Національна академія наук України
Інститут кібернетики імені В.М.Глушкова
Київський національний університет імені Тараса Шевченка

А.Л. Головинський

курс лекцій
”ВЕЛИКІ ДАНІ”

Київ – 2020

ЗМІСТ

ВСТУП

Лекції на github: https://github.com/agolovynskyi1/bigdata-lecturenotes

1 ВЕЛИКІ ДАНІ

1.1 Вступ

Означення 1 *Великі дані (Big Data) це неструктуровані, слабкоструктуровані та структуровані дані, з якими не можна працювати стандартними математичними методами. Слово "великі" означає, що їх обробка також потребує спеціальних методів паралельного програмування та відповідно багатопроцесоних обчислювальних систем.*

Це маркетинговий термін, який дозволяє розвивати і продавати математичні методи бізнесу.

Прикладами великих даних є архіви інтернет-форумів, записи даних відеоспостереження торгового центру, дані з чеків покупок да ідентифікатори дисконтних карток, літературні тексти, звукові файли тощо.

Не відносяться до великих даних, незалежно від розміру, добре структуровані дані з чіткою і відомою моделлю фізичного процесу, в яких можна напряму застосувати зрілу математичну теорію, з широкого асортименту розроблених на останні 300 років.

Основний фокус досліждень полягає у роботі з неструктурованими даними з низькою щільністю корисної інформації та отримання з них структурованих даних, це називається Data Mining, дослівно "видобування даних".

Задачі, які вирішуються у даній області – це перевірка, тобто підтвердження чи спростування певної гіпотези. Наприклад,

– чи є певна людина потенційним покупцем?
– чи залежать дані процеси від заданого параметра?
– яка частина даних несе інформацію про дане явище чи процес?

Гіпотеза дає апріорну оцінку, що є потенційно корисною інформацією, а що ігноруються.

Наприклад, деяка компанія виробляє продукцію і продає її по всьому світу. І хоче оцінити симпатії, оцінки споживачів у різних країнах.

Прямим методом буде визначення цільової аудиторії, проведення опитування у даних країнах, аналіз отриманих анкет.

Непрямим методом буде аназіл постів у соцмережах, побудова векторних просторів (для кожної країни і мови) із спеціальною метрикою, і визначення близькості назви продукту до слів-маркерів. Це ї є великі дані. Відмітимо, що для вирішення даної задачі навіть не потрібно знати мову цих країн.

1.2 Типи даних за структурою

Означення 2 *Структуровані дані: дані, що зберігаються у рядках та стовпцях, здебільшого числові, де чітко визначено значення кожного елемента даних.*

Цей тип даних становить близько 10% від загального обсягу даних на сьогодні. Такі дані доступні через системи управління базами даних. Прикладні джерела структурованих (або традиційних) даних включають

- офіційні реєстри, які створюються урядовими установами для зберігання даних про осіб, підприємств та нерухомість;
- дані датчиків у промисловості, які збирають інформацію на заводах про процеси, для контролю руху, температури, розташування, світла, вібрації, тиску, рідини та потоку.

Означення 3 *Неструктуровані дані: дані різних форм, наприклад, текст, зображення, відео, документи тощо.*

Це можуть бути скарги клієнтів, контракти або внутрішні електронні листи. Цей тип даних становить близько 90% даних, створених у 21 столітті. Вибухове зростання соціальних медіа (наприклад, Facebook та Twitter), починаючи з середини минулого десятиліття, є причиною більшої частини неструктурованих даних, які ми маємо сьогодні. Неструктуровані дані не можуть оброблятися за допомогою методів традиційних реляційних (табличних) баз даних. Важливість неструктурованих даних полягає у схованих у них зв'язках, які можуть бути не виявлені, якщо розлядати лише структуровані дані (наприклад, інформація про людину в соціальній мережі набагато повніша, ніж анетна інформація, яка подаєтья на візу чи у відділ кадрів). Саме це

робить дані, створені в соціальних медіа, відмінними від інших типів даних, це те, що дані в соціальних медіа мають сильний відбиток особистості.

