Міністерство освіти і науки України

Чернівецький національний університет
імені Юрія Федьковича

Юрченко І.В.

DataMining з використанням Python

навчальний посібник

Чернівці – 2024

УДК 519.21

DataMining з використанням Python. Навчальний посібник // Юрченко І.В.– Чернівці: Чернівецький національний університет імені Юрія Федьковича, 2024.– 143 с.

Друкується за ухвалою редакційної колегії
Чернівецького національного університету імені Юрія Федьковича

Рецензенти:
Ясинський Володимир Кирилович, доктор фіз.-мат. наук, професор
Семчук Аркадій Романович, кандидат фіз.-мат. наук, доцент

У начальному посібнику міститься опис основних понять та тверджень теорії інтелектуального аналізу даних (DataMining). Розглянуто основні методи та алгоритми DataMining, наведено приклади їх використання. Вивчено можливості бібліотек мови Python для проведення інтелектуального аналізу даних.

Для студентів спеціальності 124 – Системний аналіз.

ЗМІСТ

ВСТУП

Data Science (наука про дані) – це наука про методи аналізу даних і видобутку з них цінної інформації, знань. Вона тісно перетинається з такими областями як машинне навчання (Machine Learning) і наука про мислення (Cognitive Science) та технологіями для роботи з великими даними (Big Data). Вона прагне зрозуміти складні структури даних, виявити приховані тренди, створити прогностичні моделі та приймати поінформовані рішення на основі аналізу даних.

Основні компоненти та характеристики Data Science:

- Збір даних. Дані отримуються з різних джерел, таких як бази даних, веб-сайти, давачі IoT, соціальні мережі та багато іншого. Дані можуть бути структурованими, напівструктурованими та неструктурованими.
- Підготовка даних (Data Preprocessing). Дані, отримані з різних джерел, можуть бути брудними або неоднорідними. На цьому етапі дані очищаються, трансформуються і доводяться до придатного для аналізу вид: видалення викидів, заповнення пропущених значень і масштабування даних.
- Аналіз даних (Data Analysis). Тут дані досліджуються та аналізуються з використанням статистичних методів та візуалізації. Мета - виявити закономірності, кореляції та тренди в даних.
- Машинне навчання (Machine Learning). Data Science застосовує методи машинного навчання для створення моделей, які можуть робити прогнози і приймати рішення на основі даних: завдання класифікації, регресії, кластеризації тощо.
- Візуалізація даних. Подання результати аналізу у зрозумілій та інформативній формі: створення графіків, діаграм і інтерактивних дашбордів.
- Прийняття рішень. На основі аналізу даних та результатів машинного навчання приймаються поінформовані рішення.

Data Science може застосовуватись у різних галузях, таких як бізнес, наука, охорона здоров'я, фінанси, маркетинг та інші. Data Science допомагає організаціям отримувати цінну інформацію з великих обсягів даних та приймати більш обґрунтовані рішення. Ця галузь стрімко розвивається та знаходить застосування у багатьох сферах, включаючи бізнес, дослідження, охорону здоров'я, фінанси, державне управління та інші.

Data Mining (дата-майнінг), також відомий як Knowledge Discovery in Databases (KDD), є процесом автоматичного виявлення цікавих і раніше невідомих закономірностей, шаблонів, трендів та інформації у великих обсягах даних. Цей процес містить використання методів і алгоритмів з області штучного інтелекту, машинного навчання, статистики і баз даних для аналізу даних і виявлення прихованих знань.

Ключові аспекти Data Mining:

- Data Mining дозволяє виявляти шаблони та закономірності даних, які можуть бути використані для прогнозування майбутніх подій або прийняття рішень.
- Цей процес автоматизований, що дозволяє обробляти більші обсяги даних ефективніше, ніж при ручній обробці.
- Data Mining використовує різноманітні методи, такі як класифікація, кластеризація, асоціація, регресія тощо.
- Data Mining знаходить застосування у багатьох галузях, включаючи бізнес, медицину, фінанси, науку, маркетинг та інші. Наприклад, у бізнесі це може використовуватися для аналізу купівельної поведінки та визначення трендів на ринку.

Інструменти та програмне забезпечення: Для Data Mining використовуються спеціалізовані інструменти та програмне забезпечення, такі як Python, R, Weka та інші, які надають різні алгоритми та функції для аналізу даних.

Процес Data Mining може бути поділений на кілька етапів, включаючи збір даних, їхню передобробку, вибір методів аналізу, виявлення закономірностей та інтерпретацію результатів. Ця технологія дозволяє організаціям та дослідникам витягувати цінну інформацію з великих обсягів даних, що може допомогти у прийнятті більш поінформованих рішень та виявленні нових можливостей.

Список використаної літератури

1. https://matplotlib.org/
2. https://seaborn.pydata.org/
3. https://www.tensorflow.org/
4. https://scikit-learn.org/stable/index.html
5. https://pytorch.org/
6. https://docs.scipy.org/doc/scipy/reference/stats.html
7. https://developer.ibm.com/articles/ba-data-mining-techniques/
8. https://openai.com/
9. Nong Ye. Data Mining: Theories, Algorithms, and Examples.– London - New York: CRC Press, Taylor& Francis Group, 2014.– 8424 p.
10. Юрченко І.В. Прикладна статистика: Методичні вказівки до лабораторних робіт.– Чернівці: Рута, 2000.– 55 с.
11. Ясинський В.К., Юрченко І.В. Прикладний статистичний аналіз. Методичні рекомендації до лабораторних робіт.– Чернівці: Рута, 2008.– 84 с.
12. Юрченко І.В. Прикладний статистичний аналіз з використанням Python: Навч. посібник. Видання 3-є, доповнене.– Чернівці: Чернівецький національний університет, 2021.– 108 с.
13. Королюк В.С., Царков Є.Ф., Ясинський В.К. Ймовірність, статистика та випадкові процеси. Теорія та комп'ютерна практика. В 3-х т.– Чернівці: Золоті литаври, 2009.
14. Королюк В.С. Портенко М.І., Скороход А.В., Турбін А.Ф. Довідник з теорії ймовірностей і математичної статистики.– Київ: Наук. думка, 1978.– 582 с.

Навчальне видання

DataMining
з використанням Python
Навчальний посібник

Юрченко Ігор Валерійович

Публікується в авторській редакції

Підписано до друку 02.04.2024 р.
Електронне видання
Виготовлено з готового оригінал-макета замовника.