Самостійна робота.

Модуль 8.
- Основи ООП: класи, конструктори, успадкування, інкапсуляція, поліморфізм, декоратори.

ОСНОВНИЙ КУРС

«Основи програмування та інженерії»
«Programming and Engineering Basics»

Лектор: Сергій Одрібець. Data scientist, Vodafone Ukraine

Тривалість курсу: 21 аудиторна година
5 - лекційних занять
7 - практичних занять
2 - лекційно практичних заняття
1 - контрольна робота/код рев'ю

Програма курсу:

Заняття 1 (лекція).
- Введення в Python. Встановлення python. IDE для python. Встановлення бібліотек.
- Python virtual Environments. Написання першої програми

Заняття 2 (практичне заняття)
- Типи та структури даних: Числові, текстові. Словники, списки та кортежи в Python

Заняття 3 (лекція)
- Оператори та вирази в Python: If, elif, for, while, break, continue, else.

Заняття 4 (практичне заняття)
- Функції в Python. Лямбда функції. Рекурсія. Декоратори функцій

Заняття 5 (практичне заняття)
- Модулі та пакети в Python. Генератор та ітератори

Заняття 6 (лекція).
- Введення в концепції ООР. Побудува класів в python. Абстракція, інкапсуляція

Заняття 6 (практичне заняття)
- ООР. Наслідування, поліморфізм в Python. Шаблони проектування. Виключення в python

Заняття 7 (практичне заняття)
- Модуль Numpy

«Основи математичного аналізу для науки про дані» («Calculus fundamentals for Data Science»)

Лектор: Чертов Олег Романович, доктор технічних наук, професор

Тривалість курсу: 12 аудиторних годин
1 лекційне заняття
6 практичних занять
1 екзаменаційна контрольна робота
Тривалість кожного заняття: 1,5 години.

Програма курсу:

Заняття 1 (лекція).
Вступ.

- Строгість, точність та абстрактність математики.
- Логічна символіка (висловлення, логічні зв'язки, квантори).
- Множини, операції над ними.
- Бінарні відношення. Відношення еквівалентності. Клас еквівалентності.
- Відношення порядку. Обмежена множина. Супремум та інфімум. Максимум та мінімум.
- Відображення (функція). Бієкція. Обернене відображення. Композиція функцій. Послідовність.

Заняття 2 (практичне заняття).
Математична індукція. Границі послідовностей.

- Метод математичної індукції, його застосування до розв'язання задач.
- Означення границі послідовності, її геометрична інтерпретація.
- Пошук границь послідовностей.

Заняття 3 (практичне заняття).
Границі функцій.

- Означення границі функції, її геометрична інтерпретація.
- Пошук границь функцій.
- Неперервні функції. Точки розриву функцій.

Заняття 4 (практичне заняття).
Похідна. Дослідження функцій.

- Означення похідної функції однієї змінної, її геометрична інтерпретація.
- Пошук похідних та диференціалів (у тому числі вищих порядків) функцій однієї змінної, функцій заданих параметрично і заданих неявно, складеної та оберненої функції.
- Дослідження функцій однієї змінної на екстремуми та опуклість, побудова графіків функцій однієї змінної.

Заняття 5 (практичне заняття).
Невизначені та визначені інтеграли.

- Властивості невизначених інтегралів. Внесення під знак диференціалу. Інтегрування частинами.
- Знаходження первісних раціональних, ірраціональних та трансцендентних функцій.
- Обчислення визначених інтегралів від функції однієї змінної, їх застосування до обчислення довжин кривих і площ фігур.

Заняття 6 (практичне заняття).
Ряди Тейлора та Маклорена. Числові та функціональні ряди.

- Побудова поліномів Тейлора і Маклорена для функцій однієї змінної і використання їх до наближення функцій.
- Поточкова та рівномірна збіжність рядів. Дослідження на збіжність числових та функціональних рядів.
- Знаходження області та радіусу збіжності степеневих рядів.

Заняття 7 (практичне заняття).
Частинні похідні та диференціал функції багатьох змінних. Градієнт.

- Знаходження частинних похідних та диференціалів функції багатьох змінних, градієнту функції.
- Поняття локального та умовного екстремуму функцій багатьох змінних.

