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 кожного з класів (події, які нас цікавлять), за умови, що ознаки мають певні значення (події, які ми спостерігаємо). Після обчислення кожної з умовних імовірностей знаходиться найбільша та робиться висновок про належність спостереження до кожного з класів. Саме тому 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 імовірністю $P(A \mid B)$ *називається ймовірність настання події A за умови, що подія B настала.*

Наприклад, 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 і $P(B)$ — імовірність настання події B, причому $P(B) \neq 0$. Тоді умовна ймовірність $P(A \mid B)$ обчислюється так:

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

Оскільки при обчисленні умовних імовірностей використовується теорема Баєса, такий імовірнісний класифікатор називається баєсовим.

1.2. Імовірнісна модель наївного баєсового класифікатора

Як було сказано у підрозділі 1.1, для виконання класифікації необхідні такі дані: вхідні дані, тобто спостереження, та можливі значення 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:

$$p\left(C \mid \mathbf{X}\right) = p\left(C \mid X_1, \dots, X_n\right). \tag{1}$$

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:

$$p(C \mid X_1, \dots, X_n) = \frac{p(C)\, p(X_1, \dots, X_n \mid C)}{p(X_1, \dots, X_n)}. \tag{2}$$

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:

$$p(C, X_1, \dots, X_n). \tag{3}$$

Перетворюємо дану модель за допомогою визначення умовної ймовірності:

$$\begin{aligned} p(C, X_1, \dots, X_n) &= p(C)\, p(X_1, \dots, X_n \mid C) \\ &= p(C)\, p(X_1 \mid C)\, p(X_2, \dots, X_n \mid C, X_1) \\ &= p(C)\, p(X_1 \mid C)\, p(X_2 \mid C, X_1)\, p(X_3, \dots, X_n \mid C, X_1, X_2) \\ &= p(C)\, p(X_1 \mid C)\, p(X_2 \mid C, X_1)\, p(X_3 \mid C, X_1, X_2)\, p(X_4, \dots, X_n \mid C, X_1, X_2, X_3) \end{aligned}$$

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:

$$p(C \mid X_1, \dots, X_n) = p(C)\, p(X_1 \mid C)\, p(X_2 \mid C) \dots p(X_n \mid C) \tag{4}$$

$$= p(C) \prod_{i=1}^{n} p(X_i \mid C). \tag{5}$$

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

$$p(C \mid X_1, \dots, X_n) = \frac{1}{Z}\, p(C) \prod_{i=1}^{n} p(X_i \mid C), \tag{6}$$

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. Нормальний розподіл характеризується двома величинами: математичним сподіванням μ та дисперсією випадкової величини σ^2.

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.

Нехай необхідно обчислити ймовірність, що

2. Практична частина

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.