ЛЕКЦІЯ № 8
з навчальної дисципліни

Тема лекції: Обробка та аналіз даних для початківців

Питання лекції

1. П’ять питань, на які дають відповідь обробка і аналіз даних
2. Критерії оцінки даних
3. Правильна постановка правильного питання
4. Прогнозування відповіді за допомогою простої моделі

1. П’ять питань, на які дають відповідь обробка і аналіз даних

Обробка та аналіз даних можуть бути непростим завданням, тому розповімо про базові поняття в цій галузі, без будь-яких формул або комп'ютерного жаргону, зрозумілого тільки програмістам.

У класиці це називається "5 питань, на які дають відповідь обробка і аналіз даних".

Для прогнозування відповідей на питання функція обробки і аналізу даних використовує **числа** і **імена** (також відомі як **категорії** або **мітки**).

Можливо, вас це здивує, але існує тільки п'ять основних питань, на які обробка і аналіз даних дають відповідь:

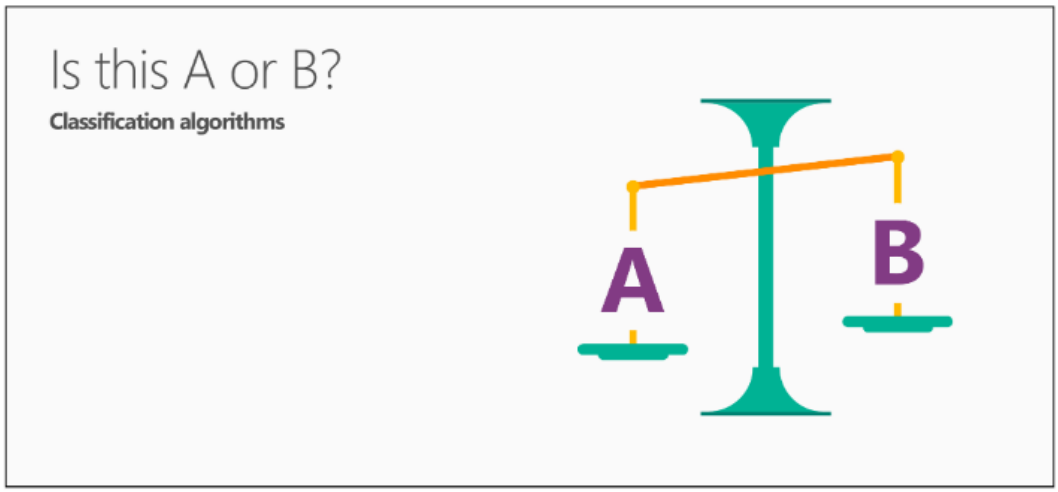
1. Це А або В?
2. Чи є це дивним?
3. Скільки?
4. Як це організовано?
5. Що робити далі?

Для відповіді на кожне з цих питань використовується окрема група методів машинного навчання, які називаються алгоритмами.

Щоб краще зрозуміти, уявіть, що алгоритм - це рецепт страви, а дані - це інгредієнти для його приготування. Алгоритм вказує, як об'єднати і змішати дані, щоб отримати відповідь. Комп'ютер можна порівняти з міксером. Більшу частину важкої роботи алгоритму комп'ютер робить за вас, і робить це досить швидко.

Питання 1 "Це А або В?" використовує алгоритми класифікації

Давайте почнемо з питання "Це А або В?"



Алгоритми класифікації: "Це А або В?"

Ця група алгоритмів називається двохкласовою класифікацією.

Її зручно використовувати для питань, які мають тільки два можливих варіанти відповіді.

Наприклад:

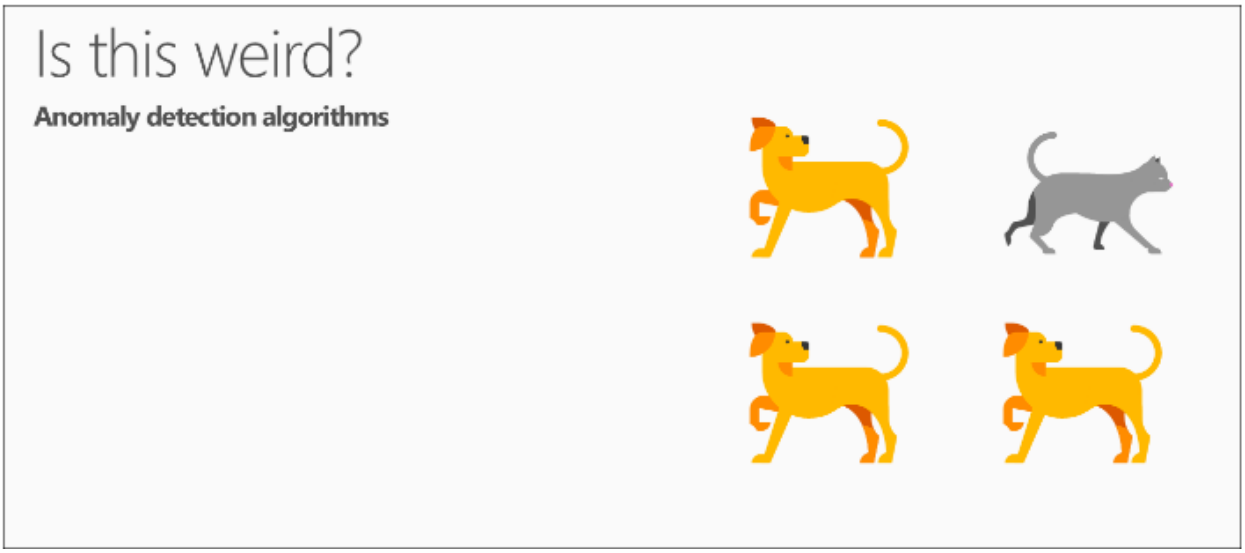
Чи вийде з ладу ця шина на наступних 1000 км: так чи ні?

Який купон привертає більше клієнтів: купон на 5 доларів або на знижку в 25%?

Це питання також можна перефразувати, щоб воно включало більше двох варіантів відповіді: "Це A, B, C або D, і т. Д.?" Це називається класифікацією по декількох класах. Її зручно використовувати, коли допускається кілька (або кілька тисяч) можливих варіантів відповіді. Класифікація по декількох класах вибирає найбільш ймовірний варіант.

Питання 2 "Чи є це дивним?" використовує алгоритми виявлення аномалій

Наступне питання, на який дають відповідь обробка і аналіз даних: "Чи є це дивним?" Для відповіді на це питання використовується група алгоритмів, що називається виявленням аномалій.



Алгоритми виявлення аномалій: "Чи є це дивним?"

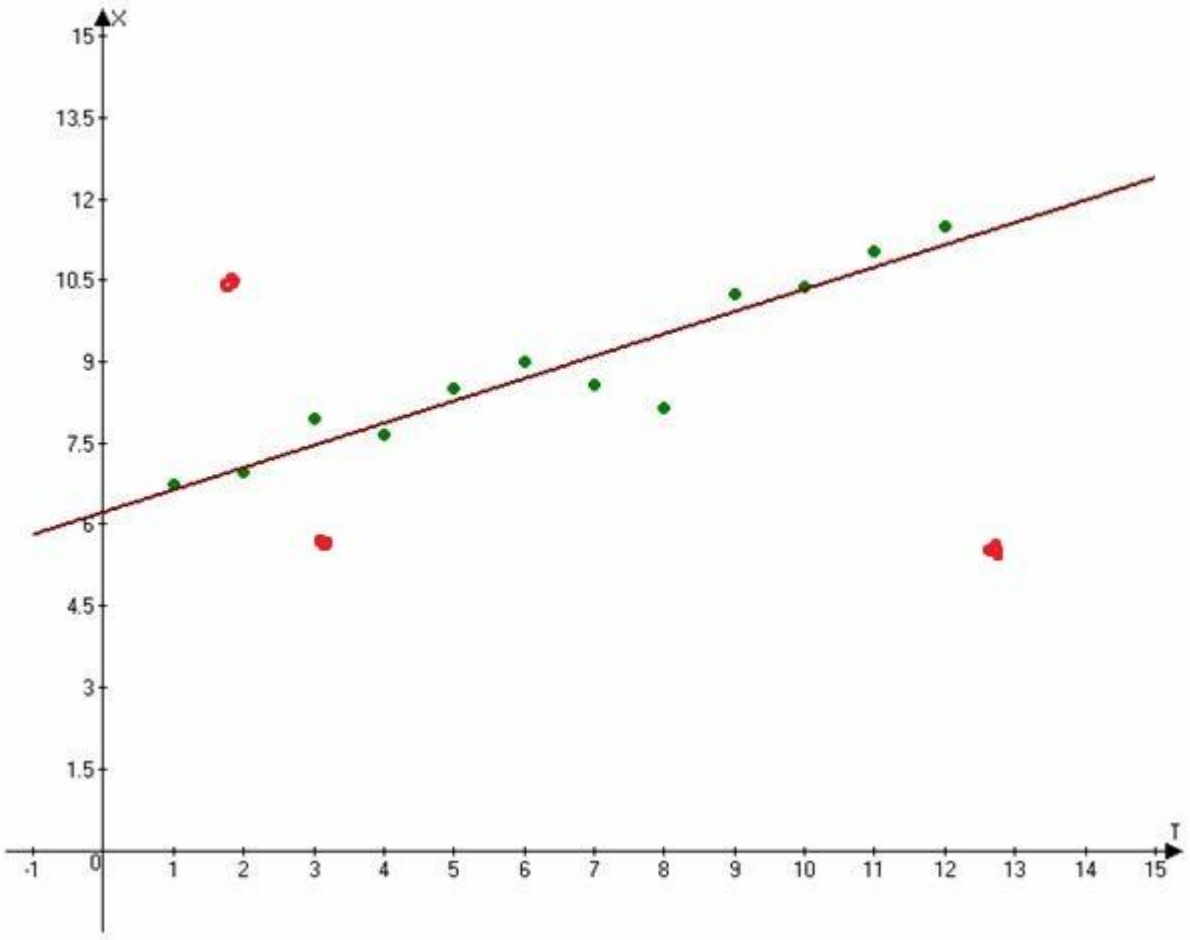
Якщо у вас є кредитна карта, ви вже отримали перевагу від виявлення аномалій. Компанія (банк), що обслуговує кредитну карту, аналізує ваші стандартні покупки і може

