«Теория принятия решений»

ст. преп. каф. СС и ПД
Владимиров Сергей Александрович

Лекция 4

Методы математической статистики в задачах принятия решений

С О Д Е Р Ж А Н И Е

Литература:

1. Гмурман В. Е. Теория вероятностей и математическая статистика . Учебное пособие для вузов – Изд.. 7-е, стер. – М.: Высш. шк. 2001.
2. Вентцель Е.С., Овчаров Л.А. Теория случайных процессов и ее инженерные приложения. Учебное пособие для вузов – Изд.. 2-е, стер. – М.: Высш. шк. 2000.
3. Вентцель Е.С., Овчаров Л.А. Теория вероятностей и ее инженерные приложения. Учебное пособие для вузов – Изд.. 2-е, стер. – М.: Высш. шк. 2000.

Введение

На прошлой лекции были рассмотрены случайные факторы, их распределения и параметры оценок величин, то есть что можно получить, имея такие данные. На этой лекции мы рассмотрим вопрос — как это делать и каким математическим аппаратом производить анализ, оценку и оптимизацию данных в задачах принятия решений.

При принятии решений проблемы с оценкой получаемых данных подразделяются на три класса:

1) хорошо структурированные или количественно сформулированные данные и соответственно анализ проблемы, в которых получают численные оценки;

2) неструктурированные или качественно выраженные данные и отсюда проблемы, в которых количественные зависимости между признаками и характеристиками совершенно неизвестны;

3) слабо структурированные или смешанные данные и следовательно проблемы, содержащие как количественные, так и качественные элементы, причем последние имеют тенденцию к доминированию.

Постановка задачи и общий алгоритм анализа случайных последовательностей с использованием методов математической статистики.

Наиболее полным описанием случайной последовательности является функция распределения вероятностей ее значений и задача анализа в общем случае сводится к получению эмпирических вероятностных характеристик по доступным выборочным данным и проверке гипотез о их соответствии некоторым стандартным характеристикам, определяющим различные классы случайных последовательностей и отдельные их свойства. Часто в качестве стандартной случайной последовательности (СП) X выступает последовательность, например, с нормальным распределением $N(M_X; D_X)$ и числовыми характеристиками: M_X - математическое ожидание и D_X - дисперсия случайной последовательности.

Общий алгоритм анализа случайной последовательности с учетом вводимой стандартной случайной последовательности может включать следующие этапы.

1. Определение эмпирических вероятностных характеристик анализируемой случайной последовательности (математического ожидания, дисперсии, корреляционного момента, вероятностей событий и функции распределения вероятностей). Важно, чтобы качество полученных эмпирических оценок соответствовало выдвигаемым априорно требованиям к допустимому отклонению от истинных значений характеристик (доверительному интервалу и доверительной вероятности), а также определялось требуемым для этого размером выборки. На основе полученных характеристик могут быть установлены свойства симметрии распределения (совпадение значений среднего, моды и медианы, либо равенство значений вероятностей превышения и

не превышения среднего значения) и близости его формы к некоторому стандартному (см. лекцию 3), например, к нормальному.

2. Построение гистограммы вероятностей и восстановление эмпирического распределения случайной последовательности на основе полученных вероятностных характеристик и выдвижение гипотезы о виде распределения СП.

3. Проверка верности выдвинутой гипотезы по критериям соответствия эмпирических и аналитических вероятностных характеристик, а также определение класса и основных свойств случайной последовательности с оценкой показателей качества полученных оценок и решений.

Основные этапы анализа случайных последовательностей в предположении выполнения условия стационарности выборочных данных.

Вероятностной характеристикой θ случайной величины X, определяемой непосредственно путем эксперимента, является некоторое число - математическое ожидание, дисперсия, вероятность события $\alpha < X < \beta$. Символ θ означает истинное значение характеристики. Путем обработки результатов экспериментального исследования X получают экспериментальное значение характеристики θ, статистическую характеристику или оценку $\tilde{\theta}$ характеристики θ.

Экспериментальное исследование случайной величины *X* с целью определения $\tilde{\theta}$ - оценки (приближенного значения) θ, заключается в проведении ***N опытов*** (испытаний, наблюдений) и получении (путем соответствующих измерений) ряда значений $x_1, \ldots, x_i, \ldots, x_N$ — реализаций *X*. В результате обработки экспериментальных данных определяется $\tilde{\theta} = \Psi_\theta(x_1, \ldots, x_i, \ldots, x_N)$ как функция эксперимента.

Если провести ***еще одну серию из N опытов***, то будет получен ряд других реализаций $x_1', \dots, x_i', \dots, x_N'$ случайной величины X и другое значение $\tilde{\theta}'$ оценки искомой характеристики θ. Значение x_i случайной величины *X*, полученное в результате i- ого опыта в серии, можно рассматривать как значение случайной величины X_i а оценку $\tilde{\theta}$ - как реализацию более общей случайной величины

$$\tilde{\Theta} = \Psi_\theta(X_1, \dots, X_i, \dots, X_N) , \qquad (1)$$

Вероятностными характеристиками системы двух случайных величин (X,Y), определяемыми непосредственно на основании эксперимента, являются математические ожидания, дисперсии, корреляционный момент, вероятность события $\alpha_x < X < \beta_x, \alpha_y < Y < \beta_y$. Эксперимент заключается в проведении N опытов и получении ряда значений $(x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)$ реализаций случайных величин *X*, *Y*. В результате обработки экспериментальных данных получается оценка

$$\tilde{\theta} = \Psi_\theta((x_1, y_1), \dots, (x_i, y_i), \dots, (x_N, y_N)) ,$$

как реализация случайной функции аналогичной (1).

$$\widetilde{\Theta}=\Psi_{\theta}\left(X_1, Y_1; \ldots, X_i, Y_i; \ldots, X_N, Y_N\right), \qquad (2)$$

Погрешность приближения оценки $\widetilde{\Theta}$ к θ равная

$$\Delta\widetilde{\Theta}=\widetilde{\Theta}-\theta, \qquad (3)$$

является, как и $\widetilde{\Theta}$, случайной величиной.

Функцию Ψ_{θ} желательно выбирать так, чтобы выполнялось три условия

1. Математическое ожидание $\Delta\widetilde{\Theta}$ равно нулю:

$$M\left[\Delta\widetilde{\Theta}\right]=0. \qquad (4)$$

2. Дисперсия $\Delta\widetilde{\Theta}$ стремится к нулю с увеличением N

$$\lim_{N\to\infty} D\left(\Delta\widetilde{\Theta}\right)=0. \qquad (5)$$

3. Дисперсия $D\left(\Delta\widetilde{\Theta}\right)$ при данной Ψ_{θ} должна быть наименьшей.

Определения

При выполнении условия (4) **оценка** $\widetilde{\Theta}$ называется ***несмещенной***,

условий (4), (5) - ***состоятельной***,

всех трех условий - ***эффективной***.

Вследствие случайного характера погрешности (3) для характеристики точности приближенного равенства $\widetilde{\theta} = \theta$ необходимо располагать вероятностью p_∂ того, что абсолютное значение погрешности не превзойдет некоторого предела

$$P\left(|\Delta\widetilde{\Theta}| \le \Delta_\theta\right) = p_\partial. \qquad (6)$$

Интервал от $\widetilde{\theta} - \Delta_\theta$ до $\theta + \Delta_\theta$, в котором с вероятностью p_∂ находится истинное значение Θ, называется *доