Анналін Нг та Кеннет Су

ОПАНУЙ ЧИСЛА!

НАУКА ПРО ДАНІ ДЛЯ НЕФАХІВЦІВ

[без складних математичних обчислень]

ВИДАВНИЧИЙ ДІМ ФАБУЛА #PRO

УДК 519.25
Н11

ВИДАВНИЧИЙ ДІМ ФАБУЛА #PRO

Оригінальна назва твору:
Numsense! Data Science for the Layman

Нг Анналін

Н11 Опануй числа! Наука про дані для нефахівців / Анналін Нг та Кеннет Су / пер. з англ. О. Буйвол. — Харків : ВД «Фабула», 2024. — 168 с.
ISBN 978-617-522-177-8

Сучасний світ перенасичений інформацією, зокрема даними. Величезною кількістю даних! Як серед них не загубитися? Як їх осягнути? Як використовувати в повсякденному житті та різноманітних галузях? Як аналізувати? Як інтерпретувати?

«Опануй числа!» — це зручний графічний опис ключових алгоритмів обробки даних, корисний як вступ для новачків у цій галузі, огляд для ділових людей, що працюють з аналітиками, чи стимул для тих, хто прагне знати, що відбувається з їхніми даними.

УДК 519.25

ISBN 978-617-522-177-8

Зміст

Вступне слово

Великі дані — це вже великий бізнес. Оскільки вони дедалі більше домінують у нашому житті, їхня монетизація зробилася основним завданням майже кожної організації. Методи розпізнавання закономірностей і прогнозування відкривають нові можливості для бізнесу. Прикладом слугують системи рекомендації товарів. Вони є безпрограшним варіантом як для покупців, так і для продавців, оскільки рекомендують першим товари, які можуть їх зацікавити, що, зі свого боку, приносить більший прибуток другим.

Але великі дані — це лише частина головоломки. Наука про дані, яка дає нам змогу аналізувати та використовувати їх, є багатогранною дисципліною, що охоплює машинне навчання, статистику та суміжні галузі математики. Зауважте, що машинне навчання посідає особливо важливе місце в цьому описі, оскільки є основним рушієм, який уможливлює розпізнавання шаблонів і прогнозування. Алгоритми машинного навчання, що лежать в основі науки про дані, у поєднанні з даними здатні призвести до неоціненних відкриттів і нових способів використання інформації, яка вже є у нас під рукою.

Аби усвідомити, як наука про дані рухає сучасну інформаційну революцію, непосвяченим слід краще зрозуміти цю сферу. Проте побоювання щодо необхідних навичок змушують декого уникати цієї галузі — і це попри високий попит на грамотність у роботі з даними.

Саме тут на допомогу приходить «Опануй числа! Наука про дані для нефахівців». Після ознайомлення з книжкою Анналін Нг і Кеннета Су я переконався, що вона цілком відповідає своїй назві. Автори висвітлюють науку про дані саме для нефахівців, тому складну математику, яку подано на високому рівні, навмисно не описано детально. Але це не означає, що зміст книжки вихолощено. Насправді інформація, що міститься в цій праці, є ґрунтовною, і її перевага полягає саме в тому, що вона чітка й лаконічна.

«Чи є користь від такого підходу?» — запитаєте ви. Насправді її дуже багато! Я стверджую, що для неспеціаліста такий підхід є найкращим. Уявіть нефахівця, який зацікавлений у вивченні роботи автомобіля. Загальний огляд складових частин автівки, імовірно, буде менш складним, ніж технічний текст із фізики горіння. Те саме стосується вступного курсу науки про дані: якщо ви зацікавлені у вивченні цієї галузі знань, легше почати із загальних понять, перш ніж заглиблюватися в математичні формули.

Вступ до книжки на кількох сторінках ознайомить непосвячених із фундаментальними поняттями. Це гарантує, що кожен почне читати, маючи загальне уявлення, що таке наука про дані. Важливі поняття, як-от вибір алгоритму, що їх часто не беруть до уваги у вступних матеріалах, також розглянуто одразу — це прищеплює читачам відчуття

нагальної потреби сформувати власне розуміння цієї галузі знань і надає всеосяжну основу для досягнення такої мети.