попередити вас про можливу спробу шахрайства. "Дивними" витратами можуть, наприклад, вважатися покупки в магазині, де ви зазвичай не буваєте, або покупка дуже дорогого товару.

Цей тип питання може бути корисним у багатьох ситуаціях, наприклад:

Якщо ваш автомобіль оснащений манометром, можливо, ви хочете знати, чи коректно він зчитує дані про тиск.

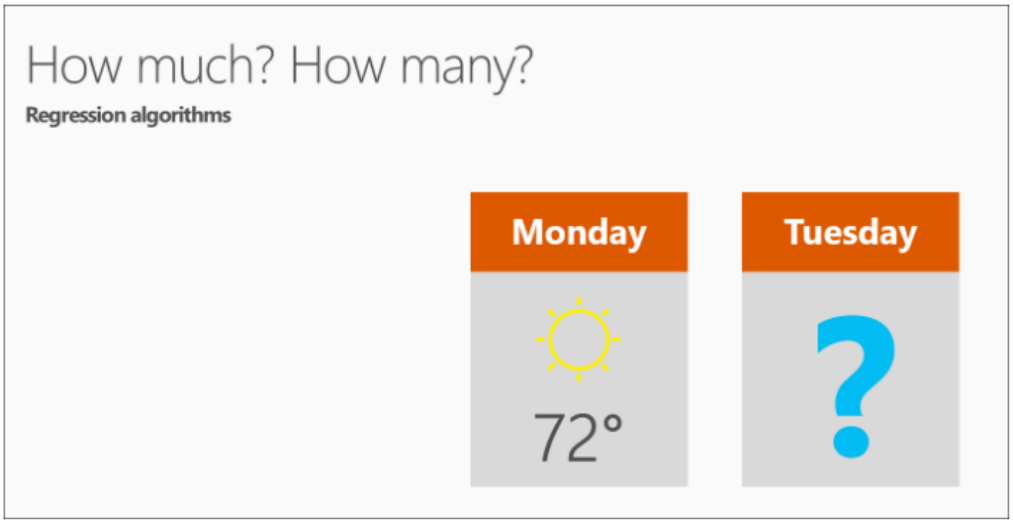
Якщо ви спостерігаєте в Інтернеті, то чи можете ви сказати: це повідомлення від звичайного Інтернету?



Функція виявлення аномалій реєструє несподівані або незвичайні події або особливості поведінки. А також підказує, де можуть бути проблеми.

Питання 3 "Скільки?" використовує алгоритми регресії

Служба машинного навчання також може прогнозувати відповідь на питання "Скільки?" Група алгоритмів, що відповідає на це питання, називається регресією.



Алгоритми регресії: "Скільки?"

Алгоритми регресії роблять числові прогнози, такі як в наведених нижче прикладах:

Якою буде температура в наступний вівторок?

Яким буде обсяг продажів за четвертий квартал?

Ці алгоритми допомагають відповісти на будь-яке питання, якщо запитується число.

Automobile price prediction › Automobile price data (Raw) › dataset

rows 205 columns 26

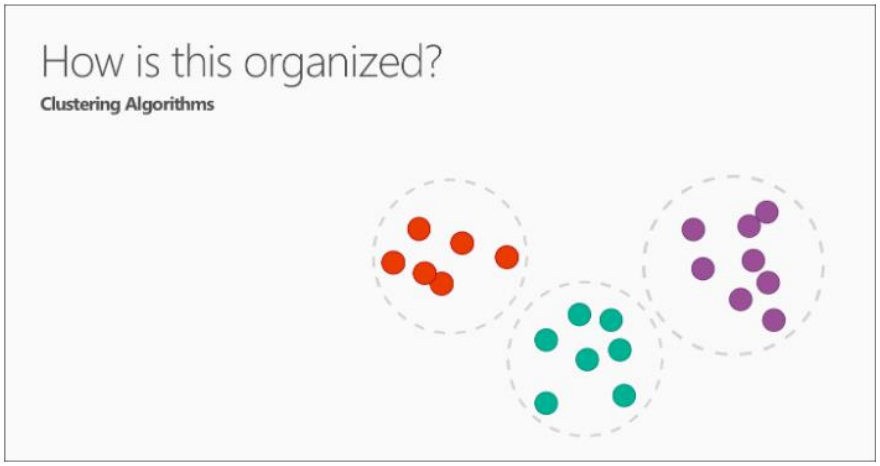
	symboling	normalized-losses	make	fuel-type	aspiration	num-of-doors	body-style	peak-rpm	city-mpg	highway-mpg	price
view as											
	3		alfa-romero	gas	std	two	convertib	5000	21	27	13495
	3		alfa-romero	gas	std	two	convert	5000	21	27	16500
	1		alfa-romero	gas	std	two	hatch	5000	19	26	16500
	2	164	audi	gas	std	four	sed	5500	24	30	13950
	2	164	audi	gas	std	four	se	5500	18	22	17450
	2		audi	gas	std	two	s	5500	19	25	15250
	1	158	audi	gas	std	four		5500	19	25	17710
	1		audi	gas	std	four		5500	19	25	18920
	1	158	audi	gas	turbo	four		5500	17	20	23875
	0		audi	gas	turbo	two		5500	16	22	
	2	192	bmw	gas	std	two		5800	23	29	16430
	0	192	bmw	gas	std	four		5800	23	29	16925
	0	188	bmw	gas	std	two		4250	21	28	20970
	0	188	bmw	gas	std	fo		4250	21	28	21105
	1		bmw	gas	std	f		4250	20	25	24565

Питання 4 "Як це організовано?" використовує алгоритми кластеризації

Останні два питання трохи складніші.

Іноді потрібно зрозуміти структуру набору даних, тобто задати питання "Як це організовано?" Для цього питання у вас немає прикладів, для яких ви вже дізналися результати.

Існує безліч способів виявлення структури даних. Одним із способів є кластеризація. Вона розділяє дані на природні групи, щоб спростити їх інтерпретацію. При використанні кластеризації немає єдиної правильної відповіді.



Алгоритми кластеризації: "Як це організовано?"

Нижче наводяться найпоширеніші приклади питань з кластеризацией.

Які глядачі віддають перевагу ті ж типи фільмів?

Які моделі принтерів виходять з ладу аналогічним чином?

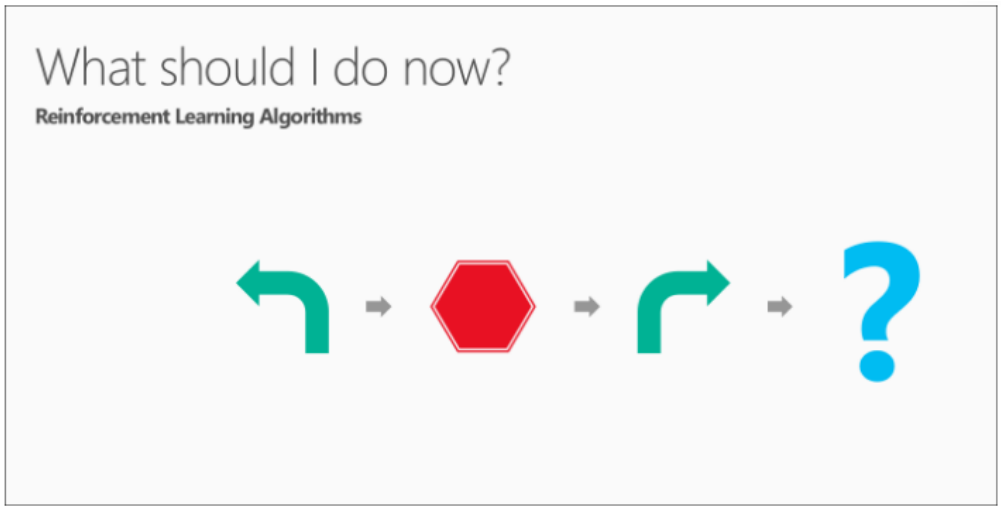
Зрозумівши, як організовані дані, ви зможете краще зрозуміти і спрогнозувати варіанти поведінки і подій.

