МАТЕМАТИЧЕСКИЕ ОСНОВЫ ТЕОРИИ ПРОГНОЗИРОВАНИЯ

Лектор
Сенько Олег Валентинович

Лекция 12

Виды задач прогнозирования

Ранее нами рассматривались разнообразные средства решения задачи распознавания и задачи прогнозирования непрерывных переменных (регрессионного анализа). Однако в различных прикладных исследованиях и практической деятельности встречаются задачи, которые не могут быть адекватно решены только лишь с помощью данных средств. К числу таких задач следует отнести задачу анализа выживаемости в медицине и биологии или задачу анализа надёжности в технике.

Задачи анализа выживаемости или надёжности

Целью таких задач является восстановление вероятности того, что ожидаемое критическое событие с исследуемым объектом произойдёт не ранее произвольного момента времени. Таким критическим событием может быть отказ изделия в технике, гибель испытуемого организма в биологии или смерть пациента в медицине.

Таким образом целью анализа является вычисление функции (кривой) выживаемости $S(t) = \Pr\{T > t\}$, где T - время наступления критического события, $\Pr$ -вероятность.

Задачи анализа выживаемости или надёжности

Следует отметить, что в большинстве практических исследованиях важно не только вычислить кривую выживаемости, но и оценить влияние на неё переменных, характеризующих исследуемые объекты. Такими переменными могут быть, например, возраст пациента и различные клинические показатели в биомедицинских исследованиях, или параметры, характеризующие условия изготовления изделия, в задачах анализа надёжности.

Задачи анализа выживаемости или надёжности

Задача расчёта кривых выживаемости и оценки влияния на них различных переменных может быть решена с помощью методов моделирования по эмпирическим данным.

Методы анализа выживаемости по эмпирическим данным тесно связаны с цензурированностью информации. Наблюдение в статистике считается цензурированным, если известно не точное значение наблюдаемой величины, а только интервал, которому оно принадлежит. Данный интервал может быть как конечным, так и бесконечным (ограниченным с одной стороны).

Задачи анализа выживаемости или надёжности

В данных, связанных с анализом выживаемости или надёжности нередко цензурированной оказывается информация о наступлении критического события. Например, в анализируемой выборке может содержаться информация не только об объектах, для которых критическое событие уже наступило, и момент этого события был точно зафиксирован, но также и об объектах, для которых критическое событие на момент последнего наблюдения не произошло.

Задачи анализа выживаемости или надёжности

Выборки данных в задачах анализа выживаемости обычно имеют вид $\tilde{\mathbf{S}} = \{s_1 = (\alpha_1, t_1, \mathbf{x}_1), \ldots, s_m = (\alpha_m, t_m, \mathbf{x}_m)\}$, где t_i - время, прошедшее от начального момента (например, момент изготовления изделия) до момента последнего наблюдения за объектом, α_i - индикатор, равный 1, если в момент t_i для объекта s_i было зафиксировано критическое событие, и равный 0, если в момент t_i критическое событие не наступило, $\mathbf{x}_i$ - вектор переменных $X_1, \ldots, X_n$, которые потенциально могут оказывать влияние на форму кривой выживаемости, $i = 1, \ldots, m$.

Задачи анализа выживаемости или надёжности

Рассмотрим методы восстановления кривых выживаемости при игнорировании влияния на их форму переменных $X_1, \ldots, X_n$.

Одним из наиболее популярных методов восстановления кривых выживаемости в этих случаях является процедура Каплан-Майера, учитывающая существование цензурированных наблюдений. При отсутствии цензурирования процедура Каплан-Майера совпадает со стандартным способом оценки значения кривой выживаемости в точке t как доли объектов, для которых критическое событие произошло после момента t.

