Решение пробного итогового теста по дисциплине «Интеллектуальный анализ данных» для направлений НФИ, НПИ, НБИ

7 октября 2024 г.

1. Подготовка данных

Дан набор данных с числовыми признаками. Вычислить:

- математическое ожидание, медиану, дисперсию заданного признака

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i,\ \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \hat{\mu}\right)^2$$

$$m = \begin{cases} x_{k+1}, & n = 2k+1 \\ \frac{1}{2}\left(x_k + x_{k+1}\right), & n = 2k \end{cases},\ x_1 \leqslant x_2 \leqslant \ldots \leqslant x_n.$$

- значение эмпирической CDF признака в заданной точке

$$\hat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leqslant x)$$

- ковариацию (корреляцию) между заданными признаками

$$\hat{\sigma}_{12} = \frac{1}{n}\sum_{i=1}^{n}\left(x_{i1} - \hat{\mu}_1\right)\left(x_{i2} - \hat{\mu}_2\right)$$

$$\hat{\rho}_{12} = \frac{\hat{\sigma}_{12}}{\hat{\sigma}_1\hat{\sigma}_2} = \frac{\sum_{i=1}^{n}\left(x_{i1} - \hat{\mu}_1\right)\left(x_{i2} - \hat{\mu}_2\right)}{\sqrt{\sum_{i=1}^{n}\left(x_{i1} - \hat{\mu}_1\right)^2 \sum_{i=1}^{n}\left(x_{i2} - \hat{\mu}_2\right)^2}}$$

Задание 1. Дан следующий набор данных с двумя числовыми признаками:

$$D = \left\{(1,\ 1),\ (3,\ 4),\ (2,\ 3),\ (0,\ 4W)\right\},$$

где W – некоторое числовое значение.

Вычислить медиану первого признака и эмпирическую CDF второго признака в точке $x = 3.5$ при $W = \frac{1}{2}$.

Решение: Упорядочим значения первого признака по возрастанию и получим последовательность 0, 1, 2, 3, поэтому медиана равна среднему арифметическому двух значений в середине последовательности:

$$\frac{1+2}{2} = 1.5$$

Упорядочим значения второго признака по возрастанию и получим последовательность 1, 2, 3, 4. Для точки $x = 3.5$ три элемента последовательности из четырех меньше или равны x, поэтому эмпирическая CDF равна

$$\frac{3}{4} = 0.75$$

2. Наивный байесовский классификатор

Задание 2. Дан размеченный набор данных с одним признаком (первая компонента) и метками классов (вторая компонента):

$$\mathbf{D} = \{(0,1)\,,\,(1,1)\,,\,(2,1)\,,\,(4,2)\,,\,(6,2)\}$$

Используя (наивный) байесовский подход, найти прогнозируемый класс для заданной точки $x = 3$

Решение: Определяем априорные вероятности классов c_1 и c_2

$$n = |\mathbf{D}| = 5,$$

$$n_1 = |\mathbf{D}_1| = |\{(0,1)\,,\,(1,1)\,,\,(2,1)\}| = 3,$$

$$n_2 = |\mathbf{D}_2| = |\{(4,2)\,,\,(6,2)\}| = 2,$$

$$\mathbb{P}\left[c_1\right] = \frac{n_1}{n} = \frac{3}{5} = 0.6,\ \mathbb{P}\left[c_2\right] = \frac{n_2}{n} = \frac{2}{5} = 0.4$$

Вычисляем математические ожидания признака для классов c_1 и c_2:

$$\mu_1 = \mathbb{E}\left[X \mid c_1\right] = \frac{1}{3}\left(0+1+2\right) = 1,\ \mu_2 = \mathbb{E}\left[X \mid c_2\right] = \frac{1}{2}\left(4+6\right) = 5$$

