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

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 *імовірнісним*.

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 імовірністю* $P(A \mid B)$ називається ймовірність настання події A за умови, що подія B настала. Вона обчислюється так:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}, \tag{1.1}$$

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 $P(A \mid B)$. Для обчислення цієї імовірності необхідно використати теорему Баєса (теорема 1).

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

$$P(A \mid B) = \frac{P(B \mid A)\,P(A)}{P(B)}. \tag{1.2}$$

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

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

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

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

$$P(A \cap B) = P(A \mid B)\,P(B), \tag{1.5}$$

$$P(B \cap A) = P(B \mid A)\,P(A). \tag{1.6}$$

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

$$P(A \cap B) = P(B \mid A)\,P(A). \tag{1.7}$$

Підставимо рівність 1.7 в 1.3 і отримаємо:

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

□

Як видно, при класифікації наївним баєсовим класифікатором, у обчисленні умовних імовірностей використовується теорема Баєса, тому такий імовірнісний 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 $\Omega = D_1 \times \cdots \times D_n$. 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 $p(X \mid Y)$ позначає значення імовірностей $P(X = x_i \mid Y = y_i)$ 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 $p(C \mid \mathbf{X})$. Тобто сама модель має такий загальний вигляд:

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

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:

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

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

$$p(C, X_1, \dots, X_n). \tag{1.11}$$

Перетворюємо дану модель за допомогою визначення умовної ймовірності:

$$p(C, X_1, \dots, X_n) = p(C)\, p(X_1, \dots, X_n \mid C) \tag{1.12}$$
$$= p(C)\, p(X_1 \mid C)\, p(X_2, \dots, X_n \mid C, X_1) \tag{1.13}$$
$$= p(C)\, p(X_1 \mid C)\, p(X_2 \mid C, X_1)\, p(X_3, \dots, X_n \mid C, X_1, X_2) \tag{1.14}$$
$$= p(C)\, p(X_1 \mid C)\, p(X_2 \mid C, X_1)\, p(X_3 \mid C, X_1, X_2)$$
$$p(X_4, \dots, X_n \mid C, X_1, X_2, X_3) \tag{1.15}$$

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:

$$p(X_i \mid C, X_j) = p(X_i \mid C).$$

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

Отже, виражаємо загальну модель:

$$p(C \mid X_1, \dots, X_n) = p(C)\, p(X_1 \mid C)\, p(X_2 \mid C) \dots p(X_n \mid C) \tag{1.16}$$
$$= p(C) \prod_{i=1}^{n} p(X_i \mid C). \tag{1.17}$$

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

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

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 апріорні імовірності класів $p(C)$ та незалежні розподіли $p(X_i \mid C)$. Якщо є k класів та модель модель для $p(X_i)$ може бути виражена r параметрами, то відповідний наївний баєсовий класифікатор матиме $(k-1) + nrk$ параметрів. [3]

1.3. Оцінка параметрів

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

ознак $p(X_i \mid C)$. 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:

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

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:

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

Оцінивши усі необхідні параметри для моделі, тобто апріорні розподіли імовірності належності до класу $p(C)$ та розподіли ймовірностей ознак $p(X_i \mid C)$, 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 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:

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

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

2.2. Опис програми та роботи з нею

Реалізація класифікатора (яку далі ми називатимемо програмою) виконана у вигляді двох модулів на мові Python: `nbc.py` і `nbc_main.py`. Модуль `nbc.py` містить реалізації всіх функцій, необхідних для роботи наївного баєсового класифікатора, а модуль `nbc_main.py` містить інструкції, які використовують надані функції для класифікації наборів даних.

