

Міністерство освіти і науки України
Національний авіаційний університет
Навчально-науковий інститут комп'ютерних інформаційних технологій
Кафедра комп'ютеризованих систем управління

Курсова робота
з дисципліни «Системне програмне забезпечення»

Пояснювальна записка
Тема: реалізація наївного баєсового класифікатора
на мові програмування Python

Виконав:
студент групи СП-325
Клокун В. Д.

Київ — 2018

**Завдання на виконання курсової роботи
студента групи СП-325 Клокуна Владислава Денисовича**

1. Тема курсової роботи: реалізація наївного баєсового класифікатора на мові програмування Python для класифікації спостережень, що містять неперервні дані.
2. Термін виконання курсової роботи:
з « ____ » _____ 2018 р. по « ____ » _____ 2018 р.
3. Вхідні дані до роботи: набір даних для класифікації.
4. Етапи виконання курсової роботи:
 - Огляд теоретичних відомостей про наївний баєсов класифікатор.
 - Реалізація та тестування наївного баєсового класифікатора.
5. Перелік обов'язкових додатків і графічного матеріалу:
 - FIXME.

Завдання отримав: « ____ » _____ 2018 р.

Підпис студента: _____ (Клокун В. Д.)

Зміст

1. Теоретична частина	4
1.1. Короткі теоретичні відомості	4
1.2. Імовірнісна модель наївного баєсового класифікатора	6
1.3. Оцінка параметрів	7
1.4. Побудова класифікатора з імовірнісної моделі	8
2. Практична частина	10
2.1. Використані програмні засоби	10
2.2. Опис програми та її роботи	10
Висновки	12

1. Теоретична частина

1.1. Короткі теоретичні відомості

Припустимо, що в ході деякого експерименту проводились спостереження, під час проведення яких збирались неперервні (недискретні) дані про результат події. Також були визначені категорії (або класи), до яких ці дані можуть належати. Поставлена задача класифікувати дані спостережень. *Класифікація* — це задача визначення, до якої з категорій належить певне спостереження [1]. *Класифікатор* — це алгоритм, який виконує класифікацію [1].

Наївний баєсів класифікатор — це ймовірнісний класифікатор, який використовує теорему Баєса для класифікації спостережень. Такі класифікатори отримують на вхід спостереження, оцінюють його і роблять припущення про клас, до якого воно належить. Вхідні дані, тобто спостереження, представляються у вигляді вектора відомих значень випадкових змінних, які називаються *ознаками*. Результатом роботи класифікатора є певне значення цільової змінної або змінних, які зазвичай називаються класовими, і позначають клас, до якого належить спостереження.

Принцип класифікації полягає в обчисленні умовних імовірностей (визначення 1) того, що вхідні дані належать до певних класів (події, які нас цікавлять), за умови, що ознаки мають певні значення (події, які ми спостерігаємо). Після обчислення кожної з умовних імовірностей за обраним правилом прийняття рішення робиться висновок, до якого класу належить задане спостереження. Оскільки такий класифікатор використовує ймовірнісну модель, наївний баєсів класифікатор називають *ймовірнісним*.

Визначення 1 (Умовна ймовірність). Нехай A і B — події. Позначимо ймовірність настання кожної з них незалежно одна від одної як $P(A)$ і $P(B)$ відповідно. Тоді *умовною ймовірністю* $P(A | B)$ називається ймовірність настання події A за умови, що подія B настала. Вона обчислюється так:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}, \quad (1.1)$$

де $P(A \cap B)$ — ймовірність, що події A і B настали.

Розглянемо приклад класифікації наївним баєсовим класифікатором. Нехай подія A — дане спостереження належить до певного класу, подія B — ознаки спостереження мають певні значення. Тоді щоб знайти ймовірність,

що дане спостереження з певним значенням ознак належить до певного класу, необхідно обчислити умовну ймовірність $P(A | B)$. Для обчислення цієї ймовірності необхідно використати теорему Баєса (теорема 1).