и дисперсии признака для классов c_1 и c_2:

$$\sigma_1^2 = \frac{1}{3}\left((0-1)^2 + (1-1)^2 + (2-1)^2\right) = \frac{2}{3} = 0.667,$$

$$\sigma_2^2 = \frac{1}{2}\left((4-5)^2 + (6-5)^2\right) = \frac{2}{2} = 1$$

Для точки $x = 3$ апостериорные вероятности классов c_1 и c_2 будут пропорциональны значениям:

$$p_1 = \frac{1}{\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\}\mathbb{P}\left[c_1\right] = \frac{1}{\sqrt{\frac{2}{3}}}\exp\left\{-\frac{(3-1)^2}{2\,\frac{2}{3}}\right\}\frac{3}{5} = \frac{\sqrt{3}}{5\sqrt{2}}\exp\left\{-3\right\}$$

$$p_2 = \frac{1}{\sigma_2}\exp\left\{-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right\}\mathbb{P}\left[c_2\right] = \frac{1}{\sqrt{1}}\exp\left\{-\frac{(3-5)^2}{2}\right\}\frac{2}{5} = \frac{2}{5}\exp\left\{-2\right\}$$

Так как

$$\frac{p_1^2}{p_2^2} = \frac{\frac{3}{25\cdot 2}\exp\left\{-6\right\}}{\frac{2}{25}\exp\left\{-4\right\}} = \frac{3}{4}\exp\left\{-2\right\} < 1,$$

для заданной точки x = 3 будет спрогнозирован класс 2.

3. Поиск ассоциативных правил

Дана база транзакций D. Для заданного набора предметов X найти:

- поддержку (support)

$$\sup(X,\,\mathbf{D}) = |\{t \mid \langle t,\,\mathbf{i}(t)\rangle \in \mathbf{D},\ X \subseteq \mathbf{i}(t)\}| = |\mathbf{t}(X)|$$

- относительную поддержку (relative support)

$$\operatorname{rsup}(X,\,\mathbf{D}) = \frac{\sup(X,\,\mathbf{D})}{|\mathbf{D}|}$$

Для заданного ассоциативного правила $X \longrightarrow Y$ найти:

- поддержку (support)

$$s = \sup(X \to Y) = |\mathbf{t}(XY)| = \sup(XY)$$

- относительную поддержку (relative support)

$$\operatorname{rsup}(X \to Y) = \frac{\sup(XY)}{|\mathbf{D}|} = \mathbb{P}[X \wedge Y]$$

- достоверность (confidence)

$$c = \operatorname{conf}(X \to Y) = \mathbb{P}[Y \mid X] = \frac{\mathbb{P}[X \wedge Y]}{\mathbb{P}[X]} = \frac{\sup(XY)}{\sup(X)}$$

- лифт (lift)

$$\operatorname{lift}(X \to Y) = \frac{\mathbb{P}[XY]}{\mathbb{P}[X]\,\mathbb{P}[Y]} = \frac{\operatorname{rsup}(XY)}{\operatorname{rsup}(X)\operatorname{rsup}(Y)} = \frac{conf\,(X \to Y)}{\operatorname{rsup}(Y)}$$

- рычаг (leverage)

$$\operatorname{leverage}(X \to Y) = \mathbb{P}[XY] - \mathbb{P}[X]\,\mathbb{P}[Y] = \operatorname{rsup}(XY) - \operatorname{rsup}(X)\operatorname{rsup}(Y)$$

- убежденность (conviction)

$$\operatorname{conv}(X \to Y) = \frac{\mathbb{P}[X]\,\mathbb{P}[\neg Y]}{\mathbb{P}[X \neg Y]} = \frac{1}{\operatorname{lift}(X \to \neg Y)} = \frac{1 - \operatorname{rsup}(Y)}{1 - \operatorname{conf}(X \to Y)}$$

Задание 3. Дана база транзакций

$$\mathbf{D} = \{(1, ABDE), (2, BCE), (3, ABDE), (4, ABCE), (5, ABCDE), (6, BCD)\}$$

Для набора предметов $\{BC\}$ вычислить поддержку (support).

Решение: По определению поддержки набора предметов

$$\sup(X,\,\mathbf{D}) = |\{t \mid \langle t,\,\mathbf{i}(t)\rangle \in \mathbf{D},\ X \subseteq \mathbf{i}(t)\}| = |\mathbf{t}(X)|$$

Поэтому

$$\sup(\{BC\},\,\mathbf{D}) = |\{(2, BCE), (4, ABCE), (5, ABCDE), (6, BCD)\}| = 4$$

4. Кластеризация данных

Дано разбиение набора данных с числовыми признаками на k кластеров. Найти расстояние между кластерами: • методом одиночной связи (single link)

$$\delta\left(C_i, C_j\right) = \min_{\overline{\mathbf{x}}, \overline{\mathbf{y}}} \left\{\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) \mid \overline{\mathbf{x}} \in C_i, \overline{\mathbf{y}} \in C_j\right\}$$

