Лекция 12

Байесовские сети
Методы анализа выживаемости

Лектор – *Сенько Олег Валентинович*

Курс «Математические основы теории прогнозирования»
4-й курс, III поток

Содержание лекции

Байесовские сети

Рассмотренные ранее в курсе методы позволяют прогнозировать значения отдельных целевых переменных. Однако более высокая точность прогноза для сложных технических, биологических или социальных систем может быть достигнута на основе описания взаимодействия наборов переменных, характеризующих данные системы. Подробное наглядное описание взаимодействия больших наборов переменных может быть достигнуто с использованием современных графических вероятностных моделей. К числу подобных моделей относятся байсовские сети (БС), сочетающие графическую наглядность с математической строгостью. Аппарат байесовских сетей принципиально позволяет полностью охарактеризовать многомерное совместное распределение больших наборов переменных.

Определение 1. Байесовской сетью называется ориентированный ациклический граф, вершинам которого поставлены в соответствие случайные переменные.

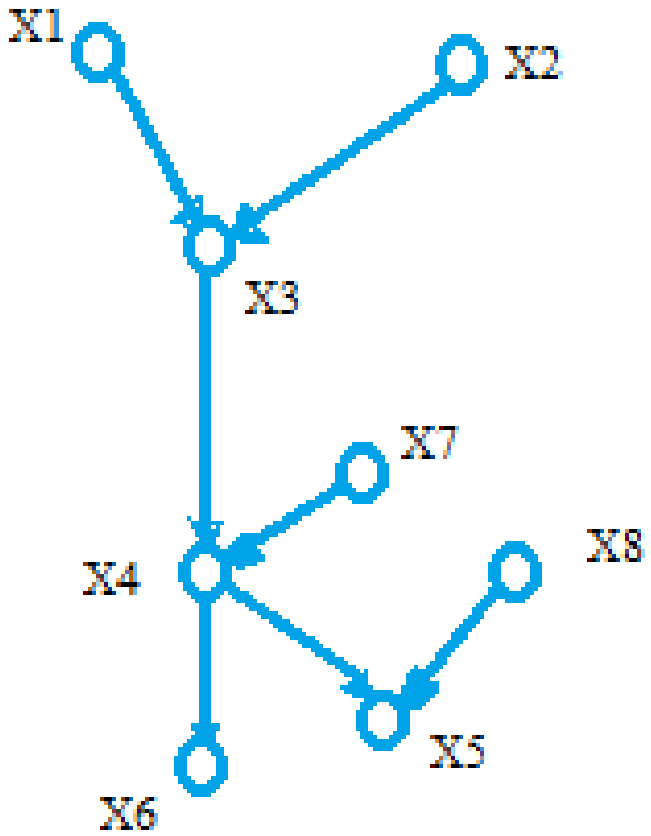

Рис.1. Пример байесовской сети.

При этом наличие ребра между двумя вершинами указывает на наличие статистической связи между соответствующими переменными. Иногда направление ребра интерпретируют как наличие причинно-следственной связи между соответствующими переменными. Для того, чтобы более точно охарактеризовать смысловую связь графической структуры БС с совместными распределениями наборов переменных, введём дополнительные определения. Вершина X_i называется предком вершины X_j, если они соединены ребром, ориентированным от X_i к X_j. Соответственно вершина X_j считается потомком вершины X_i. Вершины, не имеющие предком, называются корневыми. Обозначим через $Par(X)$ – множество предков вершины X, $V^i(X)$ - множество вершин БС, не являющихся потомками вершины X и не содержащее вершину X. Вершина называется корневой, если у неё нет вершин предков. Пример БС, описывающей взаимосвязь переменных $X_1, \ldots, X_8$, приведён на рисунке 1. Вершины, соответствующие переменным X_1, X_2, X_7, X_8 являются корневыми. Вершины, соответствующие переменным X_1 и X_2 являются предками вершины, соответствующей переменной X_3.

Вершины, соответстющие переменным X_3 и X_7, являются предками вершины, соответствующей переменной X_4 и т.д.

Условие 1. Байесовская сеть строится исходя из требования об условной независимости каждой вершины X от множества вершин $V^i(X)$ при известных значениях родителей из $Par(X)$.

Можно показать, что выполнение условия 1 эквивалентно справедливости разложения для совместного распределения вершин $X_1, \ldots, X_n$:

Условие 2.

$$P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P[X_i | Pa(X_i)]. \quad (1)$$

Из условия 2 видно, что совместная вероятность $P(X_1, \ldots, X_n)$ может быть описана с помощью n условных распределений $P[X_i|Par(X_i)]$.

Предположим, что переменные $X_1, \ldots, X_n$ являются категориальными. Тогда о параметры распределений $P[X_i|Par(X_i)]$ задаются с помощью таблицы условных вероятностей (ТУВ).

Ячейки ТУВ, соответствующей вершине (переменной) X_i, содержат вероятности каждого из возможных значений X_i при всевозможных комбинациях значений узлов, являющихся родителями X_i. В случае, когда БС является разреженной, (т.е. когда число связей между вершинами оказывается существенно ниже общего числа парных сочетаний вершин), суммарный объём ТУВ оказывается существенно меньше общего числа всевозможных комбинаций значений переменных$(X_1, \ldots, X_n)$. Построении БС производится по обучающей выборке, содержащей значения векторов переменных $X_1, \ldots, X_n$, измеренные, например, в различные моменты времени. Для оценки условной независимости используются статистические тесты. Обучающие выборки используются для вычисления ТУВ. Вместе с тем задание общего каркаса БС часто производится экспертом в области знаний,где используется БС. Построенная БС может быть использована для решения нескольких типов задач. В первую очередь необходимо отметить задачу вероятностного вывода.

Целью вероятностного вывода является оценка вероятности состояний каждой из вершин сети, исходя из известных значений переменных, соответствующих корневым вершинам. Для расчётов может быть использована представление совместной вероятности условия 2. Предположим, что переменные $X_1, \ldots, X_8$ являются бинарнымии, принимающими значения из множества $\{0, 1\}$. Рассчитаем вероятность $X_6 = 1$ при следующих условиях, наложенных на корневые переменных: $X_1 = 0$,$X_2 = 0$,$X_7 = 1$,$X_8 = 1$. Для этого достаточно вычислить, использую формулу (1) вероятность каждой из комбинаций значений переменных вида $(0, 0, u_3, u_4, u_5, u_6, 1, 1)$, где u_3, u_4, u_5, u_6 выбираются из множества $\{0, 1\}$. Обозначим множество таких комбинаций через $\widetilde{U}_f$. Разобъём множество $\widetilde{U}_f$ на подмножества $\widetilde{U}_0$, включающие комбинации из $\widetilde{U}_f$ с $u_6 = 0$, и $\widetilde{U}_1$. , включающие комбинации из $\widetilde{U}_f$ с $u_6 = 1$.

Очевидно, что вероятность множества комбинаций $\widetilde{U}$ является суммой вероятностей комбинаций, входящих в $\widetilde{U}$. Вероятность $X_6 = 1$ при условии $X_1 = 0, X_2 = 0, X_7 = 1, X_8 = 1$ очевидно равна

$$P\{\widetilde{U}_1|\widetilde{U}_f\} = \frac{P(\widetilde{U}_1)}{P(\widetilde{U}_f)}.$$

При моделировании с помощью БС сложных технических систем корневым вершинам соответствуют переменные, характеризующие внешние воздействия, или управляющие параметры. Процедура вероятностного вывода позволяет оценить распределения вероятностей для переменных, характеризующих возникновение нарушений функционирования технической системы при заданных внешних воздействиях в зависимости от значений набора управляющих параметров.

Анализ выживаемости

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. 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 $S(t) = P\{T > t\}$, где через T обозначено время наступления критического события, $P\{T > t\}$ обозначает вероятность того, что критическое событие произойдёт позже момента t.

Обычно момент t отсчитывается от от некоторой важной для изучаемого процесса точки. Такой точкой может быть, например, момент производства изделия или момент начала лечения. 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.

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). 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. Выборки данных в задачах анализа выживаемости обычно имеют вид

$$\widetilde{S} = \{s_1 = (\alpha_1, t_1, \boldsymbol{x}_1), \ldots, s_m = (\alpha_m, t_m, \boldsymbol{x}_m)\},$$

где t_i - время, прошедшее от начального момента 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 , если в момент критическое событие не наступило;
$\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$- вектор переменных $X_1, \ldots, X_n$, которые потенциально могут оказывать влияние на форму кривой выживаемости

Анализ выживаемости

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 таких наблюдений процедура Каплан-Майера эквивалентна вычислению обычных эмпирических наблюдений. Предположим, что наблюдения в некоторой выборке $\widetilde{S}$ 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

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

На рисунке 1 представлены примеры оценок кривых выживаемости по методу Каплан-Майера.

Анализ выживаемости

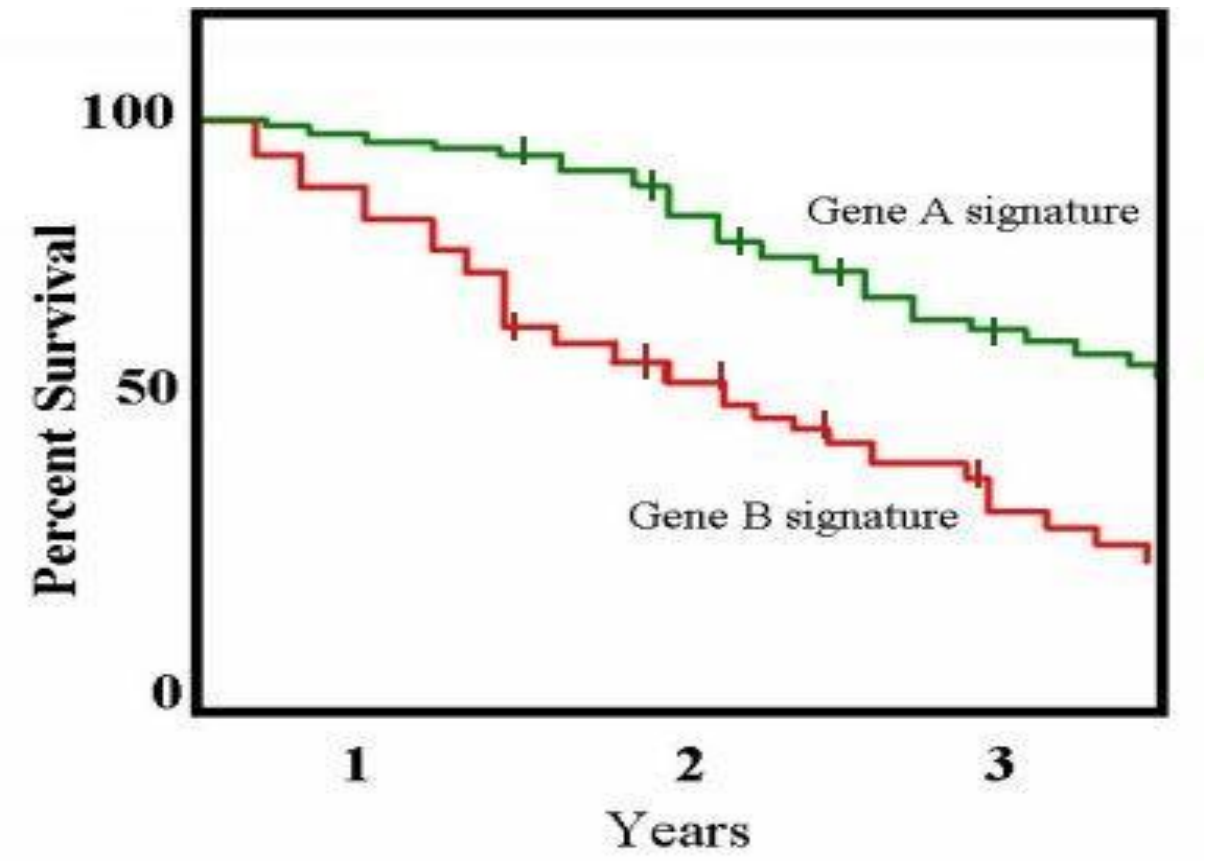

Рис. 1. Сравниваются оценки для кривых выживаемости по методу Каплан-Майера групп пациентов с двумя вариантами генотипа.