Теорема 1 (Баєса). Нехай $P(A | B)$ — умовна ймовірність настання події A за умови, що подія B настала, $P(B | A)$ — умовна ймовірність настання події B за умови, що подія A настала; $P(B)$ — ймовірність настання події B , причому $P(B) \neq 0$. Тоді умовна ймовірність $P(A | B)$ обчислюється так:

$$P(A | B) = \frac{P(B | A) P(A)}{P(B)}. \quad (1.2)$$

Доведення. Виразимо ймовірності $P(A | B)$ і $P(B | A)$ за визначенням умовної ймовірності:

$$P(A | B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0, \quad (1.3)$$

$$P(B | A) = \frac{P(B \cap A)}{P(A)}, \quad P(A) \neq 0. \quad (1.4)$$

Представимо ймовірності настання подій A і B разом, тобто $P(A \cap B)$ і $P(B \cap A)$ з рівностей 1.3, 1.4:

$$P(A \cap B) = P(A | B) P(B), \quad (1.5)$$

$$P(B \cap A) = P(B | A) P(A). \quad (1.6)$$

Оскільки ліві частини рівні за аксіомою $P(A \cap B) = P(B \cap A)$, то і праві частини також будуть рівними: $P(A | B) P(B) = P(B | A) P(A)$, отже можна записати, що:

$$P(A \cap B) = P(B | A) P(A). \quad (1.7)$$

Підставимо рівність 1.7 в 1.3 і отримаємо:

$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{P(B | A) P(A)}{P(B)}, \quad P(B) \neq 0. \quad (1.8)$$

□

Як видно, при класифікації наївним баєсовим класифікатором, у обчисленні умовних ймовірностей використовується теорема Баєса, тому такий ймовірнісний класифікатор називається *баєсовим*.

1.2. Імовірнісна модель наївного баєсового класифікатора

Як було сказано у підрозділі 1.1, наївний баєсовий класифікатор використовує ймовірнісну модель. Побудуємо та представимо її. Для виконання класифікації необхідні вхідні дані, тобто спостереження, та набір можливих значень класової змінної для позначення класів, до яких можуть належати ці дані. Позначатимемо змінні, на кшталт X_i , великими літерами, а їх значення, наприклад, x_i — малими. Вектори, на зразок \mathbf{X} — жирним шрифтом.

Вхідними даними для класифікації буде вектор ознак $\mathbf{X} = (X_1, \dots, X_n)$, де X_1, \dots, X_n — ознаки. Кожна ознака може мати значення зі своєї області визначення, яка позначається D_i . Набір усіх векторів ознак позначається як $\Omega = D_1 \times \dots \times D_n$. Для позначення класу, до якого належить спостереження, введемо випадкову змінну C , де C може приймати одне з m значень: $c \in \{0, \dots, m-1\}$.

Визначення 2 (Розподіл імовірностей). Нехай X і Y — випадкові змінні, які приймають значення x та y відповідно. Тоді розподіл імовірностей $p(X | Y)$ позначає значення імовірностей $P(X = x_i | Y = y_j)$ для кожної з можливих пар i, j . [2]

Класифікація за допомогою наївного баєсового класифікатора ставить у відповідність кожному вектору \mathbf{X} , який містить ознаки X_1, \dots, X_n , розподіли ймовірностей $p(C | \mathbf{X})$. Тобто сама модель має такий загальний вигляд:

$$p(C | \mathbf{X}) = p(C | X_1, \dots, X_n). \quad (1.9)$$

Зі зростанням кількості або можливих значень ознак, з такою моделлю неможливо працювати за допомогою таблиць імовірностей, тому переформулюємо модель, щоб зробити її зручнішою.

Використовуючи теорему Баєса, представимо її так:

$$p(C | X_1, \dots, X_n) = \frac{p(C) p(X_1, \dots, X_n | C)}{p(X_1, \dots, X_n)}. \quad (1.10)$$

Видно, що дільник не залежить від змінної C , а значення ознак X_i задані наперед, тому на практиці значення дільника постійне. Ділене рівносильне такий моделі спільного розподілу:

$$p(C, X_1, \dots, X_n). \quad (1.11)$$

Перетворюємо дану модель за допомогою визначення умовної ймовірності:

$$p(C, X_1, \dots, X_n) = p(C) p(X_1, \dots, X_n | C) \quad (1.12)$$

$$= p(C) p(X_1 | C) p(X_2, \dots, X_n | C, X_1) \quad (1.13)$$

$$= p(C) p(X_1 | C) p(X_2 | C, X_1) p(X_3, \dots, X_n | C, X_1, X_2) \quad (1.14)$$

$$= p(C) p(X_1 | C) p(X_2 | C, X_1) p(X_3 | C, X_1, X_2) \\ p(X_4, \dots, X_n | C, X_1, X_2, X_3) \quad (1.15)$$

і так далі. Тепер припускаємо, що кожна ознака X_i умовно незалежна від кожної іншої ознаки X_j , для будь-яких $j \neq i$ та заданої категорії $C = c$. Математично це означає:

$$p(X_i | C, X_j) = p(X_i | C).$$

Таке припущення є найвним, оскільки немає жодних підстав вважати, що входні ознаки дійсно незалежні одна від одної. Саме тому такий баєсовий класифікатор називається *найвним*.

