Практикум 1. Отримання навичок роботи в середовищі Python

Недашківська Н.І.

1 Варіанти завдань

Варіанти завдань вибирати відповідно до номеру в списку групи.

При виконанні завдань використовувати універсальні функції, функції транслювання (broadcasting) та агрегування бібліотеки NumPy.

1. Дано вектор y розмірності N, який відповідає деякій множині з N навчальних прикладів. Елементи вектору y приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$. Знайти значення ентропії

$$H(S) = -\sum_{i=1}^{v} \frac{k_i}{N} \log_2 \frac{k_i}{N},$$

де властивість S може приймати v різних значень, кожне з яких - в k_i випадках.

2. Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення з множини $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти ознаку x_h^*, для якої наступний вираз приймає мінімальне значення:

$$G(x_h) = \sum_{i=1}^{q_h} \frac{|T_i|}{N} H(T_i, S),$$

де T_i - підмножина прикладів, для яких ознака x_h приймає значення c_{hi}, $|A|$ - потужність множини A, $H(A, S)$ - ентропія множини A по відношенню до властивості S:

$$H(A, S) = -\sum_{i=1}^{v} \frac{k_i}{|A|} \log_2 \frac{k_i}{|A|},$$

де властивість S може приймати v різних значень, кожне з яких - в k_i випадках.

3. Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення з множини $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти ознаку x_h^*, для якої наступний вираз приймає мінімальне значення:

$$G(x_h) = \sum_{i=1}^{q_h} \frac{|T_i|}{N} H(T_i, S),$$

де T_i - підмножина прикладів, для яких ознака x_h приймає значення c_{hi}, $|A|$ - потужність множини A, $H(A, S)$ - індекс Джині множини A по відношенню до властивості S:

$$H(A, S) = 1 - \sum_{i=1}^{v} \left(\frac{k_i}{|A|} \right)^2,$$

де властивість S може приймати v різних значень, кожне з яких - в k_i випадках.

4. Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення з множини $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти ознаку x_h^* та значення цієї ознаки c_{hi}^*:

$$\mathrm{c}_{hi}^* = \arg\max_{h,i} \frac{p_2(y = s_j | x_h = c_{hi})}{p_1(x_h = c_{hi})},$$

де s_j - задано, $p_1(x_h = c_{hi})$ - кількість прикладів, для яких ознака x_h приймає значення c_{hi}, $p_2(y = s_j | x_h = c_{hi})$ - кількість прикладів, які належать класу s_j і ознака x_h приймає значення c_{hi}.

5. Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення з множини $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти ознаку x_h^* та значення цієї ознаки c_{hi}^*:

$$\mathrm{c}_{hi}^* = \arg\min_{h,i} Er(h, i),$$

$$Er(h, i) = \frac{p_3(y \neq s_j^* | x_h = c_{hi})}{p_1(x_h = c_{hi})},$$

$$s_j^* = \arg\max_{j} p_2(y = s_j | x_h = c_{hi}),$$

де $p_1(x_h = c_{hi})$ - кількість прикладів, для яких ознака x_h приймає значення c_{hi}, $p_2(y = s_j | x_h = c_{hi})$ - кількість прикладів, які належать класу s_j

і ознака x_h приймає значення c_{hi}, s_j^* - найбільш імовірний клас за умови що ознака x_h приймає значення c_{hi}.

6. Дано масив T, який складається з N рядків, які відповідають прикладам, і m стовпчиків, які відповідають ознакам. Відомо, що ознака x_h приймає значення $\{c_{h1}, c_{h2}, ..., c_{hq_h}\}$. Дано вектор y розмірності N, елементи якого приймають значення з множини $S = \{s_1, s_2, ..., s_v\}$ (мітки класів для прикладів). Знайти значення s_k^* (найбільш імовірний клас) для нового прикладу, який характеризується заданими значеннями ознак $x_1 = a_1$, $x_2 = a_2$, ..., $x_m = a_m$:

$$s_k^* = \arg\max_{s_k \in S} p(y = s_k) \prod_{i=1}^{N} p(x_i = a_i | y = s_k),$$

де a_i - задані, $p(y = s_k)$ - кількість прикладів, які належать класу s_k, $p(x_i = a_i | y = s_k)$ - кількість прикладів, у яких ознака x_i приймає значення a_i, серед тих, що належать класу s_k.

