

Статистический анализ данных. Спецкурс.

Лекция 3. Анализ нечисловых данных

Ботанический сад-институт ДВО РАН

Кислов Д.Е.
13 ноября 2016 г.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- неприменимы многие статистические понятия (среднее, дисперсия и т.п.)

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- неприменимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- неприменимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков

Обработка нечисловых данных. Специфика задач.

Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- неприменимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков
- попробовать ввести количественные меры «близости» объектов (учитывая специфику задачи)

Обработка нечисловых данных. Специфика задач.

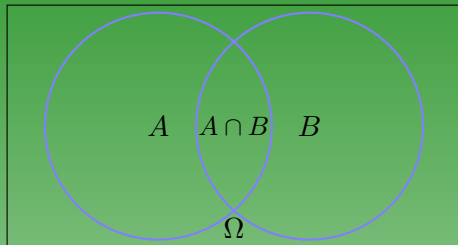
Особенности:

- нельзя применять арифметические операции (сложение, вычитание и т.п.)
- не всегда просто судить о сходстве объектов
- неприменимы многие статистические понятия (среднее, дисперсия и т.п.)

Как обрабатывать:

- проблемно-ориентированный подход: иногда можно назначать нечисловым показателям числовые метки
- можно работать с таблицами сопряженности признаков
- попробовать ввести количественные меры «близости» объектов (учитывая специфику задачи)
- оценивать вероятности наличия тех или иных признаков на основе предельных теорем

Меры сходства и различия



Основные меры:

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

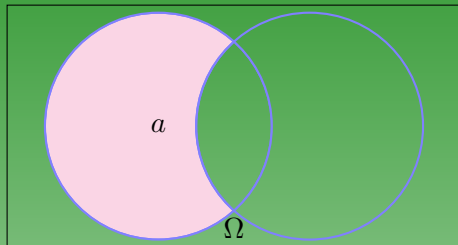
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



Основные меры:

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

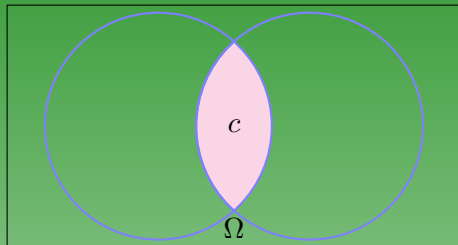
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



Основные меры:

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

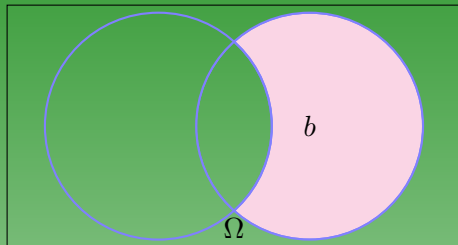
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



Основные меры:

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

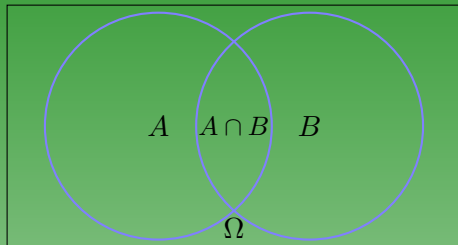
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



Основные меры:

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

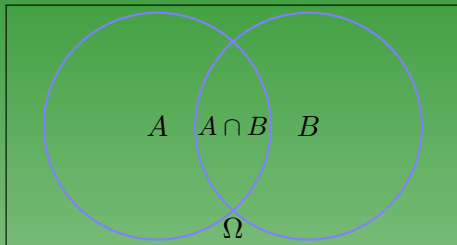
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



Основные меры:

- $\frac{c}{a + b + c}$ (Jaccard, 1901)

	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

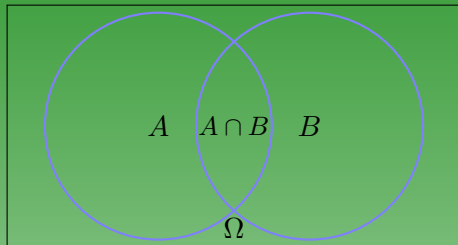
$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Меры сходства и различия



	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

$$a = |A - B|$$

$$b = |B - A|$$

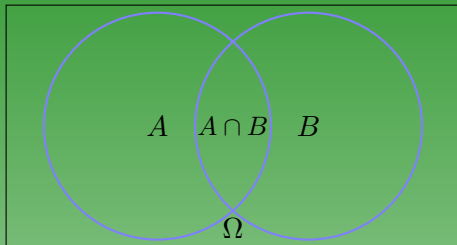
$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Основные меры:

- $\frac{c}{a + b + c}$ (Jaccard, 1901)
- $\frac{2c}{a + b + 2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)

Меры сходства и различия



	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

$$a = |A - B|$$

$$b = |B - A|$$

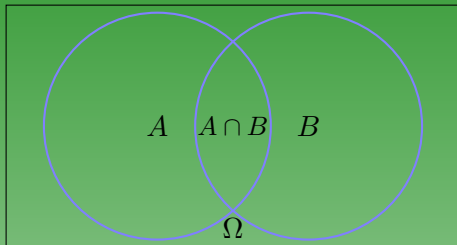
$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Основные меры:

- $\frac{c}{a + b + c}$ (Jaccard, 1901)
- $\frac{2c}{a + b + 2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)
- $\frac{c}{a + b}$ (Кульчинский, 1927)

Меры сходства и различия



	B	\overline{B}	Σ
A	c	a	$a + c$
\overline{A}	b	d	$b + d$
Σ	$b + c$	$a + d$	$a + b + c + d$

$$a = |A - B|$$

$$b = |B - A|$$

$$c = |A \cap B|$$

$$d = |\Omega - A \cup B|$$

Основные меры:

- $\frac{c}{a + b + c}$ (Jaccard, 1901)
- $\frac{2c}{a + b + 2c}$ (Чекановский, 1900; Dice, 1945; Sørensen, 1948)
- $\frac{c}{a + b}$ (Кульчинский, 1927)
- $\frac{c}{c + a}, \frac{c}{c + b}$ (Шимкевич, 1926; Simpson, 1943)

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если A – стул, B – тренога, то $a = 1$ и $b = 0$.

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если А – стул, В – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \frac{3}{1 + 0 + 3} = 0.75$

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если А – стул, В – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \frac{3}{1 + 0 + 3} = 0.75$
- мера Дайса $D = \frac{2 \cdot 3}{1 + 0 + 2 \cdot 3} = 6/7 \approx 0.857$

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если А – стул, В – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \frac{3}{1 + 0 + 3} = 0.75$
- мера Дайса $D = \frac{2 \cdot 3}{1 + 0 + 2 \cdot 3} = 6/7 \approx 0.857$
- мера Кульчинского $K = \frac{3}{1 + 0} = 3$

Стул и тренога

Имеют 4 и 3 «ноги» соответственно. Общее число ног $c = 3$.
Если А – стул, В – тренога, то $a = 1$ и $b = 0$.

- мера Жаккара $J = \frac{3}{1 + 0 + 3} = 0.75$
- мера Дайса $D = \frac{2 \cdot 3}{1 + 0 + 2 \cdot 3} = 6/7 \approx 0.857$
- мера Кульчинского $K = \frac{3}{1 + 0} = 3$

Вычисления на R

```
library(sets)
set_similarity(set(1,2,3,4),set(1,2,5),method="Jaccard")
[1] 0.4
```

Двухпараметрическое семейство мер (Б.И. Семкин, 2010):

$$K_{\tau;\eta} = \left(\frac{K_{\tau}^{\eta}(A, B) + K_{\tau}^{\eta}(B, A)}{2} \right)^{1/\eta},$$

$$K_{\tau}(A, B) = \frac{|A \cap B|}{(1 + \tau)|A| - \tau|A \cap B|},$$

$$K_{\tau}(B, A) = \frac{|A \cap B|}{(1 + \tau)|B| - \tau|A \cap B|}$$

$$-1 < \tau < \infty, -\infty < \eta < \infty$$

В этом случае $K_{0;-1}$ и $K_{1;-1}$ совпадают с коэффициентами Сёренсена-Дайса и Жаккара соответственно.

Двухпараметрическое семейство мер (Б.И. Семкин, 2010):

$$K_{\tau;\eta} = \left(\frac{K_{\tau}^{\eta}(A, B) + K_{\tau}^{\eta}(B, A)}{2} \right)^{1/\eta},$$
$$K_{\tau}(A, B) = \frac{|A \cap B|}{(1 + \tau)|A| - \tau|A \cap B|},$$
$$K_{\tau}(B, A) = \frac{|A \cap B|}{(1 + \tau)|B| - \tau|A \cap B|}$$
$$-1 < \tau < \infty, -\infty < \eta < \infty$$

В этом случае $K_{0;-1}$ и $K_{1;-1}$ совпадают с коэффициентами Сёренсена-Дайса и Жаккара соответственно.

Вывод

Используемые меры имеют много общего, они в определенном смысле «эквивалентны»

Задача

Исследуется вопрос об эффективности обработки с целью последующего проращивания жёлудей. В результате эксперимента построена следующая таблица сопряженности:

	не взошел	взошел	Σ
обработано	1	10	11
не обработано	4	3	7
Σ	5	13	18

Целесообразно ли применение данной обработки, или «увеличение» всхожести в результате обработки вполне могло возникнуть случайно?

Задача

Исследуется вопрос об эффективности обработки с целью последующего проращивания жёлудей. В результате эксперимента построена следующая таблица сопряженности:

	не взошел	взошел	Σ
обработано	1	10	11
не обработано	4	3	7
Σ	5	13	18

Целесообразно ли применение данной обработки, или «увеличение» всхожести в результате обработки вполне могло возникнуть случайно?

Решение

Нужно вычислить вероятность реализации таблицы $[(1, 10), (4, 3)]$, а также более «худшего» варианта, $[(0, 11), (5, 2)]$, т.е. когда после обработки вообще все семена взошли. Если сумма этих вероятностей будет мала, то, вероятно, что обработка (а не случайность) определяет исход проращивания.

Решение

$$\frac{C_5^1 \cdot C_{13}^{10}}{C_{18}^{11}} + \frac{C_5^0 C_{13}^{11}}{C_{18}^{11}} =$$
$$\frac{1430}{31824} + \frac{78}{31824} \approx 0.047$$

Решение

$$\frac{C_5^1 \cdot C_{13}^{10}}{C_{18}^{11}} + \frac{C_5^0 C_{13}^{11}}{C_{18}^{11}} =$$
$$\frac{1430}{31824} + \frac{78}{31824} \approx 0.047$$

Таким образом, вероятность наблюдать исход эксперимента, или даже исход, когда все желуди взошли вследствие случая (а не действия обработки), равна около 4.7%; это весьма маленькое значение, поэтому результаты наблюдений следует интерпретировать, что имеет место значимое влияние обработки на результат прорастания желудей.

Общий случай

	не взошел	взошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

$$P(a, b; c, d) = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$

$$n = a + b + c + d$$

Общий случай

	не вошел	вошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

$$P(a, b; c, d) = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$

$$n = a + b + c + d$$

Односторонний тест: «усугубление» наблюдаемой ситуации

$$\sum_{j=0}^a P(j, \tilde{b}; \tilde{c}, \tilde{d}), \text{ при условии: } j + \tilde{c} = a + c, \tilde{b} + \tilde{d} = b + d,$$

$$\tilde{b} + j = a + b, j + \tilde{b} + \tilde{c} + \tilde{d} = n$$

Общий случай

	не взошел	взошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

$$P(a, b; c, d) = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$
$$n = a + b + c + d$$

Двусторонний тест

Общий случай

	не взошел	взошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

$$P(a, b; c, d) = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$
$$n = a + b + c + d$$

Двусторонний тест: что считать «услужением»?

Общий случай

	не вошел	вошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

$$P(a, b; c, d) = \frac{C_{a+c}^a \cdot C_{b+d}^b}{C_n^{a+b}} = \frac{(a+c)!(b+d)!(a+b)!(c+d)!}{a!b!c!d!n!},$$

$$n = a + b + c + d$$

Двусторонний тест: что считать «усугублением»?

$\sum_{\tilde{a}, \tilde{b}, \tilde{c}, \tilde{d}} P(\tilde{a}, \tilde{b}; \tilde{c}, \tilde{d})$, суммирование при условиях:

$$P(\tilde{a}, \tilde{b}; \tilde{c}, \tilde{d}) \leq P(a, b; c, d), \quad \tilde{a} + \tilde{b} = a + b, \quad \tilde{c} + \tilde{d} = c + d \dots$$

Точный тест Фишера: приближенные вычисления

	не вошел	вошел	Σ
обработано	a	b	$a + b$
не обработано	c	d	$c + d$
Σ	$a + c$	$b + d$	$a + b + c + d$

Гипотеза: наблюдаемое распределение a, b, c, d результат случая

Аппроксимация распределением χ^2 (с поправкой Ейтса)

$$\chi^2_{\text{выч.}} = \frac{n \left(|ad - bc| - \frac{n}{2} \right)^2}{(a + b)(a + c)(b + d)(c + d)}$$
$$n = a + b + c + d$$

условия применимости: $a, b, c, d \geq 5, n \geq 40$

Гипотеза отвергается на уровне значимости α , если $\chi^2_{\text{выч.}} > \chi^2_{1-\alpha}(1)$ (в частности, $\chi^2_{0.96}(1) \approx 3.85$, 1 – число степеней свободы для таблицы 2×2)

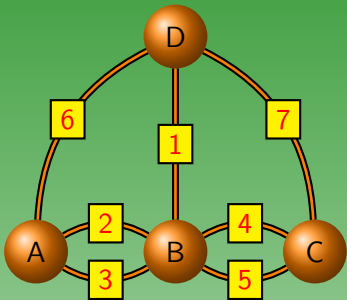
Задачи

- определение и визуализация сложных зависимостей;

Задачи

- определение и визуализация сложных зависимостей;
- идентификация и анализ сходных групп объектов зависимостей;

Пример графа (задача о Кёнигсбергских мостах)



Формулировка задачи

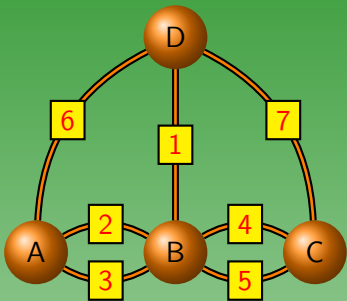
Возможно ли пройти по всем ребрам графа лишь один раз?

Ответ

Нет

Графы, для которых ответ «да», носят название эйлеровых графов.

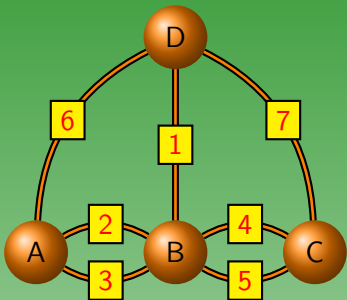
Пример графа (задача о Кёнигсбергских мостах)



Определения

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;

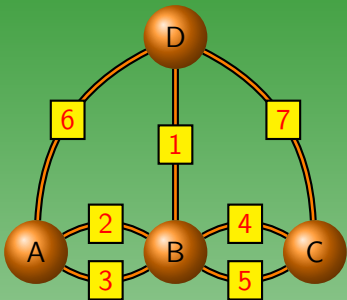
Пример графа (задача о Кёнигсбергских мостах)



Определения (нестрогие)

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;
- Граф называется взвешенным, если ребрам присвоены веса;

Пример графа (задача о Кёнигсбергских мостах)

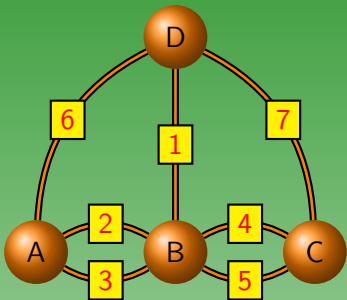


- Граф называется ориентированным, если указывается направления ребер, которые соединяют вершины;

Определения (нестрогие)

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;
- Граф называется взвешенным, если ребрам присвоены веса;

Пример графа (задача о Кёнигсбергских мостах)

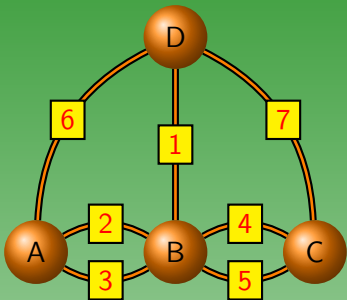


Определения (нестрогие)

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;
- Граф называется взвешенным, если ребрам присвоены веса;

- Граф называется ориентированным, если указывается направления ребер, которые соединяют вершины;
- Граф называется полным, если все вершины друг с другом соединены ребрами;

Пример графа (задача о Кёнигсбергских мостах)

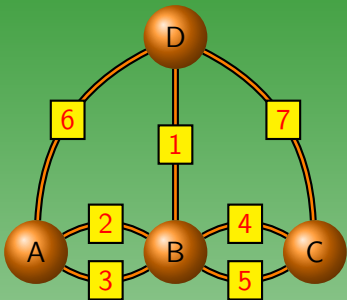


Определения (нестрогие)

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;
- Граф называется взвешенным, если ребрам присвоены веса;

- Граф называется ориентированным, если указывается направления ребер, которые соединяют вершины;
- Граф называется полным, если все вершины друг с другом соединены ребрами;
- Вершина называется изолированной, если из нее не исходит ни одного ребра;

Пример графа (задача о Кёнигсбергских мостах)



Определения (нестрогие)

- Графом называется совокупность $(\mathcal{V}, \mathcal{E})$, где \mathcal{V} – множество вершин, а \mathcal{E} – множество ребер;
- Граф называется взвешенным, если ребрам присвоены веса;

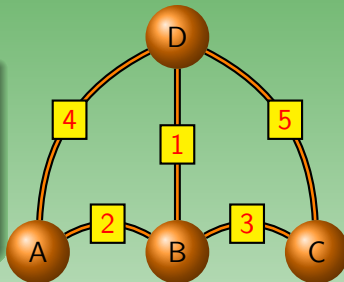
- Граф называется ориентированным, если указывается направления ребер, которые соединяют вершины;
- Граф называется полным, если все вершины друг с другом соединены ребрами;
- Вершина называется изолированной, если из нее не исходит ни одного ребра;
- Граф называется связным, если между любыми двумя вершинами существует путь по его ребрам;

Матрица смежности

Если между вершинами i и j существует ребро, то элемент $a_{i,j}$ матрицы смежности полагается равным 1, в противном случае 0.

Пример матрицы смежности

$$M = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

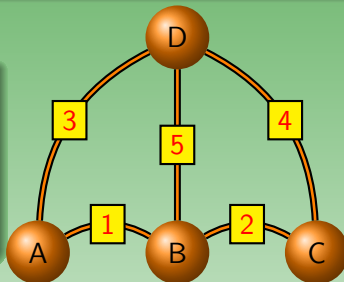


Матрица инциденций

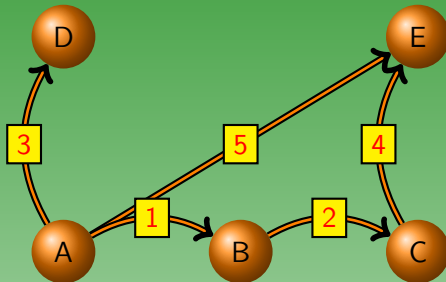
Номер столбца матрицы соответствует номеру ребра. Номер строки – номеру вершины графа. Единицы ставятся в случае, если ребро инцидентно двум вершинам (т.е. соединяет две вершины).

Пример матрицы инциденций

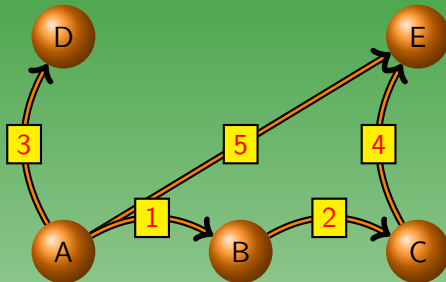
$$M = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$



Ориентированные графы



Ориентированные графы



Для представления графов в ЭВМ применяются различные варианты несимметричных матриц смежности, а также матрицы инцидентий;

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;
- 2 поиск компонент связности;

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;
- 2 поиск компонент связности;
- 3 поиск минимального остовного дерева (пути во взвешенном связном графе, проходящем через все вершины);

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;
- 2 поиск компонент связности;
- 3 поиск минимального остовного дерева (пути во взвешенном связном графе, проходящем через все вершины);
- 4 классификационные задачи;

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;
- 2 поиск компонент связности;
- 3 поиск минимального остовного дерева (пути во взвешенном связном графе, проходящем через все вершины);
- 4 классификационные задачи: деревья решений;
- 5 поиск оптимальных потоков;

Задачи связанные с теорией графов

- 1 визуализация, структуризация данных;
- 2 поиск компонент связности;
- 3 поиск минимального остовного дерева (пути во взвешенном связном графе, проходящем через все вершины);
- 4 классификационные задачи: деревья решений, дендрограммы;
- 5 поиск оптимальных потоков;

Python

- ▶ NetworkX (<https://networkx.github.io/>)
- ▶ SciPy (<https://scipy.org>)

R

- ▶ gRBase (<https://cran.r-project.org/web/packages/gRbase/vignettes/gRbase-graphs.pdf>)

Программные пакеты для работы с графами

Python

- ▶ NetworkX (<https://networkx.github.io/>)
- ▶ SciPy (<https://scipy.org>)
- ▶ Google search: python workign with graphs

R

- ▶ gRBase (<https://cran.r-project.org/web/packages/gRbase/vignettes/gRbase-graphs.pdf>)
- ▶ Search: workign with graphs, graphs