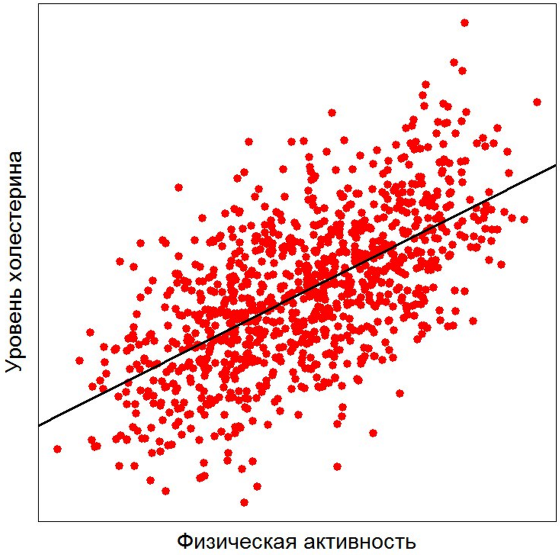
Прикладной статистический анализ данных. 11. Причинно-следственные связи.

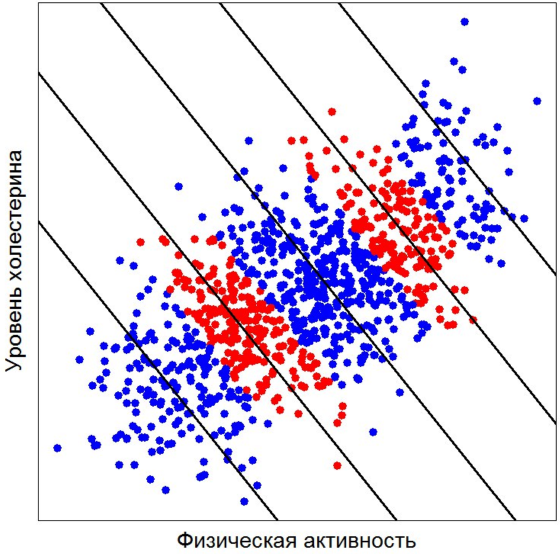
Швечиков Павел

7 апреля 2017 г.

Исследование уровня холестерина



Исследование уровня холестерина



Парадокс Симпсона

Пример 1:

Σ	лекарство	плацебо
выздоровели	273	289
не выздоровели	77	61
	78%	83%

плацебо на 5% эффективнее

мужчины	лекарство	плацебо
выздоровели	81	234
не выздоровели	6	36
	93%	87%

лекарство на 5% эффективнее

женщины	лекарство	плацебо
выздоровели	192	55
не выздоровели	71	25
	73%	69%

лекарство на 4% эффективнее

Парадокс Симпсона

Какой из двух выводов верен?

Предположение: верны выводы по отдельным подгруппам, потому что они основаны на более детальной информации.

Это предположение неверно — всё зависит от того, как признак, по которому происходит разбиение на подгруппы, связан с остальными анализируемыми признаками.

Парадокс Симпсона

Пример 2:

Σ	лекарство	плацебо
выздоровели	273	289
не выздоровели	77	61
	78%	83%

плацебо на 5% эффективнее

низкое давление в конце лечения	лекарство	плацебо
выздоровели	81	234
не выздоровели	6	36
	93%	87%

лекарство на 5% эффективнее

низкое давление в конце лечения	лекарство	плацебо
выздоровели	192	55
не выздоровели	71	25
	73%	69%

лекарство на 4% эффективнее

Причинные графы

Отношения причинности могут быть представлены в виде направленного графа, вершины которого соответствуют признакам, а наличие пути говорит о существовании причинно-следственной связи.

Путь — последовательность вершин, где каждая вершина соединена со следующей ребром.
Направленный путь — путь, в котором все ребра имеют одинаковое направление.
Путь через заднюю дверь (ПЗД) от X к Y — такой путь, который начинается с $X \leftarrow$, а заканчивается $\rightarrow Y$.

