ЛАБОРАТОРНА РОБОТА № 2

БАЙЄСІВСЬКИЙ АНАЛІЗ ДАНИХ

Теоретичні відомості

Якщо у результаті досліду може відбутися декілька подій $A_1, A_2, ..., A_k$, то має місце **сукупність** або **група випадкових подій**. **Несумісними** (або взаємно виключеними) подіями є події, які одночасно відбуватися не можуть.

Сумою двох подій A_1 і A_2 називають подію $S = A_1 + A_2$, яка полягає в тому, що в результаті досліду відбувається хоча б одна (незалежно яка) з подій A_1 або A_2 чи A_1 і A_2. Це твердження є змістом **постулату адитивності ймовірності**.

Якщо події A_1 і A_2 – несумісні, то $S = A_1 + A_2$ – подія, що полягає в появі однієї з подій A_1 або A_2 (незалежно якої). Зауважимо, що таке означення суми двох подій повністю тотожне логічній операції ***АБО***. Звідси випливає, що ймовірність появи однієї з двох несумісних подій A_1 і A_2 дорівнює сумі ймовірностей цих подій:

$$P\{S\} = P\{A_1\ або\ A_2\} = P\{A_1\} + p\{A_2\}. \qquad (1)$$

Узагальнивши цю формулу для випадку k попарно несумісних подій $A_i\ (i = 1, 2...k)$, отримаємо

$$P\{S\} = P\{A_1\ або\ A_2\} = P\{A_1\} + p\{A_2\} \qquad (2)$$

Сукупність несумісних подій $A_i\ (i = 1 \div k)$ утворює **повну групу**, якщо внаслідок досліду обов'язково відбувається одна з них, тобто якщо сума S є достовірною подією:

$$P\{S\} = \sum_{i=1}^{k} P\{A_i\} = 1 \qquad (3)$$

Співвідношення (3) виражає зміст постулату **нормування ймовірності**.

Для **двох сумісних** події A і B з ймовірностями $P\{A\}$ і $P\{B\}$ та імовірністю їхньої сумісної появи $P\{AB\}$, ймовірність суми подій $S = A + B$ дорівнює сумі ймовірностей цих подій без імовірності їх спільної появи:

$$P\{S\} = P\{A + B\} = P\{A\} + P\{B\} - P\{AB\}. \qquad (4)$$

Добутком подій A і B називають подію $M = AB$, яка полягає в тому, що в результаті досліду відбуваються як подія A, так і подія B. Означення добутку двох подій повністю співпадає з логічною операцією ***І***.

Очевидно, що якщо A і B є несумісними, то M – неможлива подія.

Нехай при N випробуваннях подія A відбулася N_A разів, подія B – N_B разів, а у N_{AB} випадках із N мали місце відразу обидві події A і B. Тоді при $N, N_A, N_B, N_{AB} \to \infty$ можна записати такі вирази для ймовірностей:

$$P\{A\} = \lim_{N\to\infty} \frac{N_A}{N}, \quad P\{B\} = \lim_{N\to\infty} \frac{N_B}{N}, \quad P\{AB\} = \lim_{N\to\infty} \frac{N_{AB}}{N}. \tag{5}$$

Останній вираз у (5) визначає ймовірність сумісної реалізації подій A і B.

Важливим є поняття **умовної ймовірності,** при якій реалізується подія B за умови, що відбулася (обов'язково!) подія A:

$$P\{B \mid A\} \equiv P_A\{B\} = \lim_{N\to\infty} \frac{N_{AB}}{N_A}.$$

Аналогічно можна записати й умовну ймовірність здійснення події A за умови, що подія B обов'язково має місце:

$$P\{A \mid B\} \equiv P_B\{A\} = \lim_{N\to\infty} \frac{N_{AB}}{N_B}.$$

Оскільки $\frac{N_{AB}}{N} = \frac{N_{AB}}{N_A}\frac{N_A}{N} = \frac{N_{AB}}{N_B}\frac{N_B}{N}$, то між умовними і звичайними (тобто безумовними) ймовірностями має місце таке співвідношення:

$$P\{M\} = P\{AB\} = P\{A \mid B\}P\{B\} = P\{B \mid A\}P\{A\}. \tag{6}$$

Умовну ймовірність називають **апостеріорною** (або «після дослідною» ймовірністю), а $P\{A\}$ – **апріорною** («перед дослідною»). З (6) отримаємо

$$P\{A \mid B\} = \frac{P\{AB\}}{P\{B\}} \; або \; P\{B \mid A\} = \frac{P\{AB\}}{P\{A\}}. \tag{7}$$

Дві події A і B називаються **незалежними**, якщо умовна ймовірність події $P\{A \mid B\}$ збігається з «безумовною» $P\{A\}$, тобто ймовірність появи події A не залежить від того, чи відбулася подія B (або навпаки ймовірність появи події B не залежить від того, чи відбулася подія A). При цьому з (6) випливає, що ймовірність добутку незалежних подій є добутком ймовірностей цих подій

$$P\{M\} = P\{AB\} = P\{A\} \cdot P\{B\}. \tag{8}$$

У загальному випадку для довільного числа незалежних подій множення ймовірностей визначається співвідношенням

$$P\{M\} = P\{A_1A_2...A_k\} = P\{A_1\} \cdot P\{A_2\}...P\{A_k\} = \prod_{i=1}^{k} P\{A_i\} \qquad (9)$$

Повна ймовірність. Формула Байєса.