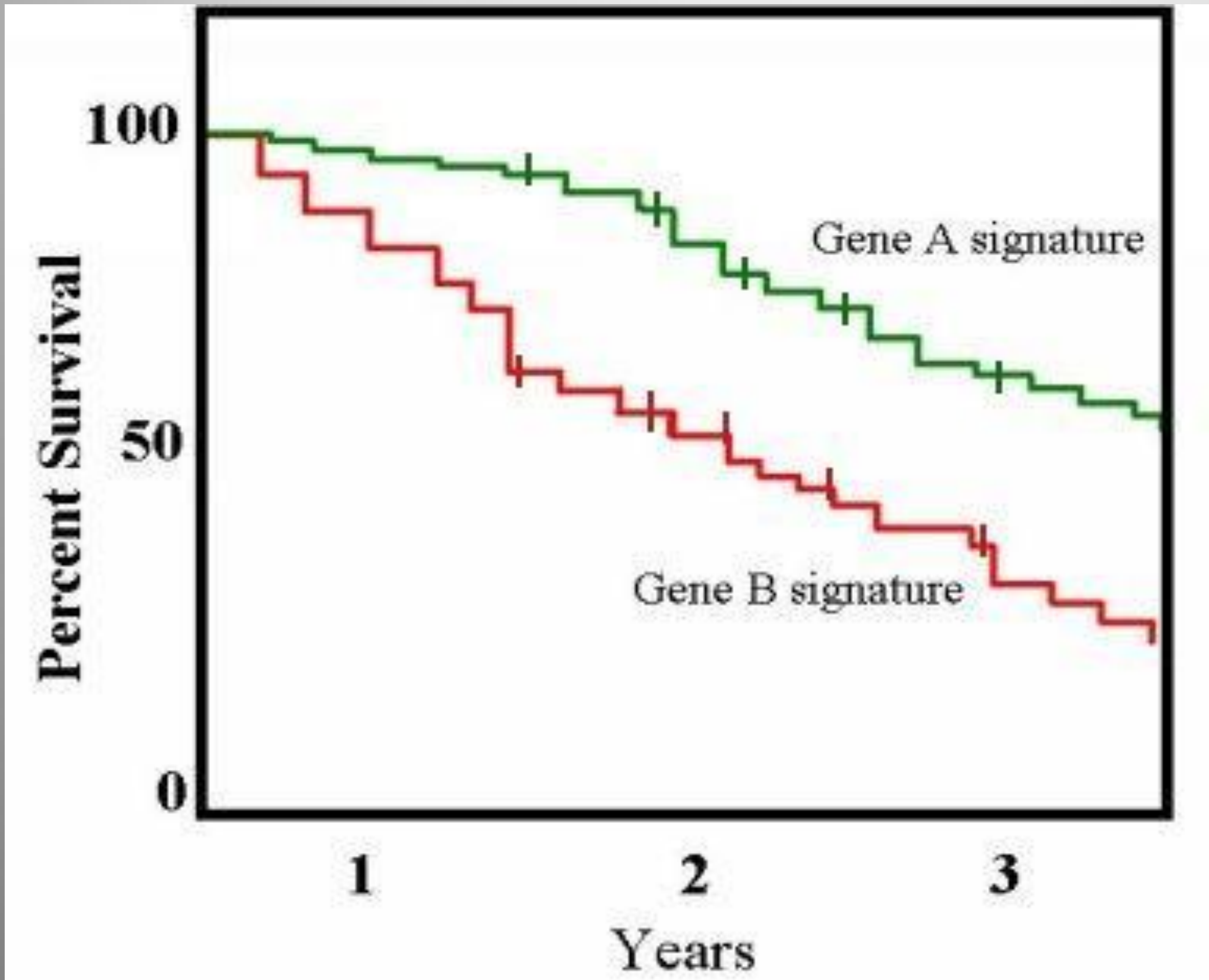
Задачи анализа выживаемости или надёжности

Предположим, что наблюдения в некоторой выборке $\tilde{\mathbf{S}}$ фиксировались в моменты $t_1, \ldots, t_N$. Пусть n_i - число объектов, для которых критический момент не наступил до момента времени t_i, d_i - число критических событий в момент t_i.

Оценка значения кривой выживаемости мо методу Каплан-Майера на полуинтервале $(t_i, t_{i+1}]$ вычисляется по формуле.

$$S(t) = \prod_{j=1}^{i} \frac{n_j - d_j}{n_j}$$

Задачи анализа выживаемости или надёжности



На рисунке представлены примеры оценок кривых выживаемости по методу Каплан-Майера для двух групп пациентов с двумя вариантами генотипа.