Отже, виражаємо загальну модель:

$$p(C | X_1, \dots, X_n) = p(C) p(X_1 | C) p(X_2 | C) \dots p(X_n | C) \quad (1.16)$$

$$= p(C) \prod_{i=1}^n p(X_i | C). \quad (1.17)$$

Це означає, що враховуючи припущення про незалежність змінних, умовний розподіл над класовою змінною C може бути виражений так:

$$p(C | X_1, \dots, X_n) = \frac{1}{Z} p(C) \prod_{i=1}^n p(X_i | C), \quad (1.18)$$

де Z — коефіцієнт, який залежить виключно від X_1, \dots, X_n . Якщо значення ознак x_1, \dots, x_n відомі, коефіцієнт Z сталий.

Таку модель значно зручніше використовувати, оскільки вони використовують апіорні імовірності класів $p(C)$ та незалежні розподіли $p(X_i | C)$. Якщо є k класів та модель для $p(X_i)$ може бути виражена r параметрами, то відповідний найвний баєсовий класифікатор матиме $(k - 1) + nrk$ параметрів. [3]

1.3. Оцінка параметрів

Описавши ймовірнісну модель, необхідно визначити її параметри, тобто апіорні ймовірності належності до класу $p(C)$ та розподіли ймовірностей

ознак $p(X_i | C)$. Для обчислення параметрів моделі використовують *тренувальний набір даних* — такий набір даних, який складається із заздалегідь класифікованих спостережень. Тобто набір даних S складатиметься з векторів $\mathbf{x} = (x_1, \dots, x_n, c)$, де c — правильне значення класової змінної.

Усі параметри моделі можна обчислити з тренувального набору даних. [3] Щоб оцінити значення параметрів, необхідно зробити припущення щодо розподілу, який характеризує дані. Припущення щодо розподілу, який характеризує дані, називають *моделлю подій*. При роботі з неперервними (недискретними) даними, зазвичай припускають, що вони розподілені за законом нормального (гаусового) розподілу. Функція густини ймовірності нормального розподілу така:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \quad (1.19)$$

де μ — математичне сподівання, σ^2 — дисперсія випадкової величини.

Наприклад, припустимо, що тренувальний набір даних містить неперервну ознаку X . Щоб обчислити розподіл імовірності, необхідно спочатку розподілити дані за класами, наданими у тренувальному наборі, та обчислити математичне сподівання μ_c і дисперсію випадкової величини σ_c^2 для кожного з класів. Нехай μ_{c_1} — математичне сподівання значень, які належать до класу c_1 , а $\sigma_{c_1}^2$ — їх дисперсія. У результаті певного спостереження отримали значення x . Тоді розподіл імовірності для значення x для класу c обчислюється так:

$$p(X = x | C = c) = \frac{1}{\sigma_c\sqrt{2\pi}} \exp\left(-\frac{(x - \mu_c)^2}{2\sigma_c^2}\right). \quad (1.20)$$

Оцінивши усі необхідні параметри для моделі, тобто апіорні розподіли імовірності належності до класу $p(C)$ та розподіли ймовірностей ознак $p(X_i | C)$, можна переходити до класифікації.

1.4. Побудова класифікатора з імовірнісної моделі

Наївний баєсовий класифікатор поєднує баєсову імовірнісну модель з правилом прийняття рішення. Одним з поширених правил є вибір найбільш ймовірної гіпотези. Такий підхід називається *правилом прийняття рішення за максимальною апостеріорною імовірністю* (maximum a posterior (MAP) decision rule). Відповідний класифікатор є функцією $\text{classify}(X_1, \dots, X_n)$, яка ви-

значається так:

$$\text{classify}(X_1, \dots, X_n) = \arg \max_c \left(p(C = c) \prod_{i=1}^n p(X_i = x_i \mid C = c) \right), \quad (1.21)$$

де $\arg \max_x f(x)$ — функція, результатом якої є множина значень x , при якому значення функції $f(x)$ максимальне (визначення 3) [3].