- методом полной связи (complete link)

$$\delta\left(C_i, C_j\right) = \max_{\overline{\mathbf{x}}, \overline{\mathbf{y}}} \left\{\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) \mid \overline{\mathbf{x}} \in C_i, \overline{\mathbf{y}} \in C_j\right\}$$

- методом средней связи (group average)

$$\delta\left(C_i, C_j\right) = \frac{1}{n_i n_j} \sum_{\overline{\mathbf{x}} \in C_i} \sum_{\overline{\mathbf{y}} \in C_j} \delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right)$$

- методом Уарда (Ward's measure)

$$\delta\left(C_i, C_j\right) = \triangle SSE_{ij} = SSE_{ij} - SSE_i - SSE_j,\ SSE_i = \sum_{\overline{\mathbf{x}} \in C_i} \left\|\overline{\mathbf{x}} - \overline{\mu}_i\right\|^2$$

$$\delta\left(C_i, C_j\right) = \frac{n_i n_j}{n_i + n_j} \left\|\overline{\mu}_i - \overline{\mu}_j\right\|^2$$

Задание 4. Дано разбиение набора данных с двумя числовыми признаками на два кластера:

- кластер 1 (C_1): $\{(0,0)\,, (0,1)\,, (1,0)\}$
- кластер 2 (C_2): $\{(2,2)\,, (3,3)\}$

Найти евклидово расстояние между кластерами методом одиночной связи (single link).

Решение: Расстояние между кластерами методом одиночной связи (single link) вычисляется по формуле

$$\delta\left(C_i, C_j\right) = \min_{\overline{\mathbf{x}}, \overline{\mathbf{y}}} \left\{\delta\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) \mid \overline{\mathbf{x}} \in C_i, \overline{\mathbf{y}} \in C_j\right\}$$

Евклидово расстояние вычисляется по формуле

$$\rho\left(\overline{\mathbf{x}}, \overline{\mathbf{y}}\right) = \sqrt{\sum_{i=1}^{d} \left(x_i - y_i\right)^2},\ \overline{\mathbf{x}} = \left(x_1, ..., x_d\right), \overline{\mathbf{y}} = \left(y_1, ..., y_d\right)$$

Наиболее близкими точками двух кластеров являются точка (0, 1) (или точка (1, 0)) и точка (2, 2), поэтому

$$\delta\left(C_1, C_2\right) = \rho\left((0,\,1)\,, (2,\,2)\right) = \sqrt{2^2 + 1^2} = \sqrt{5} = 2.236$$

5. Кластеризация данных

Дана таблица (матрица) сопряженности кластеризации. Вычислить:

- чистоту (purity) или точность (precision) кластеров и чистоту кластеризации

$$p_j = prec_j = \frac{1}{m_j} \max_{i=\overline{1,k}} n_{ij} = \frac{n_{i_j j}}{m_j},\ i_j = \max_{i=\overline{1,k}} n_{ij},\ j = \overline{1,r}$$

$$p = \sum_{j=1}^{r} \frac{m_j}{n} p_j = \frac{1}{n} \sum_{j=1}^{r} \max_{i=\overline{1,k}} n_{ij},$$

- полнота (recall) кластеров

$$recall_j = \frac{n_{i_j j}}{\left|T_{i_j}\right|} = \frac{n_{i_j j}}{n_{i_j}},\ n_{i_j} = \left|T_{i_j}\right|,\ j = \overline{1,r},$$

- F-меру кластеров и F-меру кластеризации

$$F_j = \frac{2}{\frac{1}{prec_j} + \frac{1}{recall_j}} = \frac{2n_{i_j j}}{n_{i_j} + m_j},\ j = \overline{1,r},\ F = \frac{1}{r} \sum_{j=1}^{r} F_j$$

- энтропию кластеризации (разбиения на кластеры)

$$H\left(\mathcal{C}\right) = -\sum_{j=1}^{r} p_{C_j} \log_2 p_{C_j},\ p_{C_j} = \frac{m_j}{n}$$

- энтропию разбиения на классы

$$H\left(\mathcal{T}\right) = -\sum_{i=1}^{k} p_{T_i} \log_2 p_{T_i},\ p_{T_i} = \frac{n_i}{n}$$