Виведемо вираз для обрахунку **повної ймовірності**. Вважатимемо, що подія A може відбутися тільки з однією із k несумісних подій $B_1, B_2, ...B_k$, які утворюють повну групу і розглядаються як **гіпотези**, які пов'язані з появою події A.

Оскільки події AB_i та AB_j при $i \neq j$ є несумісними, то подію A можна розглядати як суму подій $AB_1, ... AB_k$, тому з теореми додавання ймовірностей (3) отримаємо:

$$P(A) = \sum_{i=1}^{k} P(AB_i) . \qquad (10)$$

З теореми множення (8) випливає також, що $P(AB_i) = P(B_i)P(A \mid B_i)$, тому співвідношення (10) перепишеться у вигляді:

$$P(A) = \sum_{i=1}^{k} P(B_i)P(A \mid B_i) . \qquad (11)$$

Отримана формула повної ймовірності дозволяє порахувати ймовірність $P(A)$ події A, якщо є відомими ймовірності $P(B_i)$ кожної гіпотези B_i та умовна ймовірність $P(A \mid B_i)$ події A за умови, що гіпотеза B_i є правдивою.

Знаючи ймовірність $P(B_i)$ та умовні ймовірності $P(A \mid B_i)$ з теореми множення можна розрахувати умовні ймовірності $P(B_i \mid A)$ гіпотез B_i:

$$P(AB_i) = P(B_i)P(A \mid B_i) = P(A)P(B_i \mid A).$$

З останнього виразу отримуємо:

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{P(A)}.$$

Підставляючи в отриману формулу вираз (11) для $P(A)$, отримаємо відому так звану **формулу Байєса**:

$$P(B_i \mid A) = \frac{P(B_i)P(A \mid B_i)}{\sum_{i=1}^{k} P(B_i)P(A \mid B_i)} \qquad (12)$$

Як зазначалося вище, формула Байєса дає можливість оцінити апостеріорні ймовірності подій, тобто подій, які відбудуться з урахуванням попередньо проведених дослідів.

Проілюструємо застосування формули Байєса на прикладі. Розглянемо три урни, у яких є по 10 кульок білого та чорного кольору. Нехай у першій урні є 8 білих та 2 чорні кулі, у другій урні – 5 білих та 5 чорних куль, а у третій – 2 білі і 8 чорних куль. Далі проводимо експеримент, який полягає у тому, що випадково вибирається одна з урн, а потім із вибраної урни вибирається дві кульки. Завдання полягає у тому, щоб за результатами проведення експерименту встановити, яка з трьох урн була вибрана.

У розглядуваному прикладі існує три гіпотези $B_k,\ k=1,2,3$, які полягають у тому, що вибрано одну з трьох урн. Апріорні ймовірності таких подій є рівними $\frac{1}{3}$.

Нехай подія A полягає у тому, що вибрано 2 чорні кулі.

Тоді $P(A\,|\,B_1)=\frac{2}{10}\cdot\frac{1}{9}=\frac{1}{45}$, $P(A\,|\,B_2)=\frac{5}{10}\cdot\frac{4}{9}=\frac{10}{45}$, $P(A\,|\,B_3)=\frac{8}{10}\cdot\frac{7}{9}=\frac{28}{45}$.

Тому для апостеріорних ймовірностей отримуємо такі значення:

$$P(B_1\,|\,A)=\frac{\frac{1}{45}\cdot\frac{1}{3}}{\frac{1}{3}\left(\frac{1}{45}+\frac{10}{45}+\frac{28}{45}\right)}=\frac{1}{39},\quad P(B_2\,|\,A)=\frac{\frac{10}{45}\cdot\frac{1}{3}}{\frac{1}{3}\left(\frac{1}{45}+\frac{10}{45}+\frac{28}{45}\right)}=\frac{10}{39},\quad P(B_3\,|\,A)=\frac{\frac{28}{45}\cdot\frac{1}{3}}{\frac{1}{3}\left(\frac{1}{45}+\frac{10}{45}+\frac{28}{45}\right)}=\frac{28}{39}.$$

Таким чином, з ймовірністю, рівною приблизно 0,72 можна стверджувати, що справедливою є третя гіпотеза, тобто була вибрана третя урна.

Аналогічним способом можна порахувати ймовірності інших подій A, які, наприклад, полягають у тому, що було вибрано білу і чорну кульки (незалежно від порядку, подія A_1), вибрано 2 білі кульки (подія A_2) і т.д. Оцінивши для цих подій апостеріорні ймовірності, можемо зробити висновок про ймовірність реалізації тої чи іншої події $B_k,\ k=1,2,3$.

У цьому полягає суть **Байєсівського аналізу даних,** який ґрунтується на проведенні певного числа дослідів, за результатами яких обраховуються апостеріорні ймовірності гіпотез.

Приклад з дайсом (dice)

Припустимо, маємо 3 гральні кубики (далі **дайси**). Перший кубик має 4 грані, другий – 6 граней, третій – 8 граней. Вибираємо один дайс випадково, котимо і отримуємо результат – випала грань із числом 2. **Завдання полягає у**

тому, щоб вгадати, який кубик було вибрано. Для цього потрібно обчислити ймовірність того, що дайс був з 4-ма гранями (4d дайс), з 6-ма (6d дайс) чи з 8-ма гранями (8d дайс).

Оскільки маємо 3 дайси, то *початкова ймовірність* $P(B_i)$ вибору кожного з них однакова і дорівнює 1/3.

Зауважимо:

- якщо число, що випало у результаті кочення випадково вибраного дайсу, більше за розмірність дайсу, то ймовірність його появи на дайсі рівна 0 (тобто, якщо випало число 6, то ймовірність появи цього числа на 4d-дайсі нульова);
- якщо число, що випало, менше за розмірність дайсу, то ймовірність його появи на дайсі складає 1/(розмірність дайсу) (тобто, якщо випало число 6, то ймовірність появи цього числа на 8d-дайсі дорівнює 1/8).

Наше спостереження (*подія A*) – випало число 2. Якщо ми припускаємо, що кидали 4d-дайс, то ймовірність випадання числа 2 ($P(A|B_i)$) дорівнює 1/4. Якщо 6d – то 1/6. Якщо 8d – то 1/8.

Перемножуємо отримані ймовірності:

- для 4d дайса - ймовірність вибору дайса 1/3, ймовірність випадання числа 2 - 1/4. Добуток дорівнює 1/3 * 1/4 = 1/12.
- для 6d дайса: 1/3 * 1/6 = 1/18.
- для 8d дайса: 1/3 * 1/8 = 1/24.

Нормалізуємо результати. Сумарна ймовірність $P(A)$ вибору одного з 3-х дайсів і випадання числа 2 складає:

$$1/12 + 1/18 + 1/24 = 13/72.$$

Це число менше за 1, бо ймовірність випадання числа 2 менша 1. Проте ми знаємо, що число 2 вже випало. Тому треба поділити ймовірність для кожного дайса на 13/72, щоб сума усіх ймовірностей для усіх дайсів була рівна 1. Цей процес відомий як *нормалізація*.

Після нормалізації знаходимо ймовірність $P(B_i|A)$ того, що дайс є одним з вибраних:

4d дайс: (1/12) / (13/72) = (1*72) / (12*13) = 6/13

6d дайс: (1/18) / (13/72) = (1*72) / (18*13) = 4/13

8d дайс: (1/24) / (13/72) = (1*72) / (24*13) = 3/13

Отже, на початку розв'язання цієї задачі ми припустили, що ймовірність вибору кожного дайса дорівнює 33.3%. Після кочення одного з дайсів і випадання числа 2 отримали результат: ймовірність вибору 4d дайса - 46.1%, 6d дайса – 30.8%, 8d дайса – 23.1%.

Результати занесено у *таблицю:*

Дайс	Початкова ймов., $P(B_i)$	Ймовірність випадання числа 2, $P(A\|B_i)$	$P(B_i)*P(A\|B_i)$	$P(B_i\|A)$
4d	1/3	1/4	1/12	6/13
6d	1/3	1/6	1/18	4/13
8d	1/3	1/8	1/24	3/13
Сумарна ймовірність, $P(A)$			13/72	

Завдання.

1. Використовуючи описаний вище приклад із 4d-, 6d- і 8d-дайсами, складіть імовірнісну таблицю для випадання числа 5 (замість 2). Порівняйте отримані результати із прикладом для числа 2.
2. Прорахуйте варіант випадання чисел 2-5-4 у процесі трьох кочень випадково вибраного дайсу. Складіть імовірнісну таблицю, врахувавши те, що, оскільки спершу випало число 2, то в колонці $P(B_i)$ нової таблиці для числа 5 стоятимуть числа з колонки $P(B_i|A)$ таблиці для числа 2.
3. Порівняйте отримані у п.1 і п.2 результати, зробіть висновки.

4. Розгляньте випадок 6-ти дайсів: 4d, 6d, 8d, 10d, 12d та 20d. Згенеруйте випадкові результати 10 кочень випадково вибраного дайсу і знайдіть ймовірності вибору кожного конкретного дайсу. (Для обчислень можна використати Excel, щоб створити у цій програмі ймовірнісну таблицю).

5. Результати обчислень занесіть у таблицю:

		Дайс →	4d	6d	8d	
Номер кидання ↓	Число, що спостерігаємо ↓	$P(B_i)$ →	1/6	1/6	1/6	
1		$P(B_1\|A)$ →				
10						

6. Проаналізуйте отримані результати. Зобразіть графічно залежності ймовірності вибору дайса $P(B_i|A)$ від порядкового номера кочення (1÷10).
7. Зробіть висновки. Оформіть звіт.