Розроблена програма виконує класифікацію неперервних чисельних даних із набору даних, який знаходиться у вхідному файлі формату CSV. Вона запускається за допомогою командного рядка таким чином:

```
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 результа-

ти класифікації та її обчислену точність. Якщо вхідний файл не зазначений, програма попереджує про це користувача, надає інформацію про правильний формат використання та завершує роботу.

2.3. Опис процесу роботи програми

Модуль `nbc_main.py`, який використовується для запуску програми, містить інструкції для класифікації конкретного набору даних наївним баєсовим класифікатором. Він завантажує необхідні для роботи програми ресурси в область видимості (лістинг 2.1) та описує загальний процес роботи програми.

Лістинг 2.1: Модуль `nbc_main.py`: завантаження ресурсів, необхідних для роботи класифікатора

```
1 #!/usr/bin/env python3
2
3 import argparse
4 from nbc import *
```

Перш за все, у файлі модуля вказаний шлях до сумісного інтерпретатора, а саме Python 3. Це надає можливість запускати модуль у Unix-подібних операційних системах як скрипт, не вказуючи шлях до потрібного інтерпретатора явно.

Інструкція **`import argparse`** завантажує в область видимості модуль для обробки аргументів командного рядка. Інструкція **`from nbc import *`** завантажує всі функції, описані у модулі `nbc_main.py`, щоб їх можна було використовувати у поточному модулі.

Тепер, коли обробник параметрів командного рядка та всі функції класифікатора завантажені, відбувається перевірка способу запуску модуля (лістинг 2.2).

Лістинг 2.2: Модуль `nbc_main.py`: перевірка способу запуску модуля

```
1 if __name__ == '__main__':
2 parser = argparse.ArgumentParser(description = 'An implementation of
   ↪ Gaussian Naive Bayes Classifier in Python using stdlib facilities.
   ↪ Reads a CSV file containing floats.')
3
4 # Parse CSV dataset file
```

```
5 parser.add_argument('input', help = 'Path to input file')
6
7 args = parser.parse_args()
8
9 main(args)
```

Якщо модуль запущений як скрипт, наприклад, з командного рядка, як необхідно для роботи розробленої реалізації, інтерпретатор встановлює особливе значення змінної `__name__`. Тому перевіряється умова `__name__ == '__main__'`. Якщо вона виконується, ініціалізується обробник параметрів командного рядка `argparse.ArgumentParser` з коротким описом програми. Далі додається аргумент `'input'` зі стислим поясненням до нього: він містить шлях до вхідного файлу, тобто файлу з набором даних, які необхідно класифікувати.

Коли у обробнику описані бажані аргументи, створюється змінна `args`, функція `parser.parse_args()` оброблює параметри та зберігає їх значення у створеній змінній. Тепер, коли параметри оброблені, вони передаються у функцію `main()` — починаються основні етапи роботи програми (лістинг 2.3).

Лістинг 2.3: Модуль `nbc_main.py`: функція `main()`, яка описує загальний процес роботи програми

```
1 def main(args):
2     dataset = load_training(args.input)
3     dataset_train, dataset_test = split_dataset(dataset, 1/3) # split
      ↪  dataset 1 / 3
4
5     summaries = summarize_by_class(dataset_test)
6     predictions = predict_dataset(summaries, dataset_test)
7     acc = compute_accuracy(dataset_test, predictions)
8     print('Accuracy: {}'.format(acc))
```

Загалом, процес роботи програми можна умовно поділити на такі етапи:

1. Підготовка вхідних даних.
2. Класифікація.
3. Виведення результату.

Розглянемо кожен етап роботи програми по порядку.

2.3.1. Підготовка вхідних даних

Функція `main()` (лістинг 2.3), а з нею і етап підготовки вхідних даних, починається з інструкції `dataset = load_training(args.input)`, яка призначена для завантаження файлу з набором даних у змінну `dataset` за допомогою функції `load_training()` (лістинг 2.4). Розглянемо роботу цієї функції детальніше.

Лістинг 2.4: Модуль `nbc.py`: функція `load_training()` для завантаження набору даних у пам'ять