- условную энтропию относительно кластера

$$H\left(\mathcal{T} \mid C_j\right) = -\sum_{i=1}^{k} \left(\frac{n_{ij}}{n_i}\right) \log_2 \left(\frac{n_{ij}}{n_i}\right)$$

- условную энтропию относительно кластеризации

$$H\left(\mathcal{T} \mid \mathcal{C}\right) = \sum_{j=1}^{r} \frac{m_j}{n} H\left(\mathcal{T} \mid C_j\right) = -\sum_{j=1}^{r} \sum_{i=1}^{k} p_{ij} \log_2 \left(\frac{p_{ij}}{p_{C_j}}\right),\ p_{ij} = \frac{n_{ij}}{n}$$

- парные меры FN, FP, TP, TN

$$TP = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j, \hat{y}_i = \hat{y}_j\right\}\right| = \sum_{i=1}^{k} \sum_{j=1}^{r} \binom{n_{ij}}{2} = \frac{1}{2} \left(\left(\sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij}^2\right) - n\right)$$

$$FN = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j) : y_i = y_j, \hat{y}_i \neq \hat{y}_j\right\}\right| = \sum_{i=1}^{k} \binom{n_i}{2} - TP = \frac{1}{2} \left(\sum_{i=1}^{k} n_i^2 - \sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij}^2\right)$$

$$FP = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j, \hat{y}_i = \hat{y}_j\right\}\right| = \sum_{j=1}^{r} \binom{m_j}{2} - TP = \frac{1}{2} \left(\sum_{j=1}^{r} m_j^2 - \sum_{i=1}^{k} \sum_{j=1}^{r} n_{ij}^2\right)$$

$$TN = \left|\left\{(\mathbf{x}_i, \mathbf{x}_j) : y_i \neq y_j, \hat{y}_i \neq \hat{y}_j\right\}\right| = N - (TP + FN + FP) =$$

$$= \frac{1}{2}\left(n^2 - \sum_{i=1}^{k} n_i^2 - \sum_{j=1}^{r} m_j^2 + \sum_{i=1}^{k}\sum_{j=1}^{r} n_{ij}^2\right)$$

- индекс Rand

$$Rand = \frac{TP + TN}{TP + TN + FP + FN}$$

- индекс Жаккара (Jaccard Index)

$$Jaccard = \frac{TP}{TP + FP + FN}$$

- индекс Фоулкса – Мэллоуса (Fowlkes-Mallows Index)

$$FM = \sqrt{\frac{TP}{TP + TN}\frac{TP}{TP + FP}}$$

Задание 5. Дана таблица сопряженности кластеризации с двумя классами и тремя кластерами:

$$\begin{pmatrix} 5 & 5 & 5 \\ 3 & 10 & K \end{pmatrix},$$

где K – параметр, принимающий целое положительное значение, причем $K > 5$.

Вычислить чистоту (purity) каждого кластера.

Решение: Если разбиение на классы задано в виде $\mathcal{T} = \{T_1, T_2, ..., T_k\}$, а разбиение на кластеры имеет вид $\mathcal{C} = \{C_1, ..., C_r\}$, то (расширенная) таблица (матрица) сопряженности имеет вид

$\mathcal{T} \setminus \mathcal{C}$	C_1	C_2	$\ldots$	C_r	$\sum$
T_1	n_{11}	n_{12}	$\ldots$	n_{1r}	n_1
T_2	n_{21}	n_{22}	$\ldots$	n_{2r}	n_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
T_k	n_{k1}	n_{k2}	$\ldots$	n_{kr}	n_k
$\sum$	m_1	m_2	$\ldots$	m_r	n

Чистота (purity) кластеров задается формулами

$$p_j = \frac{1}{m_j}\max_{i=\overline{1,k}} n_{ij} = \frac{n_{i_j j}}{m_j},\ i_j = \arg\max_{i=\overline{1,k}} n_{ij},\ j = \overline{1,r}$$

Строим расширенную таблицу сопряженности:

$$\left(\begin{array}{ccc|c} 5 & 5 & 5 & 15 \\ 3 & 10 & K & K+13 \\ \hline 8 & 15 & K+5 & K+28 \end{array}\right)$$

и определяем чистоту кластеров:

$$p_1 = \frac{5}{8} = 0.625,\ p_2 = \frac{10}{15} = 0.667,\ p_3 = \frac{K}{K+5}$$