Визначення 3 (Функція $\arg \max$). Нехай дана довільна множина X , повністю впорядкована множина Y та функція $f : X \mapsto Y$, тоді функція $\arg \max$ для певного значення x над певною підмножиною S визначається так:

$$\arg \max_{x \in S \subseteq X} f(x) := \{x \mid x \in S \wedge \forall y \in S : f(y) \leq f(x)\}. \quad (1.22)$$

Правило прийняття рішення за максимальною апостеріорною імовірністю правильно класифікує спостереження за умови, що ймовірність належності до правильного класу більша за ймовірності належності до інших класів, тому немає потреби у надточній оцінці цих імовірностей.

Таким чином ми отримали працюючу теоретичну модель для реалізації найвісного баєсового класифікатора для класифікації спостережень, які містять неперервні (недискретні) дані.

2. Практична частина

2.1. Використані програмні засоби

Для реалізації наївного баєсового класифікатора була використана мова програмування Python версії 3.7.1. Оскільки Python — інтерпретована мова програмування, для коректної роботи розробленої реалізації необхідно встановити робочий інтерпретатор Python 3, який можна завантажити на офіційному сайті за посиланням <https://python.org/downloads>. Також розроблена реалізація використовує засоби стандартної бібліотеки мови програмування Python (табл. 1).

Табл. 1: Перелік використаних модулів стандартної бібліотеки мови програмування Python

Модуль	Призначення
<code>argparse</code>	Обробка аргументів командного рядка.
<code>csv</code>	Зчитування файлів у форматі <code>.csv</code> , які містять дані, розділені комою.
<code>math</code>	Зручне обчислення математичних функцій, зокрема e^x , a^b та \sqrt{x} .
<code>random</code>	Робота з випадковими числами, а саме перемішування набору даних у випадковому порядку.

2.2. Опис програми та її роботи

Реалізація класифікатора виконана у вигляді програми на мові Python, початковий код якої знаходиться у двох файлах: `nbc.py` і `nbc_main.py`. Програма виконує класифікацію неперервних чисельних даних із набору даних, заданого вхідним файлом у форматі CSV. Запуск програми виконується за допомогою командного рядка:

```
1 python nbc_main.py input.csv
```

Після виконання вищезазначеної команди, програма зчитує набір даних, вказаний у файлі `input.csv`, оброблює та класифікує їх, виводить результати класифікації та її обчислену точність. Якщо вхідний файл не зазначений, програма попереджує про це користувача, надає інформацію про правильний формат використання та завершує роботу.

Розглянемо процес роботи програми детальніше. Почнемо з вмісту файлу `nbc_main.py` (лістинг 2.1).

Лістинг 2.1: `nbc_main.py`

```
1  #!/usr/bin/env python3
2
3  import argparse
4  from nbc import *
5
6  def main(args):
7      dataset = load_training(args.input)
8      dataset_train, dataset_test = split_dataset(dataset, 3) # split
9          ↪ dataset 1 / 3
10
11     summaries = summarize_by_class(dataset_test)
12     predictions = predict_dataset(summaries, dataset_test)
13     acc = compute_accuracy(dataset_test, predictions)
14     print('Accuracy: {}'.format(acc))
15
16 if __name__ == '__main__':
17     parser = argparse.ArgumentParser(description = 'An
18         ↪ implementation of Gaussian Naive Bayes Classifier in Python
19         ↪ using stdlib facilities. Reads a CSV file containing
20         ↪ floats.')
21
22     # Parse CSV dataset file
23     parser.add_argument('input', help = 'Path to input file')
24
25     args = parser.parse_args()
26
27     main(args)
```

На початку виконання програми виконується перевірка, як саме запускається модуль: якщо модуль є основною програмою, то інтерпретатор встановлює змінній `__name__` значення `'__main__'`, умова `if __name__ == 'main'` виконується, і починається обробка аргументів програми.

Висновки

Література

1. Statistical classification. — URL: https://en.wikipedia.org/wiki/Statistical_classification (дата зверн. 20.11.2018).
2. *Stuart Russell, Peter Norvig*. Artificial Intelligence: A Modern Approach. — 3-е вид. — Prentice Hall, 2010. — (Prentice Hall Series in Artificial Intelligence). — ISBN 9780136042594.
3. *Prof. M. Narasimha Murty, Dr. V. Susheela Devi (auth.)* Pattern Recognition: An Algorithmic Approach. — 1-е вид. — Springer-Verlag London, 2011. — (Undergraduate Topics in Computer Science 0). — ISBN 978-0-85729-494-4.