Элементы причинного графа

$$X \to Y \to Z$$ — **цепочка**

Пример:

- X — бюджет школы
- Y — средний балл учеников
- Z — доля поступающих в ВУЗы

Свойства:

1. X и Y, Y и Z — зависимы:
   $\exists x, y : \mathbf{P}(Y = y \,|X = x\,) \neq \mathbf{P}(Y = y)$
   $\exists y, z : \mathbf{P}(Z = z \,|Y = y\,) \neq \mathbf{P}(Z = z)$
2. Z и X — скорее всего, зависимы
3. $Z \perp X|Y$ — условно независимы: $\forall x, y, z$

$$\mathbf{P}(Z = z \,|X = x, Y = y\,) = \mathbf{P}(Z = z \,|Y = y\,)$$

(если Y фиксировано, то X и Z независимы)

Элементы причинного графа

$$Y \leftarrow X \rightarrow Z \text{ — } \textbf{вилка}$$

Пример:

- X — продажи мороженого
- Y — средняя дневная температура воздуха
- Z — число преступлений

Свойства:

1. X и Y, X и Z — зависимы
2. Y и Z — скорее всего, зависимы
3. $Y \perp Z | X$ — условно независимы

Элементы причинного графа

$$Y \to X \leftarrow Z$$ — **коллайдер**

Пример (парадокс Монти-Холла):

- X — выбор ведущего
- Y — выбор игрока
- Z — положение приза

Свойства:

1. Y и X, Z и X — зависимы
2. Y и Z — независимы
3. $Y \not\perp Z|X$ — условно зависимы

d-разделимость

Путь P **блокируется** переменной Z, если:

1. P содержит $A \to B \to C$, $A \leftarrow B \to C$, $B \in Z$
2. P содержит $A \to B \leftarrow C$, $B \notin Z$ и все потомки $B \notin Z$

Если Z блокирует все пути из X в Y, то X и Y **d-разделимы**.

d-разделимость

Пример:

$$\begin{tikzcd}
Z_1 \arrow[dd] \arrow[dr] & & Z_2 \arrow[dl] \arrow[dd] \\
& Z_3 \arrow[dl] \arrow[dr] & \\
X \arrow[r] & W \arrow[r] & Y
\end{tikzcd}$$

Упорядоченная пара вершин	d-разделяющее множество
(Z_1, W)	X
(Z_1, Y)	
(X, Y)	

d-разделимость

Пример:

$$\begin{tikzcd}
Z_1 \arrow[dd] \arrow[dr] & & Z_2 \arrow[dl] \arrow[dd] \\
& Z_3 \arrow[dl] \arrow[dr] & \\
X \arrow[r] & W \arrow[r] & Y
\end{tikzcd}$$

Упорядоченная пара вершин	d-разделяющее множество
(Z_1, W)	X
(Z_1, Y)	$\{Z_3, X, Z_2\}, \{Z_3, W, Z_2\}$
(X, Y)	$\{W, Z_3, Z_1\}$

Алгоритм индуктивной причинности

Вход: множество вершин V

1. $\forall A, B \in V$ ищем множество S_{AB}: $A \perp B | S_{AB}$, $A, B \notin S_{AB}$. Если такого S_{AB} не существует, соединяем A и B ребром.
2. $\forall A, B$, не связанных ребром и имеющих общего соседа C, проверяем: $C \in S_{AB}$? Если нет, то заменяем пару рёбер $A - C, C - B$ на пару ориентированных рёбер $A \to C, C \leftarrow B$
3. Рекурсивно применяем следующие два правила:
   - если из A в B есть ориентированный путь $A \to \cdots \to B$, то $A - B$ заменяем на $A \to B$;
   - если A и B не соединены, $A \to C$, $C - B$, то $C - B$ заменяем на $C \to B$.

Выход: ориентированный (возможно, частично) граф G.

Алгоритм индуктивной причинности

Правила (1) и (2) применять в чистом виде невозможно — числи перебираемых множеств экспоненциально растёт с числом вершин графа. Поэтому используются сокращающие перебор эвристики.

Признаки	дискретные	непрерывные
Распределение	мультиномиальное	нормальное
Критерий условной независимости	хи-квадрат для трёхмерных таблиц сопряжённости	Стьюдента для частной корреляции
Критерий качества графа	BIC	

Причинность по Грейнджеру

Между рядами $x_1, \ldots, x_T$ и $y_1, \ldots, y_T$ существует **причинная связь Грейнджера** $x_t \to y_t$, если дисперсия ошибки оптимального прогноза $\hat{y}_{t+1}$ по $y_1, \ldots, y_t, x_1, \ldots, x_t$ меньше, чем только по $y_1, \ldots, y_t$.

Причинность по Грейнджеру

- может следовать из причинно-следственной связи;
- не является достаточным условием причинно-следственной связи.