7. Дано масив $T = \{(t_i) | t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$, $x_{ik} \in R$, де приклад t_i характеризується m ознаками. Для цих даних розрахувати матриці відстаней: евклідової D_2, хемінга D_H і чебишева D_∞:

$$D_2(t_p, t_q) = \sqrt{\sum_{k=1}^{m} (x_{pk} - x_{qk})^2}$$

$$D_H(t_p, t_q) = \sum_{k=1}^{m} |x_{pk} - x_{qk}|$$

$$D_\infty(t_p, t_q) = \max_{k=1,...,m} |x_{pk} - x_{qk}|$$

8. Дано масив $T = \{(t_i) | t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$, $x_{ik} \in R$, де приклад t_i характеризується m ознаками. Для цих даних розрахувати матриці відстаней: пікову D_P та махаланобіса D_M:

$$D_P(t_p, t_q) = \frac{1}{m} \sum_{k=1}^{m} \frac{|x_{pk} - x_{qk}|}{x_{pk} + x_{qk}}$$

$$D_M(t_p, t_q) = \sqrt{(x_p - x_q)^T S^{-1} (x_p - x_q)},$$

де S - матриця коваріації.

9. Дано масив $T = \{(t_i) | t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$, $x_{ik} \in R$, де приклад t_i характеризується m ознаками. Об'єднати приклади в кластери за наступним алгоритмом:

1) $C := T$, множина кластерів C співпадає з початковою множиною прикладів,

2) Поки в C більше одного елементу:

- вибираємо два кластери $c_p, c_q \in C$, відстань між якими мінімальна,
- об'єднуємо c_p і c_q у новий кластер c_{pq}, змінюємо C за правилом:

  $$C := C \cup c_{pq} \setminus \{c_p, c_q\},$$

  Відстань між кластерами:

  $$d_{rs} = \frac{d_{ps} + d_{qs}}{2},$$

  де d_{rs} - відстань від нового кластера c_r, який утворено об'єднанням c_p і c_q, до іншого кластера c_s.

Надрукувати множину кластерів C і матрицю відстаней між отриманими кластерами.

10. Розглянути умову попередньої задачі. Надрукувати множину кластерів C і матрицю відстаней між отриманими кластерами, якщо відстань між кластерами розраховується за формулою:

$$d_{rs} = \frac{d_{ps} + d_{qs}}{2} - \frac{|d_{ps} - d_{qs}|}{2},$$

де d_{rs} - відстань від нового кластера c_r, який утворено об'єднанням c_p і c_q, до іншого кластера c_s.

11. Дано масив $T = \{(t_i)|t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$, $x_{ij} \in R$, де приклад t_i характеризується m ознаками. Задано кількість кластерів $2 \leq g \leq N$. Розрахувати центри кластерів за формулою:

$$c_k = \frac{\sum_{i=1}^{N} u_{ki} t_i}{\sum_{i=1}^{N} u_{ki}}, k = 1, ..., g,$$

де $U = \{(u_{ki})|k = 1, ..., g, i = 1, ..., N\}$ - випадковим чином задана матриця початкового розбиття, $u_{ki} \in \{0, 1\}$, $\sum_{k=1}^{g} u_{ki} = 1$, $\sum_{i=1}^{N} u_{ki} < N$.

Перерахувати матрицю розбиття:

$u_{ki} = 1$ якщо $d(t_i, c_k) = \min_{l=1,...,g} d(t_i, c_l)$,

$u_{ki} = 0$ в іншому випадку,

за умови, що $d(t_i, c_k)$ - евклідова відстань між векторами.