Питання 5 "Що робити далі?" використовує алгоритми навчання з підкріпленням

І останнє питання - що робити далі? Тут ми використовуємо серію алгоритмів для навчання з підкріпленням.

Принцип навчання з підкріпленням заснований на тому, як реагує мозок щура та людини на покарання і заохочення. Ці алгоритми роблять висновки на основі отриманих результатів, а потім приймають рішення про подальші дії.

Як правило, навчання з підкріпленням підходить для автоматизованих систем, які повинні приймати багато дрібних рішень без втручання людини.



Алгоритми навчання з підкріпленням: "Що робити далі?"

Відповіддю на питання завжди є те, яку дію необхідно виконати. Зазвичай це відноситься до машини або робота. Нижче наведені деякі приклади.

Для системи клімат-контролю будинку: змінити температуру або залишити як є?

Для автомобіля з автономним управлінням: при жовтому сигналі світлофора загальмувати або прискоритися?

Для робота-пилососа: продовжити прибирання або повернутися до зарядної станції?

Алгоритми навчання з підкріпленням збирають дані в процесі роботи, навчаючись методом проб і помилок.

Таким чином, ми ознайомилися з основними питаннями, на які дають відповідь обробка і аналіз даних.

2. Критерії оцінки даних

"Чи готові ваші дані до обробки і аналізу?".

Щоб алгоритми обробки і аналізу даних могли відповідати на поставлені запитання, необхідно надати їм високоякісну "сировину", з яким вона буде працювати. Це схоже на процес приготування піци: чим якісніше інгредієнти, тим краще кінцевий продукт.

Критерії для даних

У сфері обробки і аналізу даних існують певні компоненти, які повинні бути:

релевантними (відповідними);

підключеними;

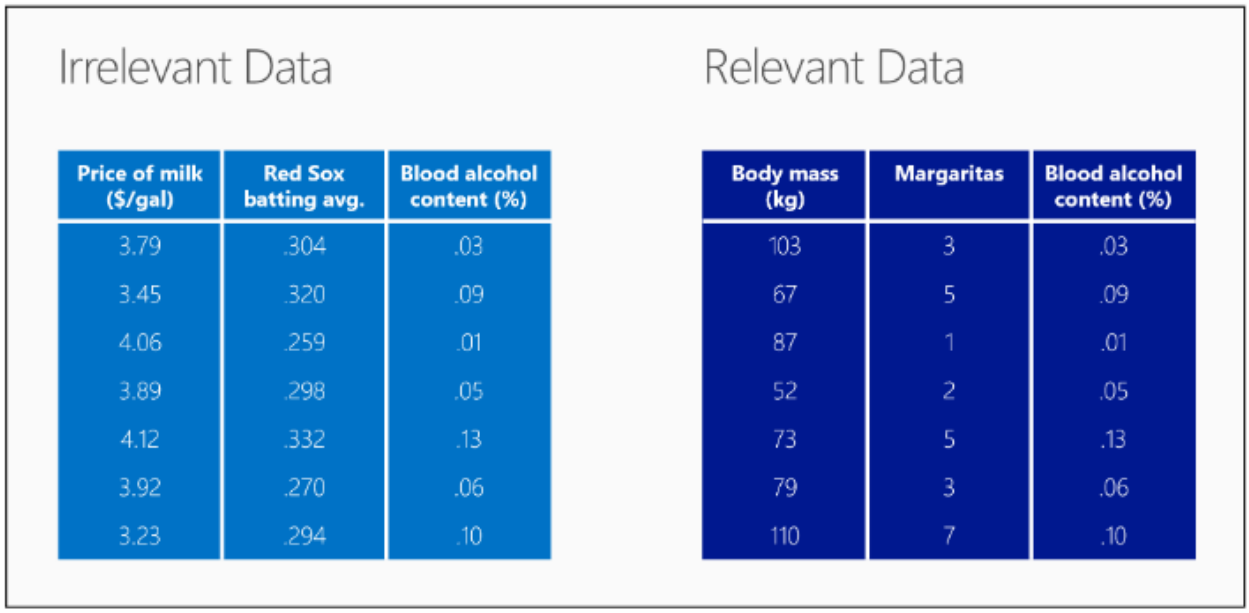
точними;

достатніми для використання в роботі.

Чи є дані релевантними (відповідними)?

Отже, перший критерій - дані повинні бути відповідними.

Порівняння відповідних і невідповідних даних - оцінка даних



Зліва в таблиці представлені наступні показники: рівень вмісту алкоголю в крові сімох людей, протестованих на виході з бару в Бостоні, середнє число успішних подач бейсбольної команди Red Sox в останній грі, а також ціна на молоко в найближчому гастрономі.

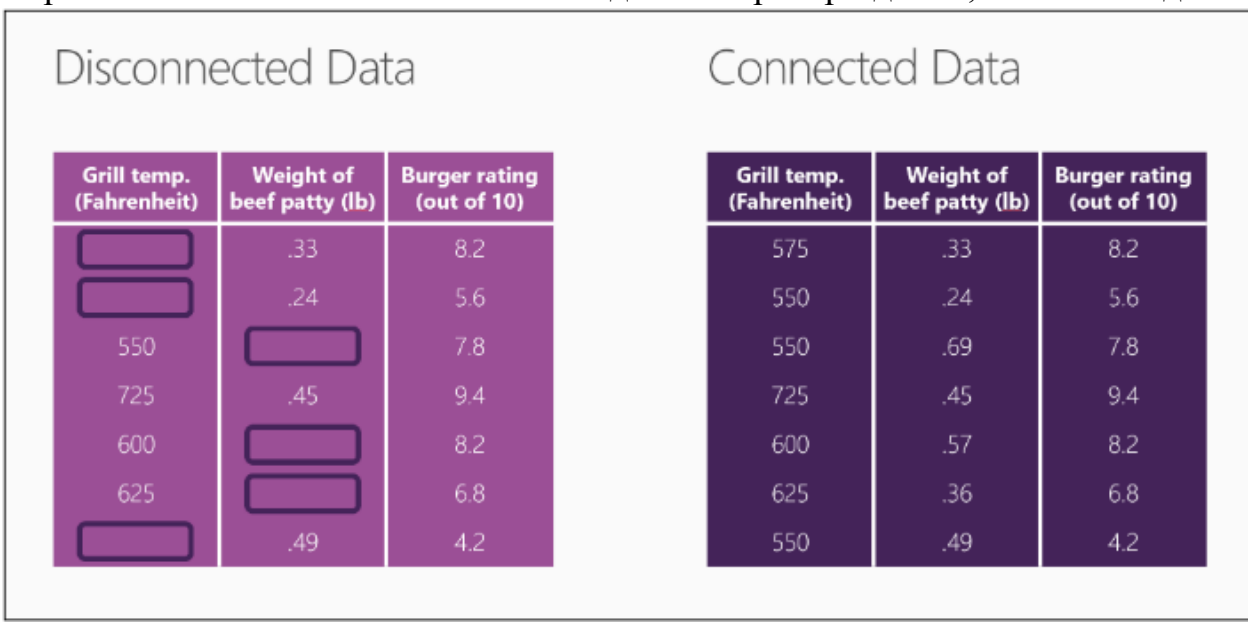
Всі ці дані, безумовно, є допустимими. Проблема лише в тому, що вони не є відповідними. Між цими числами не існує очевидного зв'язку. Якщо якась людина повідомить вам ціну на молоко і середнє число успішних подач команди Red Sox, це ніяк не допоможе вам визначити рівень вмісту алкоголю у нього в крові.

Тепер погляньте на таблицю в правій частині малюнка. На цей раз ми зважили кожного відвідувача і порахували кількість випитих ними напоїв. Тепер числа у всіх рядках відповідають один одному. Якщо я повідомлю вам свою вагу і кількість випитих алкогольних коктейлів, ви зможете приблизно вгадати рівень вмісту алкоголю у мене в крові.

Чи є дані підключеними?

Наступним критерієм є підключення даних.

Порівняння пов'язаних і непов'язаних даних - критерії даних, готовність даних



Ось деякі відповідні дані стосовно якості гамбургерів: температура гриля, вага котлети і рейтинг в місцевому спеціалізованому журналі. Але зверніть увагу на незаповнені клітинки в таблиці зліва.

У більшості наборів даних відсутні якісь значення. Такі прогалини часто зустрічаються, і існують способи їх усунення. Але якщо відсутня занадто багато, то дані починають виглядати як швейцарський сир.

Якщо поглянути на таблицю в лівій частині малюнка, то можна помітити, що багато даних відсутні. При цьому складно знайти яку-небудь взаємозв'язок між температурою гриля і вагою котлети. Цей приклад відключених даних.

Таблиця в правій частині малюнка вся заповнена - і це приклад підключених даних.

Чи є дані точними?

Наступний критерій - це точність. Тут потрібно забезпечити наступне.

Порівняння точних і неточних даних - критерії даних

Погляньте на мішень в правому верхньому куті. Все попадання щільно згруповані навколо центру. Звичайно, це приклад точних даних. Як не дивно, але на мові обробки і аналізу даних попадання, зображені на мішені в правому нижньому кутку, також вважаються точними.

Якщо вирахувати центр цих влучень, то він виявиться дуже близько до центру мішені. Попадання розкидані по всій мішені і не вважаються влучними, але всі вони знаходяться навколо одного центру мішені, тому вважаються точними.

Тепер погляньте на мішень в лівому верхньому куті. Тут все попадання знаходяться близько один до одного, тобто щільно згруповані. Вони влучні, але неточні, тому що зосереджені далеко від центру мішені. У лівому нижньому кутку показана мішень з влученнями, які не є ні достовірними, ні точними. Цьому лучники потрібно більше тренуватися.

Чи достатньо даних для використання в роботі?

Нарешті, останній компонент - достатній обсяг даних.

Чи достатньо даних для аналізу?

Уявіть собі, що кожна точка даних в таблиці - це мазок пензля на картині. Якщо таких мазків буде зовсім небагато, то картина вийде нечіткою і буде важко сказати, що на ній зображено.

У міру додавання штрихів картина стає все більш і більш чіткою.

Тільки при достатній кількості штрихів видно вже досить для того, щоб зробити якісь загальні висновки. Це місце, куди б я хотів поїхати? Виглядає яскраво, схоже на чисту воду ... Та, мабуть, я поїду сюди у відпустку.

Додаючи додаткові дані, картина стає чіткішою і можна робити більш конкретні висновки. Тепер можна розглянути три готелі на лівому березі. Можна особливо відзначити архітектуру готелю на передньому плані. Можливо, ви навіть виберете залишитися на третьому поверсі через вид, який звідти відкривається.

Таким чином, якщо дані є відповідними, підключеними, точними і їх обсяг достатній, то вони задовольняють всім критеріям для якісного виконання обробки і аналізу даних.

3. Правильна постановка правильного питання

Задайте питання, на яке можна відповісти за допомогою даних

Постановка точного питання

Ми говорили про те, що обробка і аналіз даних - це процес використання імен (також званих категоріями або мітками) і чисел для прогнозування відповіді на питання. Але це має бути не просто будь-яке питання, а точне запитання.

На загальне питання не обов'язково відповідати ім'ям або числом. А на точне запитання - обов'язково.

Уявіть собі, що ви знайшли чарівну лампу з джином, який чесно відповість на будь-яке ваше питання. Але це хитрий джин, який спробує зробити свою відповідь неясною і заплутаною, а потім зникнути.

Якщо задати йому загальне питання, а саме "Що станеться з моїми акціями?", То джин може відповісти: "Ціна зміниться". І це буде правдива відповідь, але не дуже корисна.

Але якщо поставити точне питання, а саме "Якою буде ціна продажу моїх акцій на наступному тижні?", Джин буде змушений дати конкретну відповідь і передбачити ціну продажу.

Приклади відповіді: цільові дані

Після того, як питання сформульоване, перевірте, чи містять ваші дані приклади відповідей.

Якщо наше запитання "Скільки коштуватимуть мої акції на наступному тижні?", Необхідно переконатися, що наші дані включають журнал котирування акцій.

Якщо наше запитання "Який автомобіль в моєму парку зламається першим?", Необхідно переконатися, що наші дані включають інформацію про попередні поломки.

Цільові дані - приклади відповіді.

Ці приклади відповідей називаються цільовими даними. *Цільові дані* - це те, що ми намагаємося спрогнозувати про майбутніх точках даних, будь то категорія або число.

Якщо у вас немає цільових даних, необхідно їх додати. Без цього ви не зможете отримати відповідь на питання.

Зміна формулювання питання

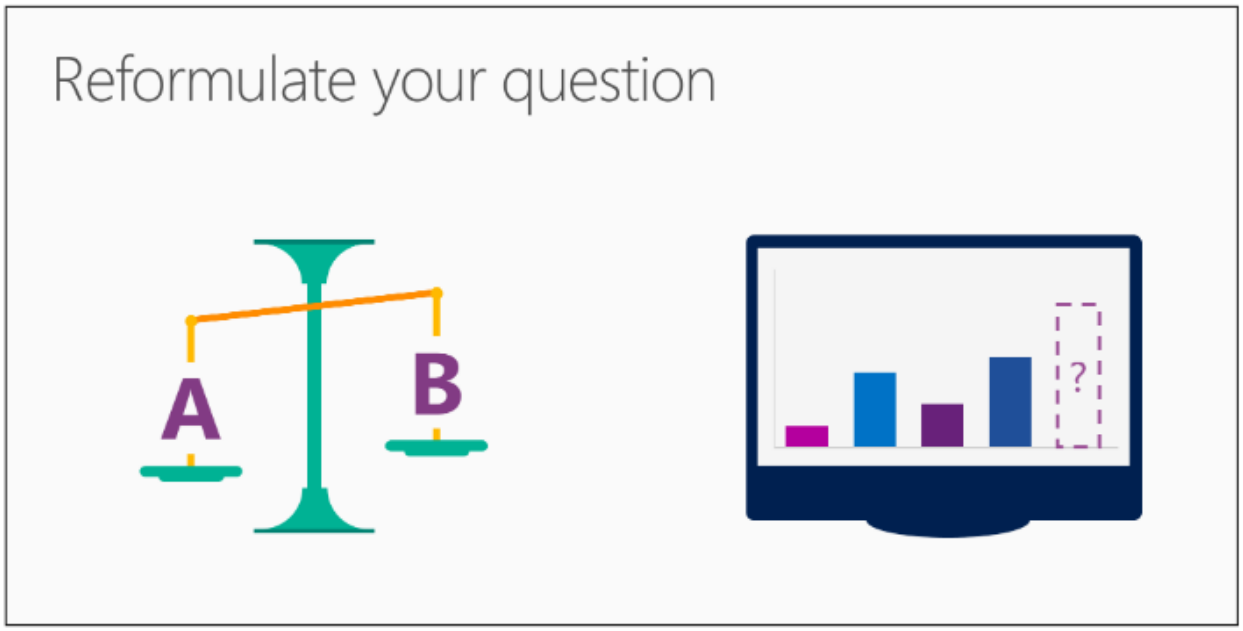
Іноді слід змінити текст питання, щоб отримати більш практичний відповідь.

Питання "*Чи вказують ці дані на A або B*?" прогнозує категорію (або ім'я або мітку) даних. Щоб відповісти на нього, використовується алгоритм класифікації.

Питання "*Скільки*?" або "*Як багато*?" прогнозує суму. Щоб відповісти на нього, використовується алгоритм регресії.

Щоб продемонструвати, як це можна змінити, давайте візьмемо для прикладу питання "Який новинний репортаж становить найбільший інтерес для читача?" Тут запитується прогноз одного варіанта з безлічі варіантів. Іншими словами: "Це A, B, C або D?" У цьому випадку використовується алгоритм класифікації.

Але відповісти на це питання буде простіше, якщо змінити його формулювання на наступну: "Наскільки цікава цьому читачеві кожна новина з цього списку?" Тепер кожній статті можна присвоїти числову оцінку, і тоді буде легко визначити статтю з найвищою оцінкою. Це приклад перефразувати питання класифікації в питання регресії або питання "Скільки?"



Змініть формулювання питання.

Те, як задається питання, є ключем до вибору алгоритму відповіді.

Можна помітити, що деякі групи алгоритмів (такі як в нашому прикладі з новинним репортажем) тісно пов'язані між собою. Формулювання питання можна змінити таким чином, щоб використовувався алгоритм, який дає найбільш практичну відповідь.

Але найголовніше - задайте точний питання, тобто питання, на яке можна відповісти даними. І переконайтеся, що у вас є необхідні дані для відповіді на нього.

Таким чином, ми обговорили деякі основні принципи формування питання, на яке можна відповісти за допомогою даних.

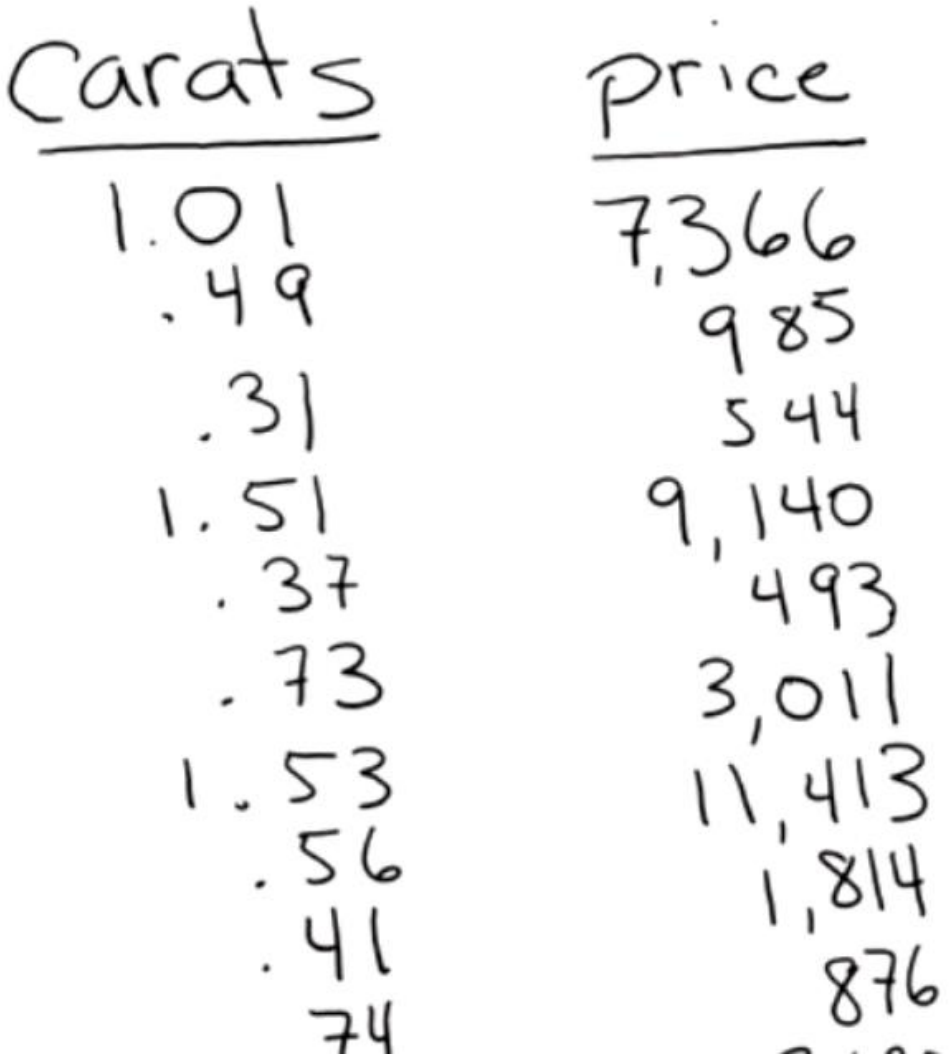
4. Прогнозування відповіді за допомогою простої моделі

Модель - це спрощена розповідь про дані.

Збір релевантних (відповідних), точних, підключених і достатніх даних.

Припустимо, ми хочемо купити діамант. У нас є кільце, яке належало моєї бабусі. Його оправа містить діамант масою 1,35 карата, і я хочу знати, скільки він буде коштувати. Я беру з собою блокнот і ручку і йду в ювелірний магазин. Там я записую всі ціни на діаманти, які є у вітрині, а також їх масу в каратах. Починаючи з першого діаманта, який має масу 1,01 карата і стоїть 7366 дол. США.

Я йду далі і роблю те ж саме для інших діамантів в магазині.



Стовпці з даними діамантів

Зверніть увагу, що наш список містить два стовпці. Кожен стовпець має свій атрибут - масу в каратах і ціну - і кожен рядок є окремим точкою даних, що представляє один діамант.

Фактично ми створили невеликий набір даних у формі таблиці. Зверніть увагу, що він відповідає нашим критеріям якості.

Дані є релевантними: маса безумовно пов'язана з ціною.

Вони точні: ми перевірили ціни, які записали.

Вони пов'язані: всі стовпці і рядки заповнені.

І, як ми побачимо далі, цих даних достатньо, щоб дати відповідь на наше питання.

Постановка точного питання

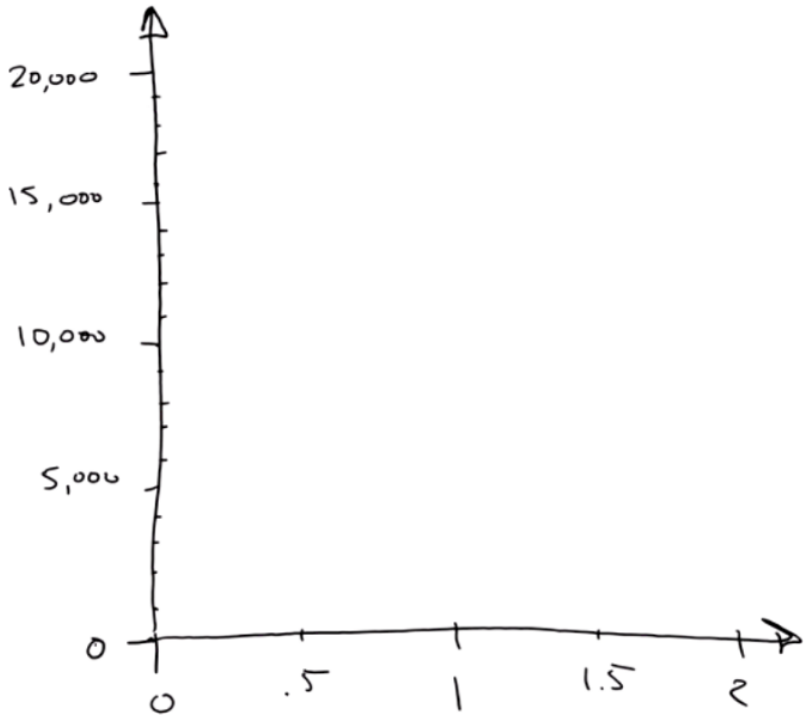
Тепер давайте точно сформулюємо наше запитання: "Скільки буде коштувати діамант масою 1,35 карата?"

У нашому списку немає діаманта масою 1,35 карата, тому необхідно використовувати наявні дані для отримання відповіді на це питання.

Побудова існуючих даних

Перше, що необхідно зробити, це намалювати горизонтальну числову лінію, яку називають віссю. На ній буде відображатися маса. Діапазон мас - від 0 до 2, тому ми намалюємо лінію, що охоплює цей діапазон, і помістимо на неї позначки для кожних 0,5 карата.

Потім ми намалюємо вертикальну вісь, на якій запишемо ціни, і з'єднаємо її з горизонтальною віссю маси. Одиницею виміру для цієї осі буде долар США. Тепер у нас є набір осей координат.

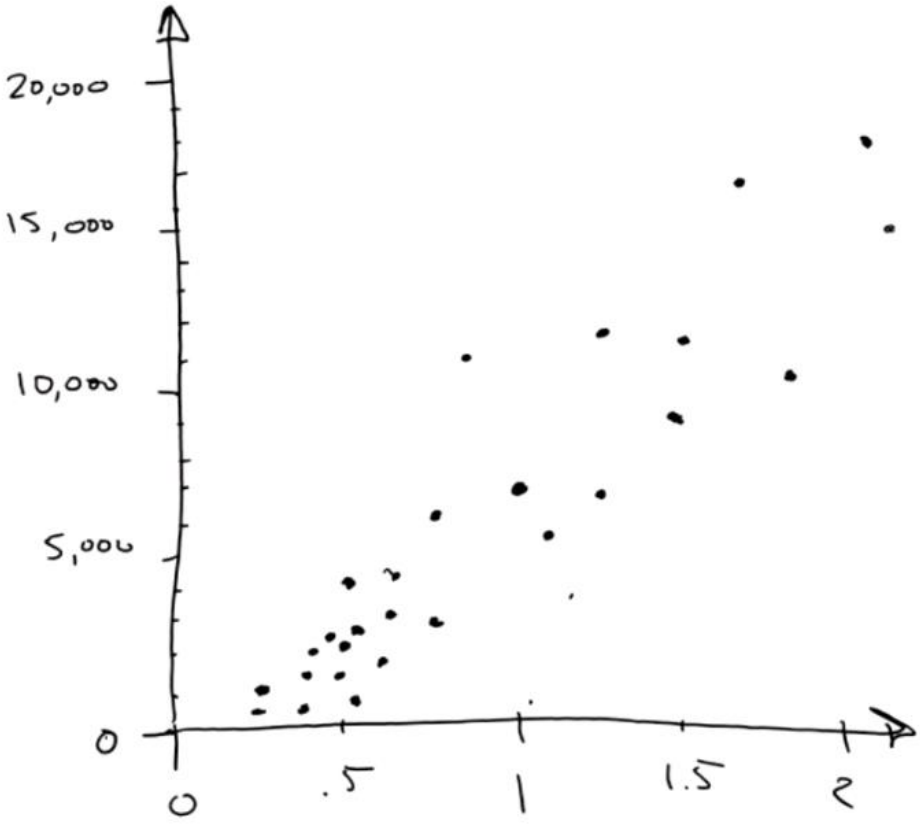


Осі маси і ціни

Тепер ми перетворимо ці дані в точкову діаграму. Це відмінний спосіб візуалізації числових наборів даних.

Беремо першу точку даних і проводимо вертикальну лінію від позначки 1,01 карата. Потім проводимо горизонтальну лінію від позначки 7366 дол. США. Там, де ці лінії перетинаються, малюємо точку. Ця точка і представляє наш перший діамант.

Тепер ми пройдемо по всім діамантів з нашого списку і зробимо те ж саме. В результаті ми отримаємо таку картину: безліч точок, кожна з яких відповідає одному діаманту.

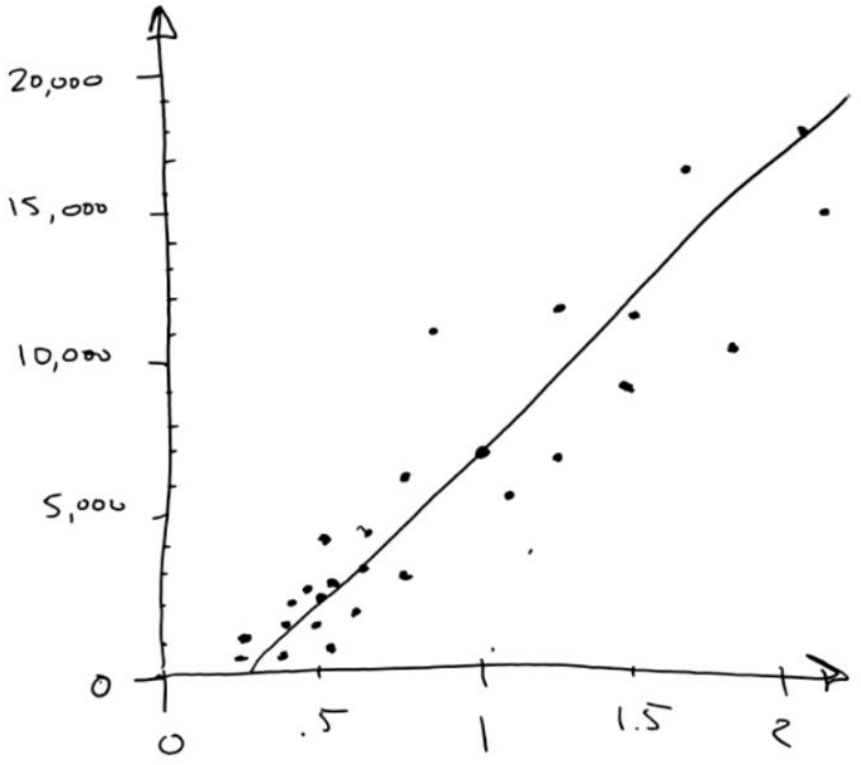


точкова діаграма

Побудова моделі на основі точок даних

Тепер, якщо поглянути на точки під невеликим кутом, то весь цей набір виглядає як товста і нечітка лінія. Можна скористатися маркером і провести через набір точок пряму лінію.

Намалювавши лінію, ми створили модель. Щоб краще зрозуміти, уявіть собі, що реальний світ зображений у спрощеній формі, як комікс. Комікс не відображає всієї дійсності: лінія не проходить через всі крапки даних. Але це корисне спрощення.



Лінія лінійної регресії

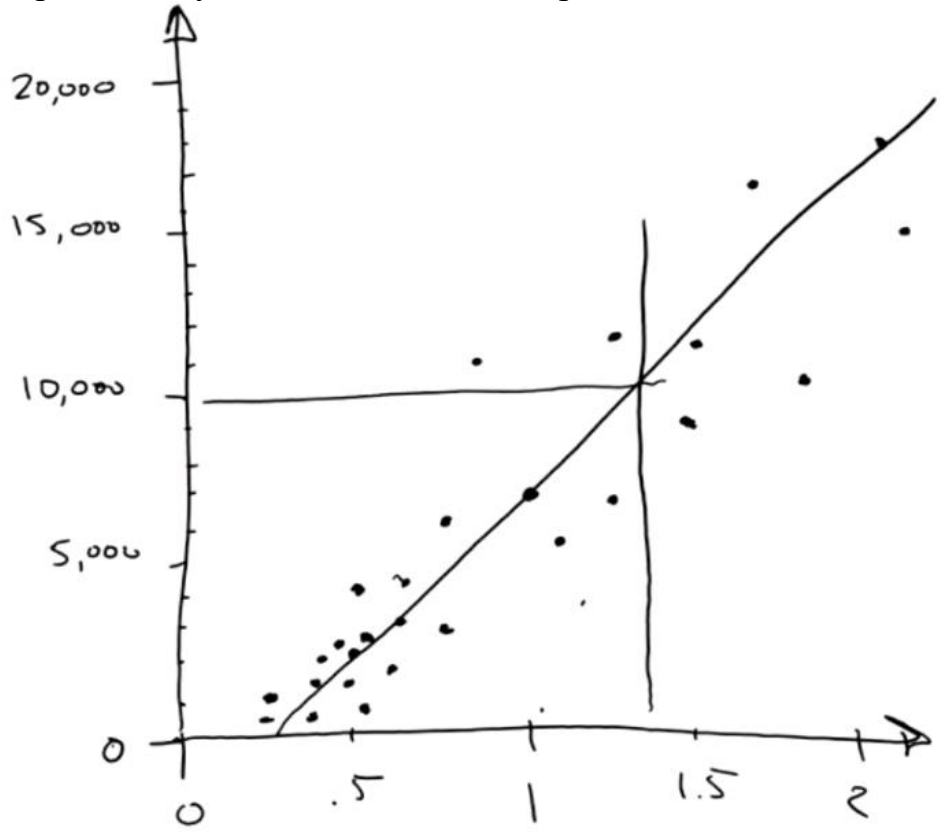
Той факт, що лінія не проходить через всі крапки, є нормальним. Фахівці з обробки даних пояснюють це так: існує модель (в нашому випадку - лінія), і кожна точка в моделі піддається впливу деякого шуму або відхилення, пов'язаного з нею. Є базова функціональна зв'язок, а є мінливий реальний світ, який додає шум і невизначеність.

Так як ми намагаємося відповісти на питання Скільки? , Це називається регресією. І оскільки ми використовуємо пряму лінію, це - лінійна регресія.

Використання моделі для пошуку відповіді

Тепер у нас є модель і ми можемо поставити їй наше запитання: "Скільки буде коштувати діамант масою 1,35 карата?"

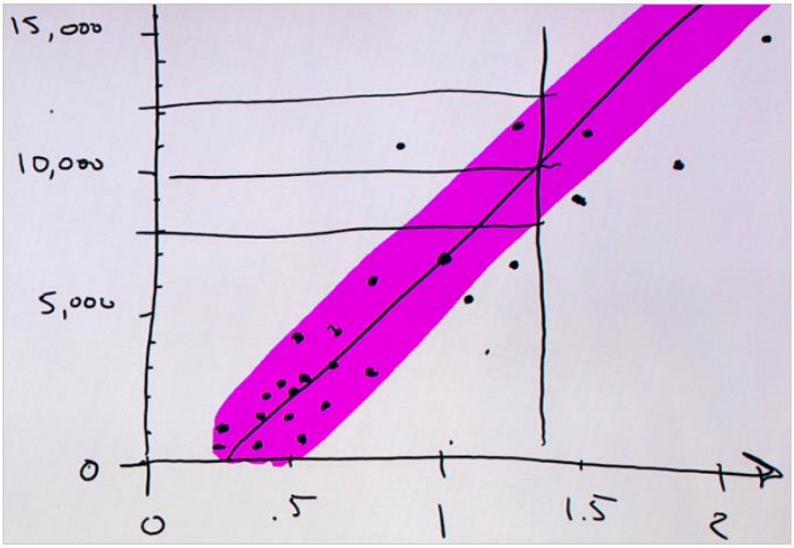
Для відповіді на питання ми проводимо вертикальну лінію від позначки 1,35 карата. Від точки її перетину з лінією моделі проводимо горизонтальну лінію до осі цін. Вона приводить нас до позначки 10 000. Ось так! Це і є відповідь: діамант масою 1,35 карата коштує близько 10 000 доларів США.



Пошук відповіді за допомогою моделі

Створення довірчого інтервалу

Було б логічним перевірити, наскільки точний цей прогноз. Корисно знати, чи буде діамант масою 1,35 карата коштувати рівно 10 000 доларів США, а може бути набагато більше або менше. Щоб це визначити, давайте намалюємо навколо лінії регресії так званий конверт, який буде включати в себе більшість точок. Цей конверт - наш довірчий інтервал. Ми можемо бути впевнені, що ціни потраплять в зазначений діапазон, адже це справедливо для більшості цін в минулому. Можна провести ще дві горизонтальні лінії від точок перетину лінії, проведеної від позначки 1,35 карата, з верхньою і нижньою межами даного конверта.



довірчі інтервали

Тепер ми можемо щось сказати про наш довірчому інтервалі. Можна з упевненістю говорити, що ціна діаманта 1,35 карата складе близько 10 000 дол. США, при цьому вона може бути не нижче 8 000 і не вище 12 000 долл. США.

Все готово, без математичних формул і комп'ютерів

Ми зробили те, за що платять фахівцям з обробки даних, при цьому ми користувалися тільки малюванням.

Ми поставили запитання, на яке можна відповісти за допомогою даних.

Ми створили модель, використовуючи лінійну регресію.

Ми зробили прогноз, доповнений довірчим інтервалом.

При цьому ми не користувалися математичними формулами або комп'ютерами.

Якби у нас були додаткові відомості, такі як:

огранювання діаманта;

відтінки кольорів (наскільки близький колір діаманта до білого);

наявність в камені сторонніх включень;

то ми б мали додаткові стовпці. В цьому випадку математика вже буде корисною. Коли є більше двох стовпців, важко намалювати точки на папері. Математика дозволить відмінно вписати отриману лінію або площину в ваші дані.

Крім того, якби замість невеликого переліку діамантів у нас було б дві тисячі або два мільйони каменів, то набагато швидше цю роботу можна було б зробити за допомогою комп'ютера.

Сьогодні ми поговорили про те, як створити лінійну регресію, а також ми зробили прогноз, використовуючи дані.

Як будується модель

Лінійна регресійна модель має наступний вигляд:

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \qquad (2.1)$$

де y– залежна змінна;

$x_1, x_2, \ldots, x_n$– незалежні змінні;

u– випадкова похибка, розподіл якої в загальному випадку залежить від незалежних змінних, але математичне очікування якої рівне нулю.

Згідно з моделлю (2.1), математичне очікування залежної змінної є лінійною функцією незалежних змінних:

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \qquad (2.2)$$

$E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u$ Вектор параметрів $\beta_0, \beta_1, \ldots, \beta_k$ є невідомим, і задача лінійної регресії полягає в оцінці цих параметрів на основі деяких експериментальних значень y і $x_1, x_2, \ldots, x_n$. Тобто, для деяких n експериментів є відомі значення $\{y_i, y_{i1}, \ldots, y_{ip}\}_{i=1}^{n}$ незалежних змінних і відповідне їм значення залежної змінної.

Згідно з визначенням моделі для кожного експериментального випадку залежність між змінними визначається формулою:

$$y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k + u \qquad (2.3)$$

або у матричних позначеннях.

На основі цих даних потрібно оцінити значення параметрів $(\beta_0, \beta_1, \ldots, \beta_k)$, а також розподіл випадкової величини u. Зважаючи на характеристики досліджуваних змінних, можуть додаватися різні додаткові специфікації моделі і застосовуватися різні методи оцінки параметрів. Серед найпоширеніших специфікацій лінійних моделей є класична модель лінійної регресії та узагальнена модель лінійної регресії.

Згідно з класичною моделлю лінійної регресії [31] додатково вводяться такі вимоги щодо специфікації моделі та відомих експериментальних даних:

$\forall i \neq jE(u_i u_j | x_i) = 0$(відсутність кореляції залишків);

$\forall iE(u_j^2 | x_i) = \sigma^2$(гомоскедастичність).

Часто додається також умова нормальності випадкових відхилень, яка дозволяє провести значно ширший аналіз оцінок параметрів та їх значимості, хоча і не є обов'язковою для можливості використання наприклад методу найменших квадратів $(u_j | x_i) \sim N(0, \sigma^2)$.

Умови гомоскедастичності та відсутності кореляції між випадковими залишками в моделі часто на практиці не виконуються. Якщо замість цих двох умов у визначенні моделі взяти загальнішу умову:

$$V(u|X) = \sigma^2 W, \qquad (2.4)$$

де W– відома додатноозначена матриця, то одержана модель називається узагальненою моделлю лінійної регресії.

Оскільки для кожної додатньо означеної матриці W існує матриця N, така, що $W^{-1} = NN$, то модель:

$$N_y = NX\beta + Nu, \qquad (2.5)$$

В залежності від об'єктів, що досліджуються за допомогою лінійної регресії, та конкретних цілей дослідження, можуть використовуватися різні методи оцінки невідомих коефіцієнтів.

Найпопулярнішим є звичайний метод найменших квадратів.

$$f = a_0 + a_1 x + \varepsilon, \qquad (2.6)$$

де a_0, a_1– постійні коефіцієнти, що називаються параметрами моделі;
ε – випадкова величина з математичним сподіванням 0 і дисперсією s2 .
В цьому випадку рівняння регресії перетворюється на рівняння прямої:

$$\overline{y}(x) = M(y/x) = a_0 + a_1 x. \qquad (2.7)$$

Передбачимо, що незалежна змінна набула значень $x_1, x_2 \dots, x_n$, внаслідок чого залежна змінна набула значень $y_1, y_2 \dots, y_n$. У припущенні лінійної залежності отримуємо n рівностей.

$$y_i = a_0 + a_1 x_i + \varepsilon_i, i = 1 \dots n, \qquad (2.8)$$

де ε_i– незалежні і розподілені так само, як ε.
Потрібно за значеннями пар (x_i, y_i) оцінити невідомі a_0, a_1. Як ми вже знаємо, кожне завдання оцінювання пов'язане з деяким критерієм якості. У теорії, що викладається нами, таким критерієм є критерій найменших квадратів:

$$\sum_{i=1}^{n} \varepsilon_i^2 - \min. \qquad (2.9)$$

Запишемо цю суму інакше, так, щоб була помітна залежність від a_0, a_1:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [\overline{y}(x) - y_i]^2 = \sum_{i=1}^{n} [y_i - a_0 - a_1 x_1]^2 \qquad (2.10)$$

Тепер остаточно приходимо до такої задачі: знайти такі значення невідомих параметрів a_0, a_1щоб функція 2.11 набула найменшого значення:

$$Q(a_0 a_1) = \sum_{i=1}^{n} [y_i - a_0 - a_1 x_1]^2 \qquad (2.11)$$

Метод розв'язання цієї задачі відомий з курсу вищої математики. Знаходимо часткові похідні функції Q і прирівнюємо їх до нуля, внаслідок чого приходимо до системи лінійних рівнянь:

$$\begin{cases} \frac{\partial Q}{\partial a_0} = -2 \sum_{i=1}^{n} [y_i - a_0 - a_1 x_i] = 0, \\ \frac{\partial Q}{\partial a_1} = -2 \sum_{i=1}^{n} [y_i - a_0 - a_1 x_i] x_i = 0. \end{cases} \qquad (2.12)$$

Після очевидних перетворень отримуємо систему:

$$\begin{cases} na_0 + a_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_{i,} \\ a_0 \sum_{i=1}^{n} x_i + a_i \sum_{i=1}^{n} {x_i}^2 = \sum_{i=1}^{n} x_i y_i \end{cases} \qquad (2.13)$$

Позначимо вибіркові середні:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \,, \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i \,, \overline{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i \, y_{i,} \overline{x}^2 = \frac{1}{n}\sum_{i=1}^{n} {x_i}^2. \qquad (2.14)$$

У цих позначеннях після ділення кожного рівняння системи на n вона набуде вигляду:

$$\begin{cases} a_0 + a_1 \overline{x} = \overline{y}, \\ a_x \overline{x} + a_1 \overline{x}^2 = \overline{xv} \end{cases} \qquad (2.15)$$

а її рішення (шукані оцінки коефіцієнтів рівняння регресії) буде таким:

$$\hat{a}_0 = \frac{\overline{x^2} \cdot \overline{y} - \overline{xy} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2} \qquad (2.16)$$
$$\hat{a}_0 = \frac{\overline{xy} - \overline{y} \cdot \overline{x}}{\overline{x^2} - (\overline{x})^2}$$

Якщо ввести ще позначення $S_x^2 = \overline{x^2} - \overline{(x)}^2$ і перетворити вираз $\hat{a}_0$:

$$\hat{a}_0 = \frac{\overline{x^2} \cdot \overline{y} - \overline{xy} \cdot \overline{x} \pm \overline{y} \cdot (\overline{x})^2}{S_{..}^2} = \frac{S_x^2 \cdot \overline{y} - \overline{x}(\overline{xy} - \overline{y} \cdot \overline{x})}{S_{..}^2} = \overline{y} - \hat{a}_1 \cdot \overline{x}, \qquad (17)$$

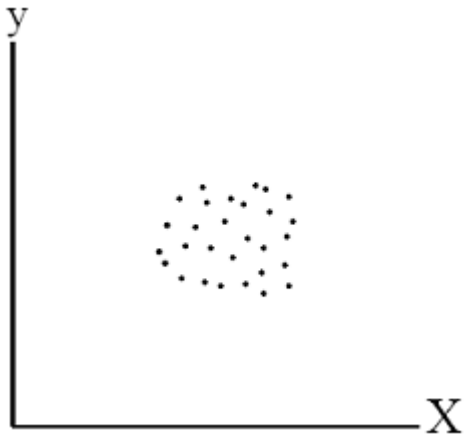
то оцінка функції регресії набуде такого вигляду:

$$\tilde{\overline{y}}(x) = \hat{a}_0 + \hat{a}_1 x = \overline{y} - \hat{a}_1 \cdot \overline{x} + \hat{a} \cdot x = \overline{y} + \hat{a}_1 (x - \overline{x}) \qquad (18)$$

Таким чином, отримали рівняння регресії, що є моделлю для опису даних.

Порівняння змінних і кореляція

Найбільш наглядний приклад показати зв'язок між двома кількісними змінними – це діаграма розсіювання. На відміну від гістограм, на осі у показують не частоту того чи іншого значення змінної по осі х, а значення іншої змінної. Крапка на діаграмі означає одночасно значення двох змінних для одного спостереження («рядок» в таблиці даних).



Діаграма розсіювання

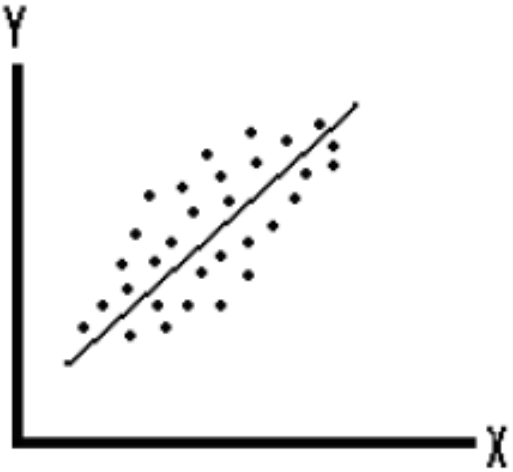
У кореляції є дві властивості – сила і напрям. Сила кореляції визначається числовим значенням, а напрям – тим, чи кореляція позитивна чи негативна.

Позитивна кореляція: обидві змінні міняються у тому ж напрямі. Тобто, якщо одна змінна зростає, друга зростає теж. Якщо одна спадає, то друга спадає так само.

Наприклад, рівень освіти – скільки років людина навчалася (в нормальних країнах) та річний заробіток корелюють між собою позитивно.

Негативна кореляція: змінні рухаються у протилежних напрямках. По мірі того, як одна змінна спадає, інша росте, і навпаки.

Наприклад – кількість годин, проведених людиною уві сні та кількість годин неспання – корелюють негативно (що очевидно – чим більше спиш – тим менше часу лишається на яві, і навпаки).



Позитивна кореляція

Негативна кореляція

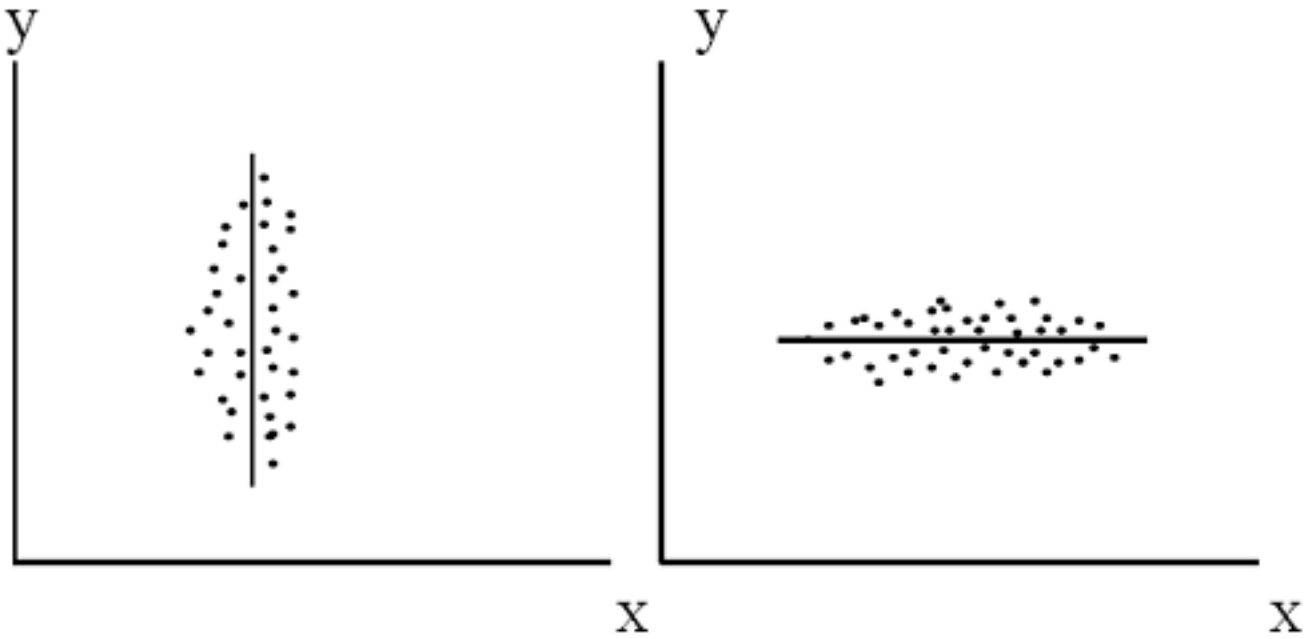
Негативна і позитивна кореляції

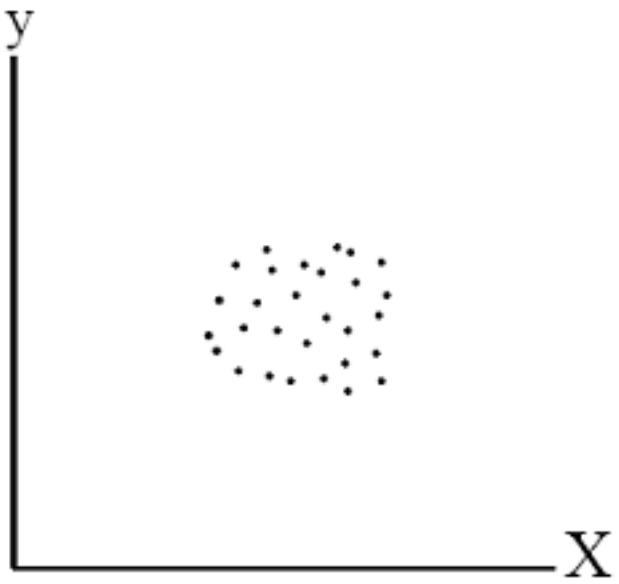
Коефіцієнт кореляції показує ступінь, до якого дві змінні пов’язані (наскільки спільно чи подібно змінюються їх значення для різних спостережень) – тобто якої сили між ними може бути зв’язок. Значення коефіцієнта кореляції може бути від -1.0 до 1.0. Якщо вирахувана кореляція більша за 1 або менша за -1 – значить десь у підрахунках сталася помилка, адже 1 – означає абсолютну пряму (позитивну) кореляцію, а -1 – абсолютну зворотню (негативну) кореляцію.

Як підраховується коефіцієнт кореляції? Дорівнює сумі добутків відхилень, поділеній на добуток їх стандартних відхилень

Що означає, коли ми кажемо, що між двома змінними нема кореляції? Це означає, що між двома змінними немає прямого зв’язку. Наприклад, немає прямої кореляції між

розміром взуття та зарплатою. Тобто, великі значення розміру взуття мають такі ж шанси зустрітися серед людей з високою зарплатою, як із низькою.



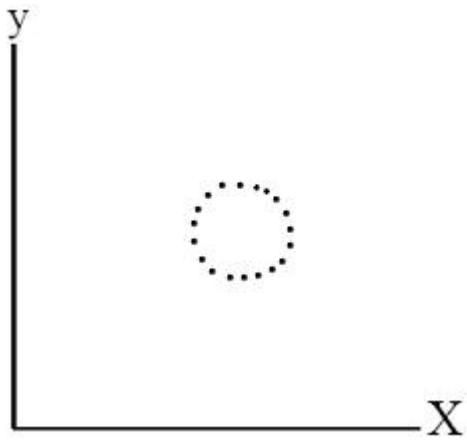


Всі три приклади показують випадки, в яких дуже мала або відсутня кореляція

Кореляція і причинно-наслідковий зв'язок. Навіть якщо дві змінні виглядають пов'язаними між собою, це не значить, що одна спричинила іншу. Класичний приклад – це кореляція між ростом злочинності та споживанням морозива протягом літніх місяців у США. Дві змінні є пов'язані між собою, але жодне явище не є причиною іншого. Насправді, обидва явища спричинені підвищенням температури повітря, а не одне одним.

Важливо також пам'ятати, що кореляція – це міра лінійного зв'язку. При цьому, кореляція не говорить нам, яка змінна впливає на яку – кореляція лише показує наявність зв'язку, але впливу. Вимірюючи кореляцію, не можна сказати – це А впливає на Б, чи Б впливає на А.

Діаграма розсіяння для двох змінних може виглядати, наприклад, так:



Для цих двох змінних кореляція буде дорівнювати нулю. Але це ще не означає, що зв’язку між змінними немає — просто він може бути не лінійним.