Прикладная статистика и анализ данных

Съезд XII

Построение графов причинно-следств. связей

Метод индуктивной причинности

Дано: признаки V

выборка $X_1, ..., X_n$, где $X_i = (X_{iv})_{v \in V}$

Выход: ориентированный (м.б. частично) граф $G = (V, E)$.

Процедура:

1. Для всех вершин $A, B \in V$ ищем множество вершин $S_{AB} \subset V$ (в т.ч. $S_{AB} = \emptyset$), при котором $A \perp\!\!\!\perp B \mid S_{AB}$ и $A, B \notin S_{AB}$.

   Если такого множества не существует,
   то соединяем A и B неориент. ребром

   Если существует, то скорее всего связь не прямая.

2. Если A и B не связны, но имеют общего соседа ($A - C - B$), то проверяем, верно ли что $C \in S_{AB}$?

   Если нет, то образуем коллайдр: $A \longrightarrow C \longleftarrow B$.

   Если та непрямая связь обусловлена C, то это только коллайдр.

Метод индуктивной причинности

Процедура:

3. Циклически устанавливаем направления еще ненаправл. ребрам:

   a). Если A и B связны $(A - B)$
   и из A в B есть неориентир. путь $A \longrightarrow \ldots \longrightarrow B$,
   то ориентируем ребро: $A \longrightarrow B$

   b). Если A и B не связны и при этом $A \longrightarrow C, C - B$,
   то ориентируем ребро: $C \longrightarrow B$

   Коллайдр не может по п. 2.

Если полный перебор затруднителен,
то используют различные эвристики.

Критерии качества: AIC, BIC.

Проверка условной независимости

Пусть X, Y, Z — связные выборки.

Как проверить по данным, что $X \perp\!\!\!\perp Y \mid Z$?

1. **Категориальные признаки**

   Критерий хи-квадрат для трехмерных таблиц сопряженности.

   Нужно спец. образом задать соотношения на вероятности и получить критерий из обобщенного крит. хи-квадрат.

2. **Вещественные признаки**

   Частная корреляция + критерий Стьюдента

3. **Временные ряды**

   Причинность по Грейнджеру

Частная корреляция

Определение. Пусть X, Y, Z — случайные величины.
Тогда частная корреляция X и Y при условии Z равна

$$corr(X, Y \mid Z) = \frac{corr(X, Y) - corr(X, Z)corr(Y, Z)}{\sqrt{[1 - corr^2(X, Z)]\,[1 - corr^2(Y, Z)]}}.$$

Утверждение

$$corr(X, Y \mid Z) = corr(\widetilde{X}, \widetilde{Y}),$$

где $\widetilde{X} = X - aZ,\ \widetilde{Y} = Y - bZ,$

числа a, b подобраны из условий $corr(\widetilde{X}, Z) = 0$ и $corr(\widetilde{Y}, Z) = 0$

Смысл: ч. к. пытается снять лин. зависимость от третьего признака.

Рекуррентная формула

$$corr(X, Y \mid Z, W) = \frac{corr(X, Y \mid W) - corr(X, Z \mid W)corr(Y, Z \mid W)}{\sqrt{[1 - corr^2(X, Z \mid W)]\,[1 - corr^2(Y, Z \mid W)]}}$$

Частная корреляция

Выборочный коэфф. корреляции

Получается заменой *corr* на $\widehat{\rho}$ — корреляция Пирсона.

Критерий Стьюдента

Пусть $X_1, X_2 \in \mathbb{R}, X_3 \in \mathbb{R}^d$,
причем (X_1, X_2, X_3) — нормальный вектор.

$\mathrm{H}_0 : corr(X_1, X_2 \mid X_3) = 0$

Тогда

$$T(X) = \frac{\widehat{\rho}_{X_1, X_2 \mid X_3} \sqrt{n - d - 2}}{\sqrt{1 - \widehat{\rho}^2_{X_1, X_2 \mid X_3}}} \overset{d_0}{\sim} T_{n-d-2}$$

Причинность по Грейнджеру

Пусть $x_1, ..., x_T$ и $y_1, ..., y_T$ — временные ряды.

Между ними существует причинно-следственная связь $x_t \longrightarrow y_t$
если дисперсия ошибки оптимального прогноза $\widehat{y}_{t+1}$
по $(x_1, ..., x_T, y_1, ..., y_T)$ *меньше чем только по* $(y_1, ..., y_T)$.

Ряды взаимосвязаны если $x_t \longrightarrow y_t$ *и* $y_t \longrightarrow x_t$.

Критерий

Обучаем линейную модель

$$y_t = \alpha + \sum_{j=1}^{k_1} \varphi_{1j} y_{t-j} + \sum_{j=1}^{k_2} \varphi_{2j} x_{t-j} + \varepsilon_t,$$

где k_1 и k_2 выбираются по информационному критерию.

Если $x_t \longrightarrow y_t$, то существует $j \in \{1, ..., k_2\}$ т.ч. $\varphi_{2j} \neq 0$.
Тогда проверяем гип. $H_0 : \varphi_{21} = ... = \varphi_{2k_2} = 0$
критерием Фишера (см. лин. регр.).

Причинность по Грейнджеру

Многомерный случай

Пусть $x_1, ..., x_T$ и $y_1, ..., y_T$ — временные ряды.

Пусть $z_1^{(\ell)}, ..., z_T^{(\ell)}$ — другие ряды, $\ell = 1, ..., m$

Добавляем в модель зависимость от остальных признаков

$$y_t = \alpha + \sum_{j=1}^{k_1} \varphi_{1j} y_{t-j} + \sum_{j=1}^{k_2} \varphi_{2j} x_{t-j} + \sum_{\ell=1}^{m} \sum_{j=1}^{k_{\ell+2}} \varphi_{\ell+2,j} z_{t-j}^{(\ell)} + \varepsilon_t,$$

и проверяем ту же гипотезу $H_0 : \varphi_{21} = ... = \varphi_{2k_2} = 0$.


ВСЁ!