Виконати декілька ітерацій з уточнення центрів кластерів.

12. Задано неорієнтовний граф G з V вершинами, де ваги дуг d_{ij} відомі для $\forall i, j = 1, ..., V$ і позначають відстані між об'єктами. Задано поріг близькості $\sigma \in [\min d_{ij}, \max d_{ij}]$. Знайти множину кластерів на основі графу G, використовуючи наступні кроки:

1) Вилучити з графа ребра, ваги яких перевищують заданий поріг близькості σ.

2) Компонента зв'язності графу – підмножина вершин графу, в якій будь-які вершини можна поєднати шляхом, який цілком належить цій підмножині.

Знайти компоненти зв'язності отриманого графа, вони і будуть шуканими кластерами.

13. Покриваючим або остовним деревом графу називається зв'язний підграф без циклів, який містить всі вершини графу. Перевірити, чи є заданий неорієнтований граф покриваючим деревом.

14. Задано неорієнтовний граф G з V вершинами, де ваги дуг d_{ij} відомі для $\forall i, j = 1, ..., V$. Побудувати підграф J графу G, використовуючи наступні кроки:

    1) Відсортувати ребра в порядку зростання їх ваг. $J := \varnothing$.

    2) Додати ребро до J, якщо воно не утворює цикл з наявними ребрами.

    3) Виконувати крок 2 до тих пір поки до J не буде додано $V - 1$ ребро.

15. Задано неорієнтовний граф G з V вершинами, де ваги дуг d_{ij} відомі для $\forall i, j = 1, ..., V$. Побудувати підграф J графу G, використовуючи наступні кроки:

    1) Вибрати будь-яку вершину графу G і додати її до J.

    2) Додати до J ребро з найменшою вагою, яке з'єднує вершину підграфу J з вершиною, яка не належить J.

    3) Виконувати крок 2 до тих пір поки до J не буде додано $V - 1$ ребро.

16. Розглянути критерій якості кластеризації - коефіцієнт розбиття:

$$PC = \frac{\sum_{j=1}^{N} \sum_{k=1}^{g} u_{kj}^2}{N},$$

    де N - задана кількість об'єктів, які кластеризуються, $1 \le g \le N$ - задана кількість кластерів, $U = \{(u_{kj}) | k = 1, ..., g, j = 1, ..., N\}$ - матриця розбиття, $u_{kj} \in \{0, 1\}$, причому $u_{kj} = 1$ означає приналежність j-го об'єкту k-му кластеру, $\sum_{k=1}^{g} u_{kj} = 1$, $\sum_{j=1}^{N} u_{kj} < N$.

    Використовуючи результати моделювання великої кількості матриць розбиття, показати, що

$$PC \in \left[\frac{1}{g}, 1\right].$$

17. Розглянути критерій якості кластеризації - ентропію розбиття:

$$PE = -\frac{\sum_{j=1}^{N} \sum_{k=1}^{g} u_{kj} \ln u_{kj}}{N},$$

де N - задана кількість об'єктів, які кластеризуються, $1 \leq g \leq N$ - задана кількість кластерів, $U = \{(u_{kj})|k = 1, ..., g, j = 1, ..., N\}$ - матриця розбиття, $u_{kj} \in \{0, 1\}$, причому $u_{kj} = 1$ означає приналежність j-го об'єкту k-му кластеру, $\sum_{k=1}^{g} u_{kj} = 1$, $\sum_{j=1}^{N} u_{kj} < N$.

Використовуючи результати моделювання великої кількості матриць розбиття, показати, що

$$PE \in [0, \ln g].$$

18. Згенерувати N об'єктів в R^2 так, щоб вони утворювали віддалені один від одного скупчення, $1 \leq g^* \leq N$ - задана кількість кластерів. В процесі генерування задати $U^* = \{(u_{kj})|k = 1, ..., g^*, j = 1, ..., N\}$ - матрицю розбиття, вона показує до якого кластеру відноситься кожний з об'єктів, $u_{kj} \in \{0, 1\}$, причому $u_{kj} = 1$ означає приналежність j-го об'єкту k-му кластеру, $\sum_{k=1}^{g^*} u_{kj} = 1$, $\sum_{j=1}^{N} u_{kj} < N$.

    Розглянути декілька результатів кластеризації цих об'єктів, які задаються матрицями розбиття:

    - еталонна кластеризація, яка задається U^* і відповідає початковим правилам генерування об'єктів,

    - зашумлені кластеризації, в яких окремі об'єкти віднесені до інших кластерів. Розглянути також випадки коли кількість кластерів g не співпадає з початково згенерованою g^*.

    Показати, що на найкращому розбитті U^* індекс чіткості CI приймає найбільше значення:

$$CI = \frac{gPC - 1}{g - 1},$$

$$PC = \frac{\sum_{j=1}^{N} \sum_{k=1}^{g} u_{kj}^2}{N}.$$

19. Розглянути умову попереднього варіанту. Дослідити, яке значення приймає модифікована ентропія розбиття PE_M на найкращому розбитті U^* :

$$PE_M = \frac{PE}{\ln g},$$

$$PE = -\frac{\sum_{j=1}^{N} \sum_{k=1}^{g} u_{kj} \ln u_{kj}}{N}.$$

20. Розрахувати індекс ефективності кластеризації:

$$PI = \sum_{j=1}^{N} \sum_{k=1}^{g} u_{kj}^2 (d^2(\bar{t}, c_k) - d^2(t_j, c_k)),$$

    - $T = \{(t_i)|t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$ - множина об'єктів, які кластеризуються, $x_{ik} \in R$,

- $\bar{t}$ - вибіркове середнє об'єктів $t_i \in T$,
- $2 \le g \le N$ - задана кількість кластерів,
- $U = \{(u_{kj})|k = 1, ..., g, j = 1, ..., N\}$ - задана матриця розбиття, $u_{kj} \in \{0, 1\}$, причому $u_{kj} = 1$ означає приналежність j-го об'єкту k-му кластеру, $\sum_{k=1}^{g} u_{kj} = 1$, $\sum_{j=1}^{N} u_{kj} < N$,
- $\{c_k|k = 1, ..., g\}$ - задані центри кластерів,
- $d^2(t_j, c_k)$ - квадрат евклідової відстані між векторами.

21. Дано масив $T = \{(t_i)|t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$ об'єктів, які потрібно кластеризувати, $x_{ik} \in R$. Задано параметр $\rho > 0$. В якості міри близькості вибрано евклідову відстань $d(t_i, t_j)$. Знайти множину кластерів за наступними етапами:

    1) Ініціалізувати множину некластеризованих точок $U := T$.

    2) Поки є некластеризовані точки, тобто $U \neq \varnothing$:

    - випадковим чином вибрати $t_0 \in U$,
    - повторювати:
        - утворити кластер – сферу з центром t_0 і радіусом ρ:

        $$C_0 := \{t_i \in T | d(t_i, t_0) \le \rho\},$$

        - помістити центр сфери в центр мас кластера:

        $$t_0 := \frac{1}{|C_0|} \sum_{t_i \in C_0} t_i,$$

    - поки центр t_0 не стабілізується,
    - відмітити всі точки множини C_0 як кластеризовані: $U := U \setminus C_0$.

22. Дано масив $T = \{(t_i)|t_i = (x_{i1}, x_{i2}, ..., x_{im}), i = 1, ..., N\}$, $x_{ij} \in R$, де приклад t_i характеризується m ознаками. Задано кількість кластерів $2 \le g \le N$ та параметр $w > 1$ - показник нечіткості, який показує розмитість кластерів. Розрахувати центри кластерів за формулою:

    $$c_k = \frac{\sum_{i=1}^{N} (u_{ki})^w \cdot t_i}{\sum_{i=1}^{N} (u_{ki})^w}, k = 1, ..., g,$$

    де $U = \{(u_{ki})|k = 1, ..., g, i = 1, ..., N\}$ - випадковим чином задана матриця початкового розбиття, $u_{ki} \in [0, 1]$, $\sum_{k=1}^{g} u_{ki} = 1$, $\sum_{i=1}^{N} u_{ki} < N$.