```
1  def load_training(filename):
2          with open(filename) as file:
3                  # ignore all lines which start with '#'
4                  reader = csv.reader(row for row in file if not
                    ↪  row.startswith('#'))
5                  dataset = []
6                  for row in reader:
7                          # comprehend each line as a list of floats and
                            ↪  append it to the dataset
8                          dataset.append([float(x) for x in row])
9
10         return dataset
```

Функція `load_training()` відкриває файл, назва якого передається в аргументі `filename`, ініціалізує CSV-зчитувач `csv.reader` у змінну `reader`. Зчитувач ініціалізується всіма рядками файлу, які не починаються з символу «#». Далі рядки, записані у зчитувач, перетворюються у вектори, які містять ознаки у форматі з плаваючою комою. Перетворені вектори записуються у змінну `dataset`. Тепер функція повертає змінну `dataset` та завершує роботу.

Далі у процесі виконання програми виконується інструкція `dataset_train, dataset_test = split_dataset(dataset, 1/3)`. Вона ділить завантажений набір даних `dataset` на тренувальний набір даних `dataset_train` і тестовий набір даних `dataset_test` за допомогою функції `split_dataset` (лістинг 2.5). Розподіл відбувається таким чином, щоб тренувальний набір даних `dataset_train` містив як мінімум `ratio` частину загального набору даних. Розбиття загального набору даних необхідне для тестування точності роботи класифікатора і є загальноприйнятою практикою в аналізі даних.

Лістинг 2.5: Модуль `nbc.py`: функція `split_dataset()` для розбиття загального набору даних на тренувальний та тестовий набори

```
1 def split_dataset(dataset, ratio):
2         dataset_shuffled = dataset
3         random.shuffle(dataset_shuffled)
4         dataset_train = dataset_shuffled[:int(len(dataset_shuffled) *
          ↪  ratio)]
5         dataset_test = dataset_shuffled[int(len(dataset_shuffled) *
          ↪  ratio):]
6
7         return dataset_train, dataset_test
```

Функція `split_dataset()` копіює отриманий набір даних `dataset` у змінну `dataset_shuffled`. Далі функція `random.shuffle()` перемішує вміст змінної `dataset_shuffled`, тобто початковий набір даних, випадковим чином, щоб усунути будь-які статистичні похибки (упередження) у даних. Перемішування відбувається безпосередньо у змінній, на місці.

Тепер дані діляться на необхідні частини: зі списку `dataset_shuffled` копіюється частина від початку до елементу з індексом `int(len(dataset_shuffled) * ratio)` і присвоюється змінній `dataset_train`. Далі з цього ж списку копіюється частина від елементу з індексом `int(len(dataset_shuffled) * ratio)` до кінця і присвоюється змінній `dataset_test` (рис. 1).

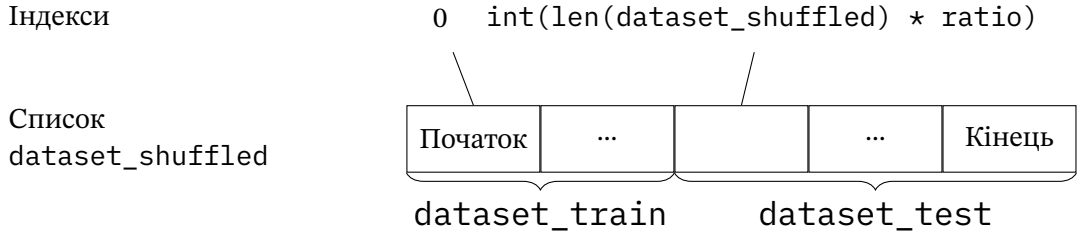

Рис. 1: Логіка розділення списку `dataset_shuffled`

Після виконання розподілу функція повертає кортеж зі змінних `dataset_train` та `dataset_test` і завершує роботу. На цьому підготовка вхідних даних завершена.

2.3.2. Класифікація

Після розбиття набору даних на тренувальний та тестовий, виконується інструкція `summaries = summarize_by_class(dataset_test)`, яка оцінює статистичні параметри значень ознак для кожного значення класової змінної за допомогою функції `summarize_by_class()` (лістинг 2.6).

Лістинг 2.6: Модуль `nbc.py`: функція `summarize_by_class()` для оцінки статистичних параметрів можливих значень кожної ознаки для усіх класів, до яких може належати спостереження

```
1 def summarize_by_class(dataset):
2     separated = separate_by_class(dataset)
3     summaries = {}
4     for class_value, data in separated.items():
5         summaries[class_value] = summarize(data)
6
7     return summaries
```

2.3.3. Виведення результату

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.