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 $\lambda(t)$ в момент t определяется как предел

$$\lim_{\Delta t \to 0} = \frac{P[T \leq (t + \Delta t) | T \geq t]}{\Delta t} = \frac{f(t)}{S(t)},$$

где $f(t)$ плотностью вероятности наступления критического события в точке t. То есть $f(t) = \frac{dF(t)}{dt}$, где $F(t) = 1 - S(t)$. Таким образом очевидна справедливость простого дифференциального уравнения

$$\lambda(t)dt = -\frac{-dS(t)}{S(t)}. \tag{2}$$

Проинтегрировав левую и правую части уравнения (1) на отрезке $[t_0, t]$ убеждаемся в справедливости равенств
$\ln[S(t)] = -\Lambda(t)$ или $S(t) = \exp[-\Lambda(t)]$ где $\Lambda(t) = \int_{t_0}^{t} \lambda(t)$.

В случае если форма кривой выживаемости зависит от переменных $X_1, \ldots, X_n$, мгновенный риск также оказывается функцией переменных $X_1, \ldots, X_n$. В основе модели Кокса (модели пропорциональных рисков) лежит предположение о возможности представления мгновенного риска для произвольного объекта s_* с описанием $\boldsymbol{x}_* = (x_1^*, \ldots, x_n^*)$ в виде произведения

$$\lambda(t|\boldsymbol{x}_*) = \lambda_0(t) \exp\left(\beta_1 * x_1^* + \ldots + \beta_n * x_n^*\right),$$

где $\lambda_0(t)$ - базовая компонента, зависящая только от времени. Пусть $S_0(t) = \exp[-\Lambda_0(t)]$, где $\Lambda_0(t) = \int_{t_0}^{t} \lambda_0(t)$. В результате получаем

$$S(t) = S_0(t)^{\left[\exp\left(\beta_1 * x_1^* + \ldots + \beta_n * x_n^*\right)\right]}.$$

Для поиска вектора параметров $(\beta_1, \ldots, \beta_n)$ используется метод максимального правдоподобия.

Модель пропорциональных рисков Кокса

Предположим, что для настройки модели пропорциональных рисков используется обучающая выборка $\widetilde{S} = \{s_1 = (\alpha_1, t_1, \boldsymbol{x}_1), \ldots, s_m = (\alpha_m, t_m, \boldsymbol{x}_m)\}$. 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

$$\frac{\lambda(t_i|\boldsymbol{x}_i)}{\sum_{t_j>t_i}\lambda(t_i|\boldsymbol{x}_j)} = \frac{\lambda_0(t_i)\exp\left(\beta_1 * x_{i1} + \ldots + \beta_n * x_{in}\right)}{\sum_{t_j>t_i}\lambda_0(t_i)\exp\left(\beta_1 * x_{j1} + \ldots + \beta_n * x_{jn}\right)} =$$

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

Функционал правдоподобия записывается в виде

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

В модели используются значения $(\beta_1, \ldots, \beta_n)$, при которых $L(\beta_1, \ldots, \beta_n)$ достигает максимума. 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 способов восстановления $S_0(t)$ является подход, основанный на аппроксимация отношения

$$\frac{S(t_i|\beta_1, \ldots, \beta_n, \boldsymbol{x}_i)}{S(t_{i-1}|\beta_1, \ldots, \beta_n, \boldsymbol{x}_i)}$$

величиной

$$1 - \frac{\exp\left(\beta_1 * x_{i1} + \ldots + \beta_n * x_{in}\right)}{\sum_{t_j > t_i} \exp\left(\beta_1 * x_{j1} + \ldots + \beta_n * x_{jn}\right)} \tag{3}$$

для произвольной пары последовательных моментов времени (t_{i-1}, t_i), для которых имели место критические события.

При этом предполагается, что вектор параметров $(\beta_1, \ldots, \beta_n)$ уже был ранее найден с помощью описанного ранее варианта метода максимального правдоподобия. Очевидно, что для вектора $\boldsymbol{x}_i$, описывающего объект s_i из обучающей выборки, справедливо равенство

$$\frac{S(t_i|\beta_1, \ldots, \beta_n, \boldsymbol{x}_i)}{S(t_{i-1}|\beta_1, \ldots, \beta_n, \boldsymbol{x}_i)} = [\frac{S_0(t_i)}{S_0(t_{i-1})}]^{exp(\beta_1 * x_{i1} + \ldots + \beta_n * x_{in})}. \quad (4)$$

Обозначим отношение $\frac{S_0(t_i)}{S_0(t_{i-1})}$ через γ_i. Из равенств (2) и (3) следует справедливость равенства

$$\gamma_i = [1 - \frac{\exp(\beta_1 * x_{i1} + \ldots + \beta_n * x_{in})}{\sum_{t_j > t_i} \exp(\beta_1 * x_{j1} + \ldots + \beta_n * x_{jn})}]^{[exp(\beta_1 * x_{i1} + \ldots + \beta_n * x_{in})]^{-1}}$$

Очевидно, величина γ_i может быть рассчитана для каждого объекта из выборки$\widetilde{}$.

Оценка базовой функции выживаемости на отрезке времени $[t_i, t_{i+1}]$ может оцениваться в виде произведения коэффициентов γ_i по всевозможным объектам $\widetilde{S}$, для которых критическое событие наступило до момента t_i. То есть

$$S_0(t_i) = \prod_{t_j < t_i} \gamma_j.$$

Временные ряды

Под временным рядом понимается множество значений некоторой переменной Z, 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. Многомерный ряд, содержащий значения переменных $Z_1, \ldots, Z_k$, может быть представлен в виде набора последовательностей:

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

$$\ldots, \ldots, \ldots, \ldots, \ldots, \ldots$$

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

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. 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 линейного тренда. Для поиска линейного тренда может быть использован метод простой одномерной регрессии с использованием в качестве прогнозирующей переменной X время t.

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

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

Процесс, отображаемый временным рядом, называется стационарным, если совместное распределение вероятности для произвольных r последовательно расположенных в ряду случайных величин

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

Совпадает с совместным распределением r случайных величин

$$\breve{Z}_{i+1+l}, \ldots, \breve{Z}_{i+r+l}, \ldots$$

при некотором 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.
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 с нулевым математическим ожиданием. Тогда случайный процесс, отображаемый временным рядо

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

оказывается стационарным.

Для прогнозирования временного ряда в произвольной точке t_i наряду с методами, основанными на выделении тренда, используются методы, основанные на поиске оптимального алгоритма A , вычисляющего оценку $Z(t_i)$ по набору предшествующих значений $\{Z(t_{j_1}), \ldots, Z(t_{j_l})\}$, где $(j_1, \ldots, j_l)$ является набором целых чисел. То есть оценка $\hat{Z}(t_i)$ вычисляется по формуле

$$\hat{Z}(t_i) = A[Z(t_{j_1}), \ldots, Z(t_{j_l})].$$

Простейшим примером такого рода прогнозирования является метод скользящего среднего, вычисляющего оценку $\hat{Z}(t_i)$ в виде

$$\hat{Z}(t_i) = \frac{1}{l} \sum_{j=1}^{l} Z(t_{i-j}).$$

Используется также метод взвешенного скользящего среднего, вычисляющего оценку $\hat{Z}(t_i)$ в виде

$$\hat{Z}(t_i) = \frac{1}{l}\sum_{j=1}^{l} c_j Z(t_{i-j}),$$

где $(c_1, \ldots, c_l)$ являются неотрицательными коэффициентами, удолетворяющими условию $\sum_{j=1}^{l} c_j = 1$.
Нетрудно видеть, что прогностическая способность метода скользящего связана с относительным постоянство математического ожидания случайных величин $\breve{Z}_{i-1}, \ldots, \breve{Z}_{i-l}, \ldots$. 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 прогнозируемая переменная Z является категориальной. Обучение алгоритма A может производится по таблице, составленный из элементов, принадлежащих известному участку временного ряда. Предположим, что в результате наблюдений стали известны значения $Z(t_1), \ldots, Z(t_N)$. По данному ряду может быть построена таблица

$$Z(t_N), Z(t_{N-1}), \ldots, Z(t_{N-l}),$$
$$Z(t_{N-1}), Z(t_{N-2}), \ldots, Z(t_{N-l-1}),$$
$$\ldots, \ldots, \ldots, \ldots, \ldots, \ldots,$$
$$Z(t_{N-l}), Z(t_{N-l-1}), \ldots, Z(t_{N-2l}).$$

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 $X_1, \ldots, X_l$. В случае многомерных временных рядов при прогнозировании некоторой переменной Z_j могут быть использованы значения и других переменных из набора $Z_1, \ldots, Z_k$. .

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:

$$Z(t_N), Z(t_{N-l}),$$

$$Z(t_{N-1}), Z(t_{N-l-1}),$$

$$\ldots, \ldots, \ldots$$

$$Z(t_{l+1}), Z(t_1).$$

Вычисляется коэффициента корреляции между столбцами. 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.