Заняття 8 (екзамен).
Підсумкова контрольна робота.

- Шість задач/вправ по темам, що були засвоєні/повторені під час курсу.

«Основи лінійної алгебри, методів оптимізації, теорії ймовірностей та математичної статистики для науки про дані» («Linear Algebra, Optimization Methods, Probability Theory and Mathematical Statistics fundamentals for Data Science»)

Лектор: Ставицький Андрій Володимирович, доктор економічних наук, доцент

Тривалість курсу: 24 аудиторні години
7 лекційних занять
8 практичних занять
1 екзаменаційна контрольна робота
Тривалість кожного заняття - 1,5 години.

Програма курсу:

Заняття 1 (лекція).
Вступ до лінійної алгебри.

- Вектори
- Матриці
- Дії з матрицями
- Знаходження визначника матриці

Заняття 2 (практичне заняття).
Робота з матрицями.

- Розв'язання задач

Заняття 3 (лекція).
Спеціальні типи матриць.

- Власні значення та власні вектори
- Симетричні матриці та ортогональна діагоналізація
- Застосування власних значень та власних векторів
- Аналіз основних компонентів
- Складні дії з матрицями

Заняття 4 (практичне заняття).
Визначення власних значень та векторів.

- Застосування власних значень та власних векторів
- Розв'язання практичних завдань
- Самостійна робота

Заняття 5 (лекція).
Огляд методів оптимізації.

- Методи одновимірної мінімізації
- Методи багатовимірної мінімізації
- Основні завдання динамічного програмування

Заняття 6 (практичне заняття).
Розв'язання типових задач оптимізації.

- Розв'язання практичних завдань
- Самостійна робота

Заняття 7 (практичне заняття).
Огляд методів комбінаторики.

- Факторіал
- Комбінації, перестановки, розміщення
- Розв'язання практичних завдань
- Самостійна робота

Заняття 8 (лекція).
Теорія ймовірностей.

- Робота з множинами
- Ймовірність та різновиди її знаходження
- Формула умовної ймовірності
- Формула повної ймовірності
- Формула Байєса

Заняття 9 (практичне заняття).
Визначення ймовірності на практиці.

- Робота з простором подій
- Розв'язання практичних завдань визначення ймовірності подій
- Самостійна робота

Заняття 10 (лекція).
Розподіли випадкових величин.

- Випадкова величина
- Математичне сподівання та дисперсія
- Формула Бернуллі та біноміальний розподіл
- Розподіл Пуассона
- Нормальний розподіл

Заняття 11 (практичне заняття).
Застосування розподілів випадкових величин на практиці.

- Розв'язання практичних завдань

Заняття 12 (лекція).
Описова статистика.

- Вибіркове середнє, медіана
- Вибіркова дисперсія
- Інтервальні та дискретні вибірки

Заняття 13 (практичне заняття).
Математична статистика.

- Визначення параметрів випадкової величини
- Побудова гістограми, емпіричної функції розподілу
- Самостійна робота

Заняття 14 (лекція).
Перевірка статистичних гіпотез.

- Види гіпотез та етапи їх перевірки
- Правила перевірки гіпотез

Заняття 15 (практичне заняття).
Перевірка статистичних гіпотез.

- Гіпотеза про значення вибіркового середнього вибірки
- Гіпотеза про різницю вибіркових середніх вибірок
- Гіпотеза про значення дисперсії вибірки
- Гіпотеза про рівність дисперсій вибірок

Заняття 16 (Підсумкова контрольна робота). Іспит.

- Розв'язання задач

«Нейронні мережі та Глибоке навчання»
«Neural networks and Deep Learning»

Лектор: Вероніка Тамайо Флорес, MSc in Business Analytics and Big Data

Тривалість курсу: 27 аудиторних годин
3 лекційних заняття
5 занять з елементами практичної роботи та одна презентація групових проектів
Тривалість одного заняття - 3 години.

Програма курсу:

Заняття 1.
Вступ.
- Основи менеджменту даних та його роль в сучасній організації
- Типи даних в програмуванні vs. типи даних в бізнесі
- Бази даних та системи управління базами даних (DBMS)
- Огляд SQL та NoSQL баз, їх переваги та недоліки. Ключові вендори

Заняття 2.
Моделі даних.
- Моделі даних
- Модель «сутність-зв'язок» (ER-модель)
- Нормалізація даних
- Практичне завдання: нормалізація даних
- Огляд індивідуального проекту

Заняття 3.
Системи збору та зберігання даних.
- Огляд основних типів архітектури для систем збору та зберігання даних
- Словник даних (data dictionary)
- Роль DBA та основи адміністрації баз даних
- Практичне завдання: побудова реляційної схеми даних
- Огляд групового проекту

Заняття 4.
SQL як мова управління даними.

- Три основні складові SQL (DBMS languages)
- Огляд базових функцій SQL
- Об'єднання даних з різних таблиць в одному запиті: JOINs
- Практичне завдання: побудова SQL-запитів

Заняття 5.
Master Data Management.

- Аудиторний тест
- Основи Master Data Management
- Дата трансформація компанії
- Життєвий цикл даних

Заняття 6.
Рух даних в компанії (data pipeline).

- Data pipeline
- ETL та ELT процеси
- Практичне завдання: побудова ETL процесу
- Здача індивідуального завдання

Заняття 7.
Аналітика даних (business intelligence).

- Огляд цілей та задач business intelligence
- OLAP vs. OLTP куби. Що це і коли їх доречно використовувати
- Огляд програм для business intelligence (лідери Gartner's Magic Quadrant)

Заняття 8.
Візуалізація даних та data storytelling.

- Основні помилки та найкращі практики візуалізації даних
- Storytelling на основі даних
- Практичне завдання: визначення цілей та задач візуалізації
- Практичне завдання: створення дашбордів

Заняття 9.
Презентація групових проектів та їх обговорення.

«Машинне навчання та інтелектуальний аналіз даних» («Machine Learning and Data Mining»)

Лектор: Чертов Олег Романович, доктор технічних наук, професор

Тривалість курсу: 33 аудиторні години
16 лекційно-семінарських занять по 1.5 (астрономічні) години кожне
5 практичних занять + 1 екзамен по 1.5 (астрономічні) години

Програма курсу:

Заняття 1.
Вступ (мотивація).

- Визначення машинного навчання. Приклади задач різних типів навчання. Базова схема машинного навчання.

Заняття 2-3.

- Класифікаційні метрики: матриця помилок (Confusion matrix) і помилки першого та другого роду, доля правильних відповідей (Accuracy), точність (Precision) і повнота (Recall), F-міри (F-Scores), коефіцієнт кореляції Метьюса (Matthews Correlation Coefficient, MCC), Precision-Recall крива (curve), ROC («reciever operating characteristic»)-крива, Area under curve (AUC), коефіцієнт Джині (Gini coefficient). Бінарний та багатокласовий випадки.

Заняття 4-5.

- Дерева прийняття рішень (Decision Trees). Ентропія Шеннона. Класифікація з дискретними та неперервними атрибутами. Виділення правил.

Заняття 6.

- Бінарні (CART) та небінарні (ID3, CHAID) види дерев. Критерії розщеплення: приріст інформації (Information Gain), нормалізований приріст інформації (Information Gain ratio), неоднорідність Джині (Gini impurity), помилка класифікації (Classification error), критерій Хі-квадрат (Chi-square test statistic for independence). Обрізання дерев (pre-pruning, post-pruning). Обробка пропущених значень, паралелізація.

Заняття 7.

- Метричні методи класифікації та регресії (Distance-Based Classification & Regression Methods). Жадібні та ліниві алгоритми (Eager vs Lazy learners). Гіпотези компактності та неперервності. Алгоритм k найближчих сусідів. Підбір вагів. Діаграма Вороного. Шумові характеристики. Нормалізація характеристик. Метрика Мінковського, косинусна міра, відстань Жаккарда, відстань Хемінга. Точні та наближені методи пошуку найближчих сусідів.

Заняття 8.

- Реалізація та дослідження методів класифікації та регресії (дерева прийняття рішень та метричні методи). Виконання індивідуальних домашніх завдань з реалізації та дослідження методів класифікації та регресії, презентація, обговорення і захист результатів.

Заняття 9.

- Лінійні моделі класифікації та регресії. Метод мінімізації емпіричного ризику. Поняття відступу (margin). Регресія зі штрафом. Різні види регуляризації. Сценарії P-більше-N. Логістична регресія. Незбалансовані класи. Регресійні метрики.

Заняття 10.

- Баєсівські методи (Bayes methods). Безумовна (апріорна) та умовна (апостеріорна) ймовірність. Теорема Баєса. Наївний Баєсів алгоритм (Naive Bayes Algorithm). Згладжування Лапласа.

Заняття 11.

- Реалізація та дослідження методів класифікації та регресії (логістична регресія та Баєсівські методи). Виконання індивідуальних домашніх завдань з реалізації та дослідження методів класифікації та регресії, презентація, обговорення і захист результатів.

Заняття 12.

- Кластеризація (Clustering). Типи кластерних алгоритмів: пласкі алгоритми (метод K-means) та алгоритми ієрархічної кластеризації (агломеративні та розділяючі). Жорстка кластеризація та м'яка (нечітка) кластеризація (метод C-means). Визначення кількості кластерів. Близькість кластерів. Моделі, засновані на щільності (DBSCAN і OPTICS).

Заняття 13.

- Реалізація та дослідження методів кластеризації. Виконання індивідуальних домашніх завдань з реалізації та дослідження методів кластеризації, презентація, обговорення і захист результатів.

Заняття 14.

- Відбір та синтез інформативних характеристик або зниження розмірності (Feature selection & extraction): відбір інформативних характеристик (feature selection), синтез інформативних характеристик (feature extraction), методи, вбудовані в модель навчання (embedded models). (Нормована) взаємна інформація. Метод головних компонент (principal component analysis, PCA). Метод незалежних компонент (independent component analysis, ICA). Лінійний дискримінантний аналіз (linear discriminant analysis, LDA).

Заняття 15.

- Реалізація та дослідження методів зниження розмірності. Виконання індивідуальних домашніх завдань з реалізації та дослідження методів відбору та/чи конструювання характеристик, презентація, обговорення і захист результатів.

Заняття 16.

- Комбінації моделей (Ensemble methods). Бутстреп-агрегування (Bootstrap method) або беггінг (bagging). Випадковий ліс (Random Forest). Boosting (підсилення). Алгоритм AdaBoost (адаптивне підсилення).

Заняття 17.

- Навчання з підкріпленням (Reinforced Learning). Базова модель навчання з підкріпленням. Марковський процес. Рівняння Беллмана. Алгоритм Q-навчання (Q-Learning).

Заняття 18.

- Реалізація та дослідження методів навчання з підкріпленням. Виконання індивідуальних домашніх завдань з реалізації та дослідження методів навчання з підкріпленням, презентація, обговорення і захист результатів.

Заняття 19.

- Асоціативні правила (Association rules). Виявлення елементів, що часто зустрічаються. Властивість антимонотонності підтримки. Замкнені набори елементів (Closed item sets). Рівень довіри (confidence). Алгоритм A-Priori. Алгоритм популярних наборів (Frequent-Pattern, FP-Growth). Підвищення ефективності обробки популярних наборів: метод локальних предметних наборів, семплінг.

Заняття 20.

- Причинно-наслідкове навчання (Causal Learning). Кореляція ≠ Причинність. Виявлення причинно-наслідкового зв’язку. Показник сили причинно-наслідкового зв’язку (Predictive Power Score, PPS).

Заняття 21.

- Навчання з гарантованою точністю (Conformal Learning). Класифікація та регресія із гарантованим навчанням (Conformal classification & regression). Функції невідповідності (Non-conformity functions). Застосування для випадку незбалансованих класів.

Заняття 22.

- Екзамен (Midterms).

«Python для машинного навчання»
«Python for Machine Learning»

Лектор: Сергій Одрібець. Data scientist, Vodafone Ukraine

Тривалість курсу: 18 аудиторних години
1 - лекційне заняття
8 - практичних занять
2 - лекційно практичних заняття
1 - контрольна робота/код рев'ю

Програма курсу:

Модуль 1 (лекція).
- Модуль Pandas. Обробка та аналіз даних. Pandas profiling та sweetviz

Модуль 2 (практичне заняття)
- Візуалізація даних в python: matplotlib, seaborn, plotly

Модуль 3 (практичне заняття)
- Scikit-learn and features engineering, побудова першої моделі

Модуль 4 (практичне заняття)
- Pipeline навчання моделей (Sklearn моделі, XGBoost, lightgbm...)

Модуль 5 (практичне заняття)
- Налаштування hyperparameters(GridSearch, RandomSearch, HyperOpt)

Модуль 6 (практичне заняття)
- Аналіз результатів моделювання. Вибір найкращої моделі

Модуль 7 (практичне заняття)
- Налаштування моделі в «prod» за допомогою flask web service. Застосування airflow

Модуль 8 (контрольна робота/код рев'ю)
- Code review

«Технології і інфраструктура Великих Даних для аналізу даних» (BDIT4DA - Big Data Infrastructure and Technologies for Data Analytics)

*Заплановані зміни в програмі. Модуль в процесі модернізації.

Лектор: Демченко Юрій Володимирович, Кандидат технічних наук

Тривалість курсу: 48 аудиторних годин
14 лекційних занять по 2 (астрономічні) години кожне
14 практичних заняття по 1 (астрономічній) годині кожне.

- Individual assignment (30%).
- Group Project (70%): including literature study 20%, project design 20%, dataflow and test application 30%.

Програма курсу:

Лекція 01
Вступ до курсу. Основи хмарних технологій.

- Моделі хмарних сервісів, хмарні ресурси, функціонування хмарних служб, багатомешканність. Віртуальний гібридний/ динамічний хмарний центр обробки даних та хмарний аутсорсинг IT-інфраструктури підприємства. Хмарні випадки використання та сценарії для підприємств.

Практика 01.

- Початок роботи з хмарою Amazon Web Services (AWS); огляд хмарних сервісів, розгортання та доступ до екземплярів EC2, S3, VM

Лекція 02
Хмарні сервіси управління даними та програмами

- Хмарні послуги зберігання даних та бази даних. Хмарний моніторинг та розгортання програм. Економічна модель хмарних сервісів та ціноутворення.

Практика 02.

- Продовження вивчення основних хмарних сервісів; інструменти моніторингу хмарних сервісів. Створення учбового веб-сайта на основі EC2 та S3 (додатково).

Лекція 03
Хмарі технології та Великі Дані, інфраструктура Великих Даних та її компоненти

- Єталонна архітектура Великих Даних та приклади використання. Функціональні компоненти та ролі. Важливі проекти та тенденції. Великі Дані в промисловості та концепція Industry 4.0. Введення до MapReduce та Hadoop.

Практика 03.

- Хмарні сервіси AWS для роботи з даними; хмарне зберігання даних та сервіси баз даних; сервіси баз даних AWS: RDS та Aurora. Запуск простих скриптів SQL.

Лекція 04

Хмарні платформи для великих даних. Огляд та порівняння.

- Огляд основних хмарних платформ Великих Даних: Amazon Web Services (AWS), Microsoft Azure, Google Cloud Platform (GCP)

Практика 04.

- Ознайомлення з хмарними сервісами AWS для Big Data; Elastic MapReduce (EMR), Data pipeline, Machine Learning.
- **Завдання 1.** Звіт по практичній роботі з сервісами AWS і відповідь на тестові питання
- **Проєкт:** Створення проектних груп, визначення теми проекту, включаючи профіль уявного підприємства, керованого даними.

Лекція 05

Екосистема Hadoop для Великих Даних.

- Компоненти екосистеми Hadoop: HDFS, HBase, MapReduce, Spark, Kafka, YARN та інші.

Практика 05.

- Початок роботи з хмарним кластером Hadoop (наприклад, AWS Elastic Map Reduce (EMR) або Azure HDInsight), ознайомлення з інтерфейсом Hue, завантаження даних і файлів.

Лекція 06

Бази даних SQL і NoSQL (1)

- Основи SQL (повторення з курсу по базам даних та SQL). Типи та огляд баз даних NoSQL. Теорема CAP і бази даних NoSQL, властивості ACID та BASE

Практика 06.

- Установка та налаштування автономного кластера Hadoop для особистого користування Cloudera Quickstart 5.13 або Hortonworks Sandbox.
- **Завдання 2.** Виконання простого завдання з MapReduce, наприклад підрахунок слів; виконати простий алгоритм ранжирування (вручну або за допомогою бібліотеки Java MapReduce).
- **Проєкт:** Робота над груповим проектом: Визначення корпоративної інфраструктури Великих Даних та потоку обробки даних.

Лекція 07

Бази даних SQL і NoSQL (2)

- Бази даних на основі стовпців та їх використання (наприклад, HBase). Сучасні великомасштабні бази даних AWS Aurora, Azure CosmosDB, Google Spanner.

Практика 07.

- До початку роботи з Hive та Pig Latin: оновлення знань з основних команд SQL, робота з базами даних SQL.

Лекція 08

Потоки даних та потоковий аналіз даних.

- Потоки даних та потоковий аналіз даних. Архітектура та компоненти Spark. Популярні платформи для Spark, DataBricks. Програмування для Spark.

Практика 08.

- Початок роботи з Hive: Виконання простих запитів; введення та перевірка даних; запуск простих запитів на наборах даних; перевірка і візуалізація вихідних даних.
- **Завдання 3.** Обробка наборів великих даних за допомогою скриптової мови Hive (HiveQL)
  Робота над груповим проектом: структури даних, програми та послуги з аналізу даних, вибір інфраструктурних послуг.

Лекція 09

Відкриті дані, доступні набори даних, збір даних, WebAPI для збору даних

- Концепція Відкритих Даних (Open Data), набори даних для аналізу даних та машинного навчання. WebAPI

Практика 09.

- Початок роботи з Pig Latin: написання простих скриптів/програм потової обробки даних.

Лекція 10

Управління великими даними та управління ними (1)

- Архітектура великих даних підприємства та управління великими масштабами. Структури даних, метадані. Хмарні послуги сховищ даних, Data Lakes. Модель зрілості управління даними та аспекти забезпечення якості даних.

Практика 10.

- Продовження роботи з Pig: Запуск простих запитів на наборах даних; перевірка вихідних даних.
- **Завдання 4.** Використання скриптової мови Pig для програмування робочих процесів для Великих Даних.
- **Проєкт:** Продовження групового проекту щодо інфраструктури Великих Даних підприємства: Вибір інфраструктурних послуг та компонентів

Лекція 11

Управління великими даними та управління ними (2)

- План управління даними (DMP). Принципи ефективності даних FAIR (Findable – Accessible – Interoperable – Reusable). Управління даними в промисловості, забезпечення довіри до даних та суверенітету даних. Модель опорної архітектури промислових IDSA RAM.

Практика 11.

- План управління даними (DMP) для групового проекту, управління метаданими (робота в групах)
- **Проєкт:** Продовження групового проекту щодо інфраструктури Великих Даних підприємства: Вибір інфраструктурних послуг та компонентів.

Лекція 12

Безпека та відповідність великих даних (1)

- Проблеми безпеки Великих Даних, захист даних. Контроль доступу та управління ідентичністю, переваги федеративної інфраструктури контролю доступу, хмарні послуги.

Практика 12.

- Продовження роботи з Планом управління даними (DMP) для групового проекту, управління метаданими (робота в групах)

Лекція 13

Безпека та відповідність великих даних (2)

- Хмарні стандарти відповідності та оцінка послуг провайдера хмарних послуг. Оцінка відповідності провайдера згідно анкети Consensus Assessment Initiative Questionnaire (CSA CAIQ

Практика 13.

- Оцінка відповідності і вимоги до інфрастуктури групового проекту (робота в групах)

Лекція 14

Інструменти управління великими даними та інструменти DataOps

- Управління проектами щодо Великих Даних. Управління життєвим циклом даних та DataOps.

Практика 14.

- Продовження оцінки відповідності і вимоги до інфрастуктури групового проекту. Обговорення вимог та рішень щодо безпеки та відповідності.
- **Проєкт:** Продовження групового проекту: питання безпеки та відповідності, захист даних.

Лекція 15

Огляд курсу та обговорення досягнутих результатів (Семінар)

- Практичні рекомендації та досвід накопичений при розробці групових проектів
- **Проєкт:** Завершення групового проекту щодо інфраструктури Великих Даних підприємства.

«Нейронні мережі та Глибоке навчання»
«Neural networks and Deep Learning»

Лектор: Софія Швець. Data scientist, Vodafone Ukraine

Тривалість курсу: 24 аудиторні години
6 - лекційних занять
9- практичних занять
4 - лекційно практичних занять
1 - контрольна робота/код рев'ю

Програма курсу:

Заняття 1 (лекція)
- Вступ та розуміння Deep learning та Neural networks.

Заняття 2 (практичне заняття)
- Neural networks на практиці, огляд популярних фреймворків, PyTorch та поняття tensor. Операції з tensor

Заняття 3 (практичне заняття)
- Pipeline навчання Neural networks. Dataset та Dataloder

Заняття 4 (лекція)
- Вступ в NLP (Natural Language Processing). Векторне представлення слів. Модель word2vec

Заняття 5 (практичне заняття)
- Розуміння рекурентної нейронної мережі. Побудова архітектури RNN (Recurrent Neural Networks), LSTM та GRU на практиці

Заняття 6 (практичне заняття)
- Згорткові нейроні мережі. Задачі комп'ютерного зору. Архітектура CNN (Convolutional Neural Networks) на практиці

Заняття 7 (контрольна робота/код рев'ю)
- Code review

«Прогнозний аналіз і часові ряди» («Predictive Analytics and Time Series Analysis»)

Лектор: Любчик Леонід Михайлович, доктор технічних наук, професор

Тривалість курсу: 24 аудиторні години
12 лекційних занять по 1.5 (астрономічні) години кожне,
4 практичних заняття по 1.5 (астрономічні) години кожне.

Програма курсу:

Заняття 1 (Лекція).
Прогнозний аналіз - принципи, методи, засоби.

- Прогнозна аналітика - цілі, завдання, етапи. Поняття прогнозу, оцінка якості прогнозування. Принципи прогнозування, прогнозуючи моделі і алгоритми. Статистичні методи. Описові статистики. Статистичні висновки. Перевірка гіпотез і обчислення статистик. Прогнозна статистика.
- Часові ряди - визначення, приклади. Структура і декомпозиція часових рядів. Тренд, сезонна, циклічна компоненти, залишки. Стаціонарні та нестаціонарні ряди. Моделі дискретних випадкових процесів. Автокореляція і спектральна щільність.

Заняття 2 (Лекція).
Оцінювання і прогнозування поліноміального тренду.

- Згладжування часових рядів. Оцінювання параметрів поліноміальних моделей тренду. Сплайнові моделі тренду. Вибір оптимального ступеня полінома. Проблема малої вибірки. Моделі тренду, оптимальні за складністю. Метод впорядкованої мінімізації ризику.
- Оцінювання нестаціонарного тренду. Ковзне середнє. Експоненціальне згладжування. Модель Брауна. Лінійний тренд, модель Хольта. Мультиплікативний тренд і сезонність, модель Хольта-Вінтерса. Адитивний лінійний тренд і сезонність, модель Тейла-Вейджа. Налаштування адаптивних моделей, модель Трігга-Ліча. Аналіз адекватності адаптивних моделей. Слідкуючий контрольний сигнал.

Заняття 3 (Лекція).
Структурний аналіз часових рядів. Прогнозування циклічного тренду.

- Динамічний метод головних компонент. Сингулярний спектральний аналіз (SSA). Побудова матриці спостережень. Аналіз головних компонент. Побудова SSA прогнозу. Алгоритм «Гусениця».
- Структурний спектральний аналіз. Полігармонічні моделі. Адаптивні гармонійні фільтри. Налаштування параметрів. Полігармонічні моделі. Прогнозування циклічних трендів.

Заняття 4 (Практика).
Дослідження методів згладжування і декомпозиції часових рядів.

- Виконання індивідуальних домашніх завдань по виділенню і прогнозування трендів часових рядів, презентація, обговорення і захист результатів.

Заняття 5 (Лекція).
Регресійні прогнозуючі моделі.

- Прогнозування на основі множинної лінійної регресії. Метод максимальної правдоподібності. Теорія нормальної регресії. Метод найменших квадратів. Узагальнені лінійні регресійні прогнозуючі моделі.
- Мультиіколлінеарніость. Регуляризація і гребнева регресія. Робастна регресія. Вибір функції втрат. Функції втрат Вапніка і Хубера. Оптимальні на класі функції втрат.

Заняття 6 (Лекція).
Авторегресійні прогнозуючі моделі.

- Методологія Бокса - Дженкінса. Стаціонарні ряди, тести на стаціонарність. Модель авторегресії - змінної середньої (ARMA). Нестаціонарні ряди. Модель авторегресії - проінтегрованої змінної середньої (ARIMA). Модель ARIMA із сезонною компонентою. Гетероскедастичні авторегресійні моделі.

Заняття 7 (Лекція).
Ідентифікація та оцінювання параметрів авторегресійних моделей.

- Постановка задачі ідентифікації. Критерії ідентифікації. Рекурентне оцінювання параметрів. Верифікація моделей. Адаптивні авторегресійні предиктори.

Заняття 8 (Практика).
Розробка і дослідження регресійних і авторегресійних моделей.

- Виконання індивідуальних домашніх завдань з розробки і дослідження регресійних та авторегресійних прогнозуючих моделей, презентація, обговорення і захист результатів.

Заняття 9 (Лекція).
Прогнозування на основі машинного навчання.

- Прогнозування як задача навчання по прецедентах. Функціонал емпіричного ризику.
  Точність прогнозування й узагальнююча здатність. Ефект перенавчання. Емпіричні методи оцінювання узагальнюючої здатності. Розбиття навчальної вибірки. Крос-валіидацiия. Процедура ковзного контролю.
- Побудова лінійних прогнозуючих моделей методами машинного навчання.

Заняття 10 (Лекція).
Прогнозування на основі методу опорних векторів.

- Метод машин опорних векторів (SVM). Лінійний SVM метод класифікації. Оптимальна розділяюча гіперплощина. Лінійний SVM метод регресії. Вибір функції втрат і методу регуляризації. Лінійні SVM - предиктори часових рядів.

Заняття 11 (Лекція).
Прогнозування на основі машинного навчання. Нелінійні прогнозуючі моделі.

- Ядерні методи машинного навчання. Ядра і спрямляючі просторий, конструктивні засоби побудови ядер. Ядерний метод опорних векторів. Ядерний трюк. Теорема Мерсера. Ядерні нелінійні SVM - предиктори часових рядів.
- Нелінійні адаптивні прогнозуючі моделі нестаціонарних часових рядів. Рекурентна ядерна ідентифікація. Зниження порядку моделі. Метод ковзного вікна. Нелінійні ядерні предиктори з кінцевоюї пам'яттю.

Заняття 12 (Практика).
Розробка і дослідження прогнозуючих моделей, що навчаються.

- Виконання індивідуальних домашніх завдань з розробки і дослідження прогнозуючих моделей, побудованих на основі методів машинного навчання, презентація, обговорення і захист результатів.

Заняття 13 (Лекція).
Прогнозування коіинтегрованих часових рядів.

- Моделі нестаціонарних коінтегрованих процесів. Коінтеграція і тренди. Коінтеграція і корекція помилок. Прогнозування коіинтегрированиных процесів.

Заняття 14 (Лекція).
Прогнозування багатовимірних часових рядів.

- Подання багатовимірних часових рядів у просторі станів. Метод рекурентного оцінювання станів. Дискретний фільтр Калмана. Прогнозування стану багатовимірних моделей.

Заняття 15 (Лекція).
Уточнення і розгортання прогнозних моделей.

- Уточнення прогнозних моделей. Валідація предикторів. Вибір моделей, колективи. Методи уточнення прогнозних моделей - скорінг моделей, інтеграція із звітністю, інтеграція із застосуваннями. Моніторинг процесів прогнозування. Особливості практичних застосувань.
- Розгортання прогнозних моделей. Сервіси прогнозування. Навчальні та прогнозні експерименти. Розгортання web-служби прогнозування.

Заняття 16 (Практика).
Презентація, обговорення і захист результатів комплексного домашнього завдання.

B I G

D A T A

L A B

vodafone

Програма навчання може змінюватись протягом курсу