Задачи анализа выживаемости или надёжности

В настоящее время существует целый ряд методов оценки влияния переменных $X_1, \ldots, X_n$ на форму кривой выживаемости.

Одной из популярных моделей до сих пор является модель Кокса, основанная на концепции мгновенного риска.

Мгновенный риск в момент t определяется как предел

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{\Pr(T \le (t + \Delta t) \,|\, T \ge t)}{\Delta t} = \frac{f(t)}{S(t)}$$

Задачи анализа выживаемости или надёжности

$f(t)$- плотность вероятности наступления критического события в точке t, где $f(t) = \frac{dF(t)}{dt}$, $F(t) = 1 - S(t)$

То есть $\lambda(t)dt = \frac{-dS(t)}{S(t)}$. Откуда $\ln S(t) = -\Lambda(t)$

Или $S(t) = \exp[-\Lambda(t)]$, где $\Lambda(t) = \int_{t_0}^{t} \lambda(t)dt$, t_0 - момент начального отсчёта, который может быть принят равным 0.

Задачи анализа выживаемости или надёжности

В случае если форма кривой выживаемости зависит от переменных $X_1, \ldots, X_n$, мгновенный риск также оказывается функцией $X_1, \ldots, X_n$

В основе модели Кокса (модели пропорциональных рисков) лежит предположение о возможности представления мгновенного риска для произвольного объекта s_* с описанием $\mathbf{x}_* = (x_1^*, \ldots, x_n^*)$ в виде произведения

$$\lambda(t \mid \mathbf{x}_*) = \lambda_0(t)\exp(\beta_1 x_1^* + \ldots + \beta_n x_n^*)$$

, где $\lambda_0(t)$ - базовая компонента, зависящая только от времени.

Задачи анализа выживаемости или надёжности

Пусть $S_0(t) = \exp[-\Lambda_0(t)]$, где $\Lambda_0(t) = \int\limits_{t_0}^{t} \lambda_0(t) dt$. Откуда

следует, что $S(t) = S_0(t)^{\exp(\beta_1 x_1^* + \ldots + \beta_n x_n^*)}$

Для поиска параметров $(\beta_1, \ldots, \beta_n)$ используется метод максимального правдоподобия. Предположим, что для настройки модели пропорциональных рисков используется обучающая выборка

$$\tilde{\mathbf{S}}_t = \{s_1 = (\alpha_1, t_1, \mathbf{x}_1), \ldots, s_m = (\alpha_m, t_m, \mathbf{x}_m)\}$$

Задачи анализа выживаемости или надёжности

Предположим, что критическое событие для объекта s_i произошло в момент времени t_i. Вероятность того, что среди всех объектов, для которых критическое событие до момента t_i не наступало, это событие в момент t_i произошло именно с s_i оценим с помощью отношения

$$\frac{\lambda(t_i \mid \mathbf{x}_i)}{\sum_{t_j > t_i} \lambda(t_i \mid \mathbf{x}_j)} = \frac{\lambda_0(t_i)\exp(\beta_1 x_{i1} + \ldots + \beta_n x_{in})}{\sum_{t_j > t_i} \lambda_0(t_i)\exp(\beta_1 x_{j1} + \ldots + \beta_n x_{jn})} =$$

$$= \frac{\exp(\beta_1 x_{i1} + \ldots + \beta_n x_{in})}{\sum_{t_j > t_i} \exp(\beta_1 x_{j1} + \ldots + \beta_n x_{jn})}$$

Задачи анализа выживаемости или надёжности

Функционал правдоподобия записывается в виде

$$L(\beta_1,\ldots,\beta_n)=\prod_{i=1}^{m}\left\{\frac{\exp(\beta_1 x_{i1}+\ldots+\beta_n x_{in})}{\sum\limits_{t_j>t_i}\exp(\beta_1 x_{j1}+\ldots+\beta_n x_{jn})}\right\}^{\alpha_i}.$$

В модели используются значения $(\beta_1,\ldots,\beta_n)$, при которых

достигает максимума $L(\beta_1,\ldots,\beta_n)$.