Означення 4 *Географічні дані: дані, пов'язані з дорогами, будівлями, озерами, адресами, людьми, робочими місцями та транспортними маршрутами, які генеруються з географічних інформаційних систем.*

Ці дані пов'язують місце, час та атрибути (тобто описову інформацію). Географічні дані, які є цифровими, мають величезні переваги перед традиційними джерелами даних, такими як карти, письмові звіти дослідників та розмовні записи, в яких цифрові дані легко копіювати, зберігати та передавати. Що ще важливіше, їх легко трансформувати, обробляти та аналізувати. Такі дані корисні для містобудування та моніторингу впливу на навколишнє середовище. Гілка статистики, яка бере участь у просторових або просторово-часових даних, називається Геостатистика.

Означення 5 *Мультимедійні дані: потокове передавання в реальному часі живих або збережених медіа-даних.*

Особливою характеристикою засобів масової інформації в режимі реального часу є величезні об'єми відео, зображень та аудіо, які в майбутньому будуть лише зростати. Одним з основних джерел медіа-даних є такі сервіси, як, наприклад, YouTube, відеоконференції.

Означення 6 *Тести природньою мовою: дані, що генеруються людьми у текстовій формі.*

Такі дані різняться за рівнем абстракції та рівнем редакційної якості. До джерел даних природної мови відносяться пристрої збору мовлення, наземні телефони, мобільні телефони та Інтернет речей, які створюють великі розміри текстового зв'язку між пристроями.

Означення 7 *Часовий ряд: послідовність точок даних (або спостережень), як правило, що складається з послідовних вимірювань, проведених за часовий інтервал.*

Мета - виявити тенденції та аномалії, визначити контекст та зовнішні впливи та порівняти індивіда з групою або порівняти окремих людей у різний час. Існує два види даних часових рядів:

– безперервне, де ми спостерігаємо в кожен момент часу,
– де ми спостерігаємо через (зазвичай регулярно) проміжки інтервалів.

Прикладами таких даних є океанські припливи, кількість сонячних плям, вартість валют, цінних паперів на біржі, та вимірювання рівня безробіття кожного місяця року.

Означення 8 *Дані про події: дані, згенеровані в результаті зіставлення зовнішніх подій із часовим рядом.*

Це вимагає відокремлення важливих подій від неважливих. Наприклад, інформацію, пов'язану з аваріями транспортних засобів або аваріями, можна збирати та аналізувати, щоб зрозуміти фактори, які спричинили подію і її наслідки. Дані в цьому прикладі формуються датчиками, закріпленими в різних місцях кузова транспортного засобу. Дані про подію складаються з трьох основних фрагментів інформації:

– дія, яка є самою подією,
– часова мітка, час, коли ця подія сталася,
– стан, який описує всю іншу інформацію, що стосується цієї події.

Дані про події зазвичай мають різноманітну структуру, неунормовані значення, вкладену структуру.

Означення 9 *Дані про мережу: дані стосуються структури зв'язів дуже великих мереж, таких як соціальні мережі (наприклад, Facebook і Twitter), інформаційні мережі (наприклад, всесвітня павутина), біологічні мережі (наприклад, біохімічні, екологічні та нейронні мережі) та технологічні мережі (наприклад, Інтернет). Такі дані представлені графом із вузлами, з'єднаними зв'язками різних типів. Об'єктом дослідження стає саме структура мережі та характер зв'язків.*

1.3 Типи даних за значеннями

Якщо розглянути величини елементів даних, то вони поділяються на категорні (або іменні), порядкові та числові. Нижче ми визначимо ці терміни та пояснимо, чому вони важливі.

Означення 10 *Категорна величина (іноді її називають іменною) - це така, яка має дві або більше категорій, але немає внутрішнього упорядкування категорій.*

Наприклад, стать - це категоріальна змінна, що має дві категорії (чоловіча та жіноча), і не має внутрішнього впорядкування між ними. Колір волосся - це також категорна величина, що має ряд категорій (блондинка, каштанова, брюнетка, руда тощо), і знову ж таки, немає змістовного способу їх впорядкування від найвищого до нижчого.

Означення 11 *Порядкова величина схожа на категорну величину. Різниця між ними полягає в тому, що є чітке впорядкування змінних.*

Наприклад, припустимо, що ви маєте змінний економічний статус з трьома категоріями (низький, середній та високий). Окрім того, що ви можете класифікувати людей на ці три категорії, ви можете впорядкувати категорії як низькі, середні та високі. Тепер розглянемо таку величину, як навчальний досвід (із такими значеннями, як випускник початкової школи, випускник середньої школи, якийсь випускник коледжу та коледж). Їх також можна замовити як початкову школу, середню школу, якийсь коледж, так і випускник коледжу. Незважаючи на те, що ми можемо упорядкувати їх від найнижчого до найвищого, інтервал між значеннями може бути не однаковим для рівнів величин.