Існує безліч концепцій, які Анналін і Кеннет могли би вважати гідними висвітлення у своїй книжці, а також чимала кількість способів їх презентації. Автори зосередилися переважно на алгоритмах машинного навчання, що мають найбільше значення для науки про дані, а також додали кілька сценаріїв на основі конкретних завдань — і це було чудовим рішенням. Випробувані та перевірені алгоритми, як-от кластеризація методом *k*-середніх, дерево ухвалення рішень і метод *k*-найближчих сусідів, отримали належну оцінку. Також пояснено новітні алгоритми класифікації та ансамблеві алгоритми, наприклад, метод опорних векторів, який часто відлякує своєю складною математикою, і метод випадкових лісів. Також розглянуто нейронні мережі, які є рушійною силою нинішнього шаленого захоплення глибоким навчанням.

Поєднання алгоритмів з інтуїтивно зрозумілими прикладами використання — ще одна перевага цієї книжки. Під час пояснення методу випадкових лісів у контексті прогнозування злочинності або ж методу кластеризації в контексті профілювання кіноманів обрані приклади забезпечують чіткість і практичне розуміння. Водночас відсутність будь-яких згадок про вищу математику підтримує зацікавленість і мотивацію до того, що, як передбачається, є першим кроком читача у вивченні науки про дані.

Я дуже рекомендую книжку «Опануй числа! Наука про дані для нефахівців» новачкам, які шукають точку входу в науку

про дані або алгоритми, що нею керують. Мені не спадає на думку інше подібне видання. З появою цієї книжки більше немає причин, аби математика завадила вам здобувати знання.

Метью Мейо, науковець із даних
і редактор сайту KDnuggets
@mattmayo13

Передмова

«Опануй числа! Наука про дані для нефахівців» написано двома ентузіастами науки про дані — Анналін Нг (Кембриджський університет) і Кеннетом Су (Стенфордський університет).

Ми помітили: хоча науку про дані дедалі частіше використовують для полегшення ухвалення рішень безпосередньо на робочому місці, багато хто й досі мало знає про цю галузь знань. Тому ми зібрали навчальні інструкції в одній книжці, щоби більше людей мали змогу навчатися — байдуже, чи то студент, підприємливий бізнесмен або ж будь-хто з допитливим розумом.

Кожна така інструкція пояснює важливі функції та передумови використання окремого методу науки про дані, не використовуючи складної математики й зайвої наукової термінології. Ми також ілюструємо ці методи даними та прикладами з реального життя.

Ми не змогли б написати цієї книжки самотужки.

Висловлюємо вдячність нашій редакторці й добрій подрузі Соні Чан за те, що вона вміло поєднала наші різні стилі написання та забезпечила плавність викладу матеріалу.

Ми дякуємо Дорі Тан, нашому талановитому графічному дизайнеру, за макет книжки та дизайн обкладинки оригінального видання.

Щира подяка нашим друзям Мішель По, Деннісу Чу та Маркові Хо за безцінні поради щодо того, як поліпшити зрозумілість контенту.

Ми також дякуємо професору Лонгу Нгуену (Мічиганський університет, місто Анн-Арбор), професору Персі Лянгу (Стенфордський університет) і доктору Михалу Косінському (Стенфордський університет) за те терпіння, яке вони виявляли до нас, а ще за те, що ділилися з нами своїми експертними порадами.

Нарешті, ми хотіли би подякувати одне одному за те, що час від часу сперечалися, як це роблять хороші друзі, але завжди залишалися разом до кінця, аби завершити розпочату справу.

Чому наука про дані?

Уявіть себе молодим лікарем.

До вас звертається пацієнт зі скаргами на задишку, біль у грудях і періодичну печію. Ви його обстежуєте. Виявляється, його кров'яний тиск і частота серцевих скорочень у нормі, а ще в нього не було жодних попередніх захворювань.

Під час огляду ви виявляєте, що пацієнт повненький. Оскільки його симптоми характерні для людей з надмірною вагою, ви запевняєте його, що все під контролем, і рекомендуєте знайти час для фізичних вправ.

Подібна ситуація занадто часто призводить до встановлення хибного діагнозу, оскільки пацієнти із серцево-судинними захворюваннями мають симптоми, схожі на ознаки звичайного ожиріння. Лікарі часто не проводять подальшого обстеження, яке могло би виявити більш серйозне захворювання.

Судження кожної людини обмежені її суб'єктивним досвідом і неповнотою знань. Це погіршує процес ухвалення рішень і може, як у випадку з недосвідченим лікарем, завадити подальшому обстеженню, яке здатне допомогти дійти більш точних висновків.

Саме тут стане в пригоді наука про дані.