Задачи анализа выживаемости или надёжности

Наряду со значением параметров $(\beta_1,\ldots,\beta_n)$ неизвестным параметром модели пропорциональных рисков является форма базовой функции выживаемости $S_0(t)$. Одним из возможных

походов является аппроксимация для произвольного момента времени t_i, для которого имело место критическое событие, отношения $\dfrac{S(t_i \mid \beta_1,\ldots,\beta_n,\mathbf{x}_i)}{S(t_{i-1} \mid \beta_1,\ldots,\beta_n,\mathbf{x}_i)}$ величиной

$$1-\frac{\exp(\beta_1 x_{i1}+\ldots+\beta_n x_{in})}{\sum_{t_j>t_i}\exp(\beta_1 x_{j1}+\ldots+\beta_n x_{jn})}, \qquad (1)$$

где t_{i-1} - предыдущий момент.

Задачи анализа выживаемости или надёжности

Предполагается, что параметры $(\beta_1, \ldots, \beta_n)$ уже найдены с помощью метода максимального правдоподобия.

Очевидно, что

$$\frac{S(t_i \mid \beta_1, \ldots, \beta_n, \mathbf{x}_i)}{S(t_{i-1} \mid \beta_1, \ldots, \beta_n, \mathbf{x}_i)} = \left\{ \frac{S_0(t_i)}{S_0(t_{i-1})} \right\}^{\exp\{\beta_1 x_{i1} + \ldots + \beta_n x_{in}\}} \qquad (2)$$

Обозначим $\dfrac{S_0(t_i)}{S_0(t_{i-1})}$ через γ_i

Задачи анализа выживаемости или надёжности

Из (1) и (2) следует , что

$$\gamma_i = \left\{1 - \frac{\exp(\beta_1 x_{i1} + \ldots + \beta_n x_{in})}{\sum\limits_{t_j > t_i} \exp(\beta_1 x_{j1} + \ldots + \beta_n x_{jn})}\right\}^{\{\exp(\beta_1 x_{i1} + \ldots + \beta_n x_{in})\}^{-1}}$$

Оценка базовой функции выживаемости на полуинтервале $(t_i, t_{i+1}]$

вычисляется по формуле $S_0(t) = \prod\limits_{j=1}^{i} \gamma_j$

Временные ряды

Под временным рядом понимается множество значений неко-торой переменной, измеренных в моменты времени, разделённые одинаковыми интервалами.

$$\ldots, Z(t_{i-1}), Z(t_i), Z(t_{i+1}), \ldots$$

Временной ряд считается многомерным, если в каждый момент времени измеряются значения нескольких переменных

$$\ldots, Z_1(t_{i-1}), Z_1(t_i), Z_1(t_{i+1}), \ldots$$

$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$

$$\ldots, Z_k(t_{i-1}), Z_k(t_i), Z_k(t_{i+1}), \ldots$$

Временные ряды

Основной задачей анализа временных рядов является поиск алгоритма, позволяющего предсказывать значения переменной Z или значения переменных из некоторого подмножества $\{Z_1, \ldots, Z_k\}$ в ещё не наступившие моменты времени. Дополнительными задачами анализ временных рядов является поиск существующих эмпирических закономерностей, включая поиск циклических изменений переменных.

Прогнозирование временного ряда производится с помощью алгоритма, обученного по доступному в результате наблюдений участку временного ряда достаточной длины.

Временные ряды

Одним из способов прогнозирования временных рядов является использование одномерной регрессионной функции $f(t)$, зависящей от времени. В тех случаях, когда прогностическая способность $f(t)$ является статистически достоверной, а функция $f(t)$ является линейной, говорят о наличии во временном ряду тренда.

Значения переменной Z в различных точках временного ряда

$$\ldots, Z(t_{i-1}), Z(t_i), Z(t_{i+1}), \ldots$$

могут рассматриваться как реализации случайных функций

$$\ldots, \breve{Z}_{i-1}, \breve{Z}_i, \breve{Z}_{i+1}, \ldots$$

.

Временные ряды

Процесс, отображаемый временным рядом, называется стационарным, если совместное распределение вероятности для произвольных r случайных величин $\breve{z}_{i+1}, \breve{z}_{i+2}, \ldots, \breve{z}_{i+r}$

Совпадает с совместным распределением r случайных величин $\breve{z}_{i+1+l}, \breve{z}_{i+2+l}, \ldots, \breve{z}_{i+r+l}$ при произвольном целом l. Очевидно, что процесс является стационарным, если переменные $\ldots, \breve{z}_{i-1}, \breve{z}_i, \breve{z}_{i+1}, \ldots$ являются независимыми и одинаково распределёнными.

Временные ряды

Предположим, что функция $f(t)$ полностью характеризует процесс. Это означает, что $Z(t_i) = f(t_i) + \varepsilon_i$, где

$\ldots, \varepsilon_{i-1}, \varepsilon_i, \varepsilon_{i+1}, \ldots$ - независимые и одинаково распределённые ошибки. Тогда случайный процесс, отображаемый временным рядом,

$$\ldots,[Z(t_{i-1}) - f(t_{i-1})],[Z(t_i) - f(t_i)],[Z(t_{i+1}) - f(t_{i+1})],\ldots$$

оказывается стационарным.

Временные ряды

Другим способом прогнозирования временного ряда в произвольной точке t_i является использование алгоритма A, вычисляющего оценку переменной Z по набору предшествующих значений - $[Z(t_{i-j_1}),\ldots,Z(t_{i-j_n})]$

То есть $\hat{Z}(t_i) = A[Z(t_{i-j_1}),\ldots,Z(t_{i-j_n})]$, где $(j_1,\ldots,j_n)$- натуральные числа

Временные ряды

Простейшим примером такого рода прогнозирования является

метод скользящего среднего, вычисляющего оценку Z в виде

$$\hat{Z}(t_i) = \tfrac{1}{n}\sum_{j=1}^{n} Z(t_{i-j})$$

А также метод взвешенного скользящего среднего

,

$$\hat{Z}(t_i) = \tfrac{1}{n}\sum_{j=1}^{n} c_j Z(t_{i-j})$$

где

$$\sum_{j=1}^{n} c_j = 1,\ c_j \geq 0, \quad j = 1, \ldots, n$$

Временные ряды

Нетрудно видеть, что прогностическая способность метода скользящего связана с относительным постоянство математического ожидания случайных величин $\breve{z}_{i-n}, \ldots, \breve{z}_i$.

Метод скользящего среднего используется для “сглаживания” временных рядов, фильтрации высокочастотной шумовой составляющей.

В общем случае для обучения алгоритма A могут быть использованы всевозможные методы регрессионного анализа и распознавания, если переменная Z категориальная.

Временные ряды

При этом обучение может производится по таблице, составленный из элементов, принадлежащих известному участку временного ряда. Предположим, что в результате наблюдений стали известны значения

$$Z(t_1), Z(t_2), \ldots\ldots, Z(t_N)$$

По данному ряду может быть построена таблица

$$Z(t_N), Z(t_{N-1}), \ldots\ldots, Z(t_{N-n})$$
$$Z(t_{N-1}), Z(t_{N-1}), \ldots\ldots, Z(t_{N-n-1})$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Z(t_{n+1}), Z(t_n), \ldots\ldots, Z(t_1)$$

Временные ряды

При этом первый слева элемент в каждой строке рассматривается в качестве прогнозируемой величины Y. Далее последовательно слева направо значения переменной Z в строке рассматриваются в качестве значений прогнозирующих переменных $X_1, \ldots, X_n$.

В случае многомерных временных рядов при прогнозировании некоторой переменной Z_j могут быть использованы значения и других переменных из набора .

$\{Z_1, \ldots, Z_k\}$

Временные ряды

Для поиска циклических (сезонных) колебаний переменной Z могут быть использованы методы корреляционного анализа. Для каждой предполагаемой длины цикла l строится таблица, состоящая из двух столбцов:

$$Z(t_N), Z(t_{N-l})$$
$$Z(t_{N-1}), Z(t_{N-1-l})$$
$$\ldots\ldots\ldots\ldots\ldots\ldots\ldots$$
$$Z(t_{l+1}), Z(t_1)$$

Вычисляется коэффициент корреляции между столбцами

Временные ряды

Реально существующему циклу длины l^* соответствует максимальная величина коэффициента корреляции для таблицы, построенной по сдвигу l^*, по отношению к коэффициентам корреляции для таблиц, построенным исходя из других величин сдвига.