Скажімо, ми присвоюємо бали 1, 2, 3 і 4 цим чотирьом рівням освітнього досвіду, і ми порівнюємо різницю в освіті між категоріями першою і другою з різницею навчального досвіду між категоріями два і три, або різницю між категоріями три і чотири. Різниця між категоріями першої та другої (початкова та середня школа), ймовірно, набагато більша, ніж різниця між категоріями дві та три (середня школа та деякі коледжі). У цьому прикладі ми можемо упорядкувати людей за рівнем освітнього досвіду, але розмір різниці між категоріями невідповідний (оскільки інтервал між категоріями одна і дві більший, ніж категорії дві та три). Якби ці категорії були однаково розташовані, то змінна була б числовою змінною.

Означення 12 *Числова величина схожа на порядкову змінну, за винятком того, що інтервали між значеннями числової змінної однаково розташовані.*

Наприклад, припустимо, у вас є така величина, як річний дохід, яка вимірюється в доларах, а у нас є троє людей, які заробляють 10 000, 15 000 і 20 000 доларів. Друга

людина заробляє на 5000 доларів більше, ніж перша особа, і 5000 доларів менше, ніж третя особа, і розмір цих інтервалів однаковий. Якби були ще двоє людей, які б заробляли 90 000 доларів і 95 000 доларів, розмір цього інтервалу між цими двома людьми також був би однаковий (5000 доларів).

Одна і та сама величина може бути різною за типом, відповідно до контексту. Наприклад, число може відноситить як до числової величини, так і до категорної, якщо це номер комунікаційного порту, ідентифікатор транспортного засобу тощо.

1.4 Вектори і тензори

Означення 13 *Скаляр (від лат. scalaris — східчастий) – величина, кожне значення якої може бути виражене одним числом.*

Означення 14 *Вектор (від лат. vector, «той що несе») – (послідовність, кортеж) однорідних елементів.*

Означення 15 *Матриця – сукупність математичних величин, певним способом розміщених у прямокутній таблиці.*

Означення 16 *Тензор (від лат. tendere, «тягнутись, простиратися») – тензор представляється у вигляді багатовимірної таблиці, заповненої числами.*

Для об’єктів однакової розмірності природньо визначаються операції додавання і множення на скаляр.

1.5 Кодування

1.5.1 Унітарний код

Нехай є набір X з n елементів.

Унітарний код (англ. one-hot encoding) – це таке відображення, коли кожному елементу $x \in X$ ставиться у відповідність вектор $e \in R^n$, що містить тільки одну 1, яка відповідає позиції x наборі.

1.5.2 Кодування зображень

Для кодування кольорових графічних зображень застосовується принцип декомпозиції кольору на основні складові: червоний (Red), зелений (Green) і синій (Blue). Цей принцип базується на тому, що будь-який колір можна отримати шляхом змішування трьох зазначених кольорів. Система кодування за першими літерами назв основних змішуються квітів називається системою RGB і описує поведінку адитивної моделі кольорів. Якщо для кодування яскравості кожної складової кольору використовувати 256 градацій (8-розрядне число), то для кодування кольорової точки досить 24-розрядного двійкового числа.

Таким чином, зображення може бути представлене тензором $H \cdot W \cdot R \cdot G \cdot B$. Кожна величина у тензорі має числовий тип.

Lab – система задання кольорів, що використовує як параметри світлосилу, відношення зеленого до червоного та відношення синього до жовтого. Ці три параметри утворюють тривимірний простір, точки якого відповідають певним кольорам.

Приклад. Представлення зображень поняття “яблуко” у тензори.

Афінський філософ Платон (427-347 рр. до н.е.). Основну частину всієї філософії Платона займає теорія ідей. Філософ визначає наявність існування двох світів – ідей та речей. Ідеї (від грец. ейдос) є ніщо інше як прообрази або ж інакше кажучи – витоки речей. У свою чергу Платоном висувається думка стосовно того, що в основі безлічі речей, які утворені від безформної матерії покладені ідеї. Вони виступають джерелом усього, а матерія здатності породження немає.

1.5.3 Тексти природньою мовою

Означення 17 *Нехай множина Ω – це набір символів, будемо називати її алфавітом.*

Означення 18 *Словом з символів $l_i \in \Omega$ є скінченні послідовності $w = (l_1, l_2, \ldots, l_k), k \in \mathbb{N}$.*

Означення 19 *Нехай W деяка множина скінченних слів w, будемо називати її словником.*

Означення 20 *Мовою L над словником W є множина всіх скінченних послідовностей $s = (w_1, w_2, \ldots, w_n), w_i \in W, n \in \mathbb{N}$.*

Означення 21 *Корпусом $K \subset L$ називається деяка підмножина тестів мови.*

1.5.4 Модель “Лантух із словами”

Тексти природньою мовою мають довільну структуру, мають багатозначні слова і звороти. Математичні методи не можуть працювати з тестами напряму, їх потрібно переводити у вектора чисел та тензори.

Популярним і простим методом кодування текстік вектори є модель “Лантух із словами” (англ. Bag of Words). Ця модель перетворює текст у вектор, який визначає кількість входжень слів у текст.

Це відбувається у два етапи:

- фомується словник;
- визначаєся кількість входжень кожного слова у тексті.

Саме тому ця модель називається “лантухом”, тому що вся інформація про порядок слів у тексті відкидається, і лише має значення, чи зустрічається слово у тексті, і скільки разів.

Нехай K корпус текстів деякої мови. За даним корпусом будується словник W із слів, які зустрічаються у текстах корпусу. Рахується частота використання слів словника у корпусі текстів.

Далі словник обробляється, з нього видаляються слова, які використовуються дуже часто і дуже рідко. Це робиться для зменшення розміру. У природних мовах словники можуть мати розмір 200 – 300 тис. слів, а для машинного навчання потрібні словники розміру 5 – 10 тис. слів.

Слова словника є категорними величинами, і тому кодуються унітарним кодуванням. Кожен текст розбиваться на слова, які представляються векторами, далі вектори для кожного слова додаються і отримується вектор, який має по кожній координаті кількість входжень відповідного слова.

Оскільки тексти можуть мати різну довжину, отримані вектори доцільно унормувати, поділивши на загальну кількість слів відповідного текста.

Практичне завдання: Imdb, оцінка емоційного забарвлення твітів.

Завдання:

Система контекстного пошуку статей і індивідуальних вподобань статей для сайту

Розумний бібліотекар: класифікація книжок за жанром.

Аналізатор системного журналу суперкомп’ютера.

1.6 Базові методи роботи з даними

1.6.1 Візуалізація даних

У 3D Python

1.6.2 Зниження розмірності векторного простору

1.6.3 Статистичні оцінки

Нехай X – набір даних $x_i \in X, i \in \mathbb{N}$, представлений у вигляді тензора.

Означення 22 *Математичне сподівання $M(X) = \frac{1}{n}\sum_{i=1}^{n} x_i$. Оскільки операції у тензорах x_i покомпонентні, то результатом є тензор тієї ж розмірності, з елементами $M(x_j)$, де $j \in \mathbb{N}$ пробігає по компонентам тензора.*

1.6.4 Хеммінга

Означення 23 *Нехай w_1, w_2 два вектори однакової довжини. Відстанню Хеммінга між ними буде кількість компонент, у яких їх значення від*

1.6.5 Метод головних компонент

Метод головних компонент (англ. principal component analysis) це перетворення координат простору таким чином, що перша головна компонента, у якої найбільша дисперсія, переходить у першу координату, друга у другу і так далі.

Нехай X – набір даних з n елементів, кожен елемент це вектор з $\mathbb{R}^p$. Він може бути представлений у вигляді матриці, у якої у строках знаходяться елементи $x_i \in X, x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$. Вхідні дані перетворені таким чином, що по кожному ствобчику матриці її середнє, дорівнює нулю, тобто від кожної компоненти вектора відняли її середнє і $\sum x_i = 0$.

Приклади. Модуль PCA у Python

Автоенкодери

1.6.6 Визначення залежностей

Регресія Логістична регресія

1.6.7 Кластеризація даних

Метод К-середніх

1.7 Вади даних

Нехай ми маємо вибірку даних.
Skin color matters: gender recognition error rate
Amazon recruiting tool biased towards women
Methods for Fixing Data Samples
Fixing dataset, make equal amount of classes Remove parasite correlations Correctly treat special data types (integer vs categorical) Missing data phenomen What if there are no more data to fulfill dataset?