    Перерахувати матрицю розбиття:

    $$u_{ki} = \frac{1}{\sum_{v=1}^{g} \left(\frac{d^2(t_i, c_k)}{d^2(t_i, c_v)} \right)^{\frac{1}{w-1}}},$$

використати $d^2(t_i, c_k)$ - квадрат евклідової відстані між векторами.

Виконати декілька ітерацій з уточнення центрів кластерів.

23. Задано неорієнтовний граф J з V вершинами, де ваги дуг d_{ij} відомі для $\forall i, j = 1, ..., V$. Побудувати підграф G графу J за наступними етапами:

   1) Ініціалізувати граф $G := T$ з множиною ребер $E := \varnothing$.

   2) Поки G не зв'язний:

   - Ініціалізувати допоміжну множину ребер $U := \varnothing$.
   - Для кожної компоненти зв'язності графу G:
     - Ініціалізувати допоміжну множину ребер $S := \varnothing$.
     - Для кожної вершини вибраної компоненти зв'язності додати в S найкоротше ребро, яке поєднує цю вершину з якою-небудь вершиною другої компоненти.
     - Додати в U найкоротше ребро з S.
   - $E := E \cup U$.

   Надрукувати граф G.

2 Контрольні питання для захисту роботи

1. **Основи роботи в бібліотеці NumPy**

   - Типи даних в Python.
   - Масиви NumPy:
     - Індексація масива. Доступ до окремих елементів багатовимірних масивів.
     - numpy.reshape. Навести приклади.
     - numpy.newaxis. Навести приклади.
     - Зрізи масивів: доступ до підмасивів.
     - Маскування з використанням булевих масивів.
     - numpy.concatenate. Навести приклади для одновимірного та двовимірного масивів.
     - numpy.vstack і numpy.hstack. Навести приклади.
     - numpy.split, numpy.hsplit, numpy.vsplit. Навести приклади.
     - Операція reduce. Навести приклади.
     - numpy.sum. Навести приклади.
     - numpy.prod. Навести приклади.
     - numpy.mean. Навести приклади.
     - numpy.var. Навести приклади.

  - numpy.amin, numpy.amax. Навести приклади.

- Універсальні функції над масивами в NumPy:
  - Поняття універсальної функції. Навіщо вони потрібні.
  - Арифметичні універсальні функції для масивів.
  - Правила транслювання (broadcasting).
  - Сортування масивів з використанням np.sort.
- Створення структурованих масивів в NumPy.

2. **Оперування даними за допомогою Pandas**

- Створення об'єкту Series бібліотеки Pandas.
- Об'єкт Series як словник.
- Об'єкт Series як одновимірний масив.
- Створення об'єкту DataFrame бібліотеки Pandas.
- Об'єкт DataFrame як словник.
- Об'єкт DataFrame як двовимірний масив.
- Застосування універсальних функцій до об'єктів Series і DataFrame.
- Застосування функцій агрегування до об'єктів Series і DataFrame.
- Опрацювання онлайн-документації бібліотеки Pandas (http://pandas.pydata.org/)

3. **Візуалізація за допомогою Matplotlib**

- Побудова графіків із сценарію. Функція matplotlib.pyplot.show()
- Побудова графіків із блокноту IPython. Функція matplotlib.pyplot.plot().
- Побудова графіку функції $y = f(x)$ за допомогою matplotlib.pyplot.
- Налаштування кольору, стилю ліній, міток на графіках, легенди засобами matplotlib.pyplot.
- Опрацювання онлайн-документації бібліотеки Matplotlib (https://matplotlib.org/)
- Опрацювання онлайн-документації бібліотеки Seaborn (https://seaborn.pydata.org/)