$x_1, \ldots, x_T$ и $y_1, \ldots, y_T$ **взаимосвязаны**, если $x_t \to y_t$ и $y_t \to x_t$.

Критерий Грейнджера

$$y_t = \alpha + \sum_{i=1}^{k_1} \phi_{1i} y_{t-i} + \sum_{i=1}^{k_2} \phi_{2i} x_{t-i} + \varepsilon_t.$$

k_1 и k_2 выбирается по информационному критерию.

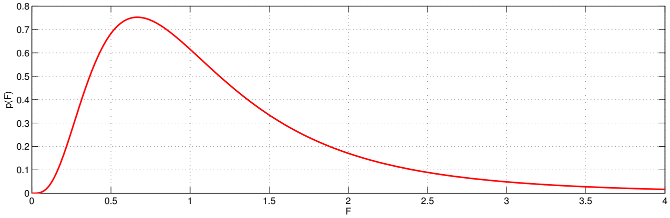
$$x_t \to y_t \quad \Rightarrow \quad \exists \phi_{2i} \neq 0.$$

нулевая гипотеза: $H_0\colon \phi_{21} = \dots = \phi_{2k_2} = 0$;

альтернатива: $H_1\colon H_0$ неверна;

статистика: $F = \frac{(RSS_r - RSS_{ur})/k_2}{RSS_{ur}/(T-k_1-k_2-1)}$;

$F \sim F(k_1, T-k_1-k_2-1)$ при H_0.



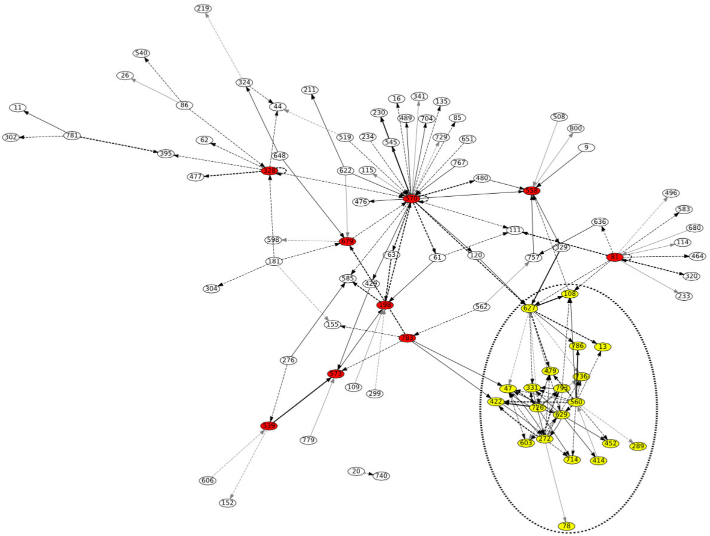
Многомерный критерий Грейнджера

Зависимость между признаками x и y может оцениваться с учётом возможной зависимости от всех остальных признаков:

$$y_t = \alpha + \sum_{i=1}^{k_1} \phi_{1i} y_{t-i} + \sum_{i=1}^{k_2} \phi_{2i} x_{t-i} + \sum_{j=1}^{m} \sum_{i=1}^{k_{j+2}} \phi_{(j+2)i} z^j_{t-i} + \varepsilon_t.$$

Для задач с большим количеством признаков могут использоваться регуляризаторы (лассо, ридж).

Граф причинности по Грейнджеру



Критерий Грейнджера + поправка на множественную проверку гипотез

Интервенция

X коррелировано с $Y \nRightarrow X$ влияет на Y.
Влияние обычно оценивают в эксперименте, когда объектам искусственно назначают разные уровни X, но эксперимент можно провести не всегда:

- погода $\to$ лесные пожары — не можем управлять X
- теленасилие $\to$ жестокость — тяжело фиксировать уровень X и создать условия для измерения Y
- потребление алкоголя $\to$ успеваемость школьников — неэтично

В таких случаях мы вынуждены использовать обзервационные данные, по которым мы хотим оценить эффект **интервенции**: что будет с Y, если мы установим значение X равным x?
Обозначение: $do(X = x)$.

Интервенция

$$\begin{tikzcd}
 & Z \text{ (пол)} \arrow[dl] \arrow[dr] & \\
X \text{ (лечение)} \arrow[rr] & & Y \text{ (исход)}
\end{tikzcd}$$

Оценку эффективности лекарства можно сформулировать в терминах интервенций:

$$\begin{aligned} ACE = &\mathbf{P}(Y = \text{выздоровление} \,|\, do\,(X = \text{лекарство})\,) - \\ &- \mathbf{P}(Y = \text{выздоровление} \,|\, do\,(X = \text{плацебо})\,)\,. \end{aligned}$$

(average conditional effect).

Хирургия графа

Хирургия графа — удаление всех ребер, входящих в X.

Пример 1, исходный граф G:

$$
\begin{tikzcd}
& Z\ (\text{пол}) \arrow[dl] \arrow[dr] & \\
X\ (\text{лечение}) \arrow[rr] & & Y\ (\text{исход})
\end{tikzcd}
$$

Оперированный граф G_m:

$$
\begin{tikzcd}
& Z\ (\text{пол}) \arrow[dr] & \\
X\ (\text{лечение}) \arrow[rr] & & Y\ (\text{исход})
\end{tikzcd}
$$

$$\mathbf{P}(Y = y \,|\, do\,(X = x)\,) = \mathbf{P}_m(\,Y = y|\,X = x)$$

Хирургия графа

В оперированном графе:

$$\mathbf{P}_m(Z=z)=\mathbf{P}(Z=z)\,,$$
$$\mathbf{P}_m(Y=y|\,X=x,Z=z)=\mathbf{P}(Y=y\,|X=x,Z=z\,)\,,$$

так как рёбра, входящие в Z и Y, не изменились $\Rightarrow$

$$\begin{aligned}\mathbf{P}(Y=y\,|do\,(X=x)\,)=\mathbf{P}_m(Y=y|\,X=x)=\\ =\sum_z \mathbf{P}_m(Y=y|\,X=x,Z=z)\,\mathbf{P}_m(Z=z)=\\ =\sum_z \mathbf{P}(Y=y\,|X=x,Z=z\,)\,\mathbf{P}(Z=z)\,.\end{aligned}$$

Хирургия графа

В примере 1 по полученной формуле:

$\mathbf{P}(Y = \text{выздоровление}\,|do\,(X = \text{лекарство})\,) = 0.832,$

$\mathbf{P}(Y = \text{выздоровление}\,|do\,(X = \text{плацебо})\,) = 0.7818$

$\Rightarrow ACE = 0.05.$

В примере 2 $G = G_m$:

$$\begin{tikzcd}
 & Z \text{ (давление)} \arrow[dr] & \\
X \text{ (лечение)} \arrow[ur] \arrow[rr] & & Y \text{ (исход)}
\end{tikzcd}$$

Значит,

$\mathbf{P}(Y = y\,|do\,(X = x)\,) = \mathbf{P}_m(Y = y|\,X = x) = \mathbf{P}(Y = y\,|X = x\,)$

$\mathbf{P}(Y = \text{выздоровление}\,|do\,(X = \text{лекарство})\,) = 0.78,$

$\mathbf{P}(Y = \text{выздоровление}\,|do\,(X = \text{плацебо})\,) = 0.83$

$\Rightarrow ACE = -0.05.$

Поправочная формула

Поправочная формула позволяет вычислить эффект интервенции обуславливанием по вершинам Z:

$$\mathbf{P}(Y = y \,|do\,(X = x)\,) = \sum_z \mathbf{P}(Y = y \,|X = x, Z = z\,)\,\mathbf{P}(Z = z)\,.$$

Что это за вершины?

Формула причинного эффекта:

$$\mathbf{P}(Y = y \,|do\,(X = x)\,) = \sum_z \mathbf{P}(Y = y \,|X = x, PA = z\,)\,\mathbf{P}(PA = z)\,,$$

где PA — родители вершины X.

Неизвестные родители

$$\begin{tikzcd} Z \text{ (соц.-экон. статус)} \arrow[r] \arrow[d] & W \text{ (вес)} \arrow[d] \\ X \text{ (лечение)} \arrow[r] & Y \text{ (исход)} \end{tikzcd}$$

Социоэкономический статус — ненаблюдаемая величина; как оценить эффект интервенции по X?

Критерий задней двери (КЗД)

Для упорядоченной пары вершин (X, Y) в ациклическом графе G множество вершин Z удовлетворяет **критерию задней двери**, если:

- Z не содержит потомков X
- Z блокирует все пути между X и Y, содержащие $X \leftarrow$.

Если Z удовлетворяет КЗД для (X, Y), то

$$\mathbf{P}(Y = y \,|\, do\,(X = x)\,) = \sum_z \mathbf{P}(Y = y \,|\, X = x, Z = z\,)\,\mathbf{P}(Z = z)$$

(формула задней двери).

Критерий задней двери (КЗД)

Чтобы вычислять меньше условных вероятностей, ФЗД можно упростить:

$$\mathbf{P}(Y = y \,|\, do\,(X = x)\,) = \sum_z \mathbf{P}(Y = y \,|\, X = x, Z = z\,)\,\mathbf{P}(Z = z) =$$

$$= \sum_z \frac{\mathbf{P}(X = x, Y = y, Z = z)}{\mathbf{P}(X = x \,|\, Z = z\,)}$$

В таком виде

- метод называется **обратное вероятностное взвешивание**
- знаменатель $\mathbf{P}(X = x \,|\, Z = z\,)$ — propensity score.

Неизвестные родители

Вызывает ли курение рак?

$$\begin{tikzcd} & U \text{ (генотип)} \arrow[dl] \arrow[dr] & \\ X \text{ (курение)} \arrow[rr] & & Y \text{ (рак)} \end{tikzcd}$$

Σ	курильщики	некурящие
нет рака	341	59
есть рак	39	361
	15%	90.25%

курильщики болеют на 75.25% реже

Курение

$$\begin{tikzcd}
 & U \text{ (генотип)} \arrow[dl] \arrow[dr] & \\
X \text{ (курение)} \arrow[r] & Z \text{ (смола)} \arrow[r] & Y \text{ (рак)}
\end{tikzcd}$$

смола	курильщики	некурящие
нет рака	323	1
есть рак	57	19
	15%	95%

курильщики болеют на 80% реже

нет смолы	курильщики	некурящие
нет рака	18	38
есть рак	2	342
	10%	90%

курильщики болеют на 80% реже

Курить полезно?

Курение

У курильщиков смола в 95% случаев вместо 5%; у курильщиков смола увеличивает риск рака с 10% до 15%; у некурящих — с 90% до 95%. Курить вредно?

Поможет граф!

Курение

Поправочная формула (КЗД для пустого множества и для X):

$$\mathbf{P}(Z = z \,|\, do\,(X = x)) = \mathbf{P}(Z = z \,|\, X = x)\,,$$
$$\mathbf{P}(Y = y \,|\, do\,(Z = z)) = \sum_{x'} \mathbf{P}\left(Y = y \,\middle|\, Z = z, X = x'\right) \mathbf{P}\left(X = x'\right)$$

$$\begin{array}{ccc} & U \text{ (генотип)} & \\ \swarrow & & \searrow \\ X \text{ (курение)} \longrightarrow & Z \text{ (смола)} & \longrightarrow Y \text{ (рак)} \end{array}$$

$$\mathbf{P}(Y = y \,|\, do\,(X = x)) =$$
$$= \sum_{z} \mathbf{P}(Y = y \,|\, do\,(Z = z))\, \mathbf{P}(Z = z \,|\, do\,(X = x)) =$$
$$= \sum_{z} \sum_{x'} \mathbf{P}\left(Y = y \,\middle|\, Z = z, X = x'\right) \mathbf{P}(Z = z \,|\, X = x)\, \mathbf{P}\left(X = x'\right).$$

Критерий передней двери (КПД)

Для упорядоченной пары вершин (X, Y) в ациклическом графе G множество вершин Z удовлетворяет **критерию передней двери**, если:

- Z перекрывает все направленные пути из X в Y
- нет незакрытых путей через заднюю дверь из X в Z
- все пути через заднюю дверь из Z в Y блокируются X

Если Z удовлетворяет КПД для (X, Y), то

$$\mathbf{P}(Y = y \,|do\,(X = x)\,) =$$
$$= \sum_z \mathbf{P}(Z = z\,|X = x\,) \sum_{x'} \mathbf{P}\left(Y = y\,\middle|X = x', Z = z\right) \mathbf{P}\left(X = x'\right)$$

(формула передней двери).

Литература

- причинные графы и выводы по ним — Pearl
- восстановление графов по статическим данным — Nagarajan, глава 2
- причинность по Грейнджеру — Kirchgassner, глава 3

Kirchgassner G., Wolters J., Hassler U. *Introduction to modern time series analysis*, 2013.

Nagarajan R., Scutari M., Lebre S. *Bayesian Networks in R with Applications in Systems Biology*, 2013.

Pearl J., Glymour M., Jewell N.P. *Causal Inference in Statistics: A Primer, 2016.*