Biased data, samples with missing data 2D and 3D objects and notions immersed to, for example, 5D space (RGB images) are much more complicated indeed Neural network as a function can get any value outside dataset

1.8 Аудіодані

Аналіз спектра, спектр від часу
LibROSA: `https://github.com/librosa/librosa/blob/master/examples/LibROSA%20demo.ipynb`

Задача1. Записати голоси 3-х людей і вібразити їх у вигляді PCA, t-sne, umap. Чи вони розрізняються?

Задача2. Умовний стук. Це вже часовий ряд

Задача3. Зробити систему, яка впізнає мелодії.

1.9 Часові ряди

1.9.1 Кореляція і автокореляція

Нехай є дві вибірки скалярних величин $X \in \mathbb{R}, Y \in \mathbb{R}, x_i \in X, y_i \in Y, i \in 1, 2, \ldots, n$, із середніми відповідно $\overline{x}, \overline{y}$ і стандартними відхиленнями σ_x, σ_y.

Означення 24 *Коваріацією виборок X, Y буде величина*

$$cov(X, Y) = M((X - \overline{X})(Y - \overline{Y}) = \frac{1}{n} \sum_{i=1} n(x_i - \overline{x})(y_i - \overline{y}) \tag{1.1}$$

це міра спільної мінливості двої випадкових величин.

Використовуючи властивість лінійності математичного сподівання

$$\begin{aligned} cov(X, Y) = M((X - M(X))(Y - M(Y)) = \\ = M(XY - XM(Y) - M(X)Y + M(X)M(Y)) = \\ = M(XY) - M(X)M(Y) - M(X)M(Y) + M(X)M(Y) = \\ = M(XY) - M(X)M(Y) = \frac{1}{n} \sum_{i=1}^{n} x_i y_i - \overline{xy} \end{aligned} \tag{1.2}$$

Якщо X та Y є незалежними, то їхня коваріація є нульовою. Це випливає з того, що за незалежності

$$M(XY) = M(X)M(Y) \tag{1.3}$$

Якщо $X \in \mathbb{R}^k, Y \in \mathbb{R}^m$ випадкові величини у векторній формі, аналогічно

визначимо взаємно-коваріаційну матрицю

$$cov(X,Y) = M((X - M(X))(Y - M(Y)^T) = M(XY^T) - M(X)M(Y)^T \quad (1.4)$$

(i,j)-тий елемент цієї матриці дорівнює коваріації $cov(X_i, Y_j)$ між i-тою скалярною складовою X та j-тою скалярною складовою Y.

Означення 25 *Кореляцією виборок $X \in \mathbb{R}, Y \in \mathbb{R}$ буде величина*

$$corr(X,Y) = \frac{cov(X,Y)}{\sigma_x \sigma_y} = \quad (1.5)$$

це коваріація, унормована на стандартні відхилення.

Нехай є випадкова величина $X \in \mathbb{R}$, для спрощення формул нехай $\overline{X} = 0, \sigma_X = 1$, тобто дані є центрованими і нормованими.

Визначимо вектор $X_i = (x_i, x_{i+2}, \ldots, x_{i+n})$ з n елементів, $x_i \in X$ і зміщення $\tau \in \mathbb{N}$.

Означення 26 *Автоковаріацією називається величина*

$$cov(X_i, X_{i+\tau}) = M(X_i X_{i+\tau}^T) = \frac{1}{n} \sum_{i=1}^{n} x_i x_{i+\tau} \quad (1.6)$$

Пошук максимума функції автоваріації за τ дозволяє знайти точні позиції шаблону у послідовності даних.

Якщо є гіпотеза, що випадкова величина має періодичний характер, але не відомими є величини періодів, при дослідженні залежності величини автоковаріації від ширини вікна n і $\tau = n$, максимуми будуть відповідати прихованим періодам.

Задача 1. Пошук прихованої перідодичності у даних, представлених у вигляді часових рядів. Дані брати пов'язані із людською активністю.

1.10 Робота з зображеннями і відео

SIFT

Конволюційні нейронні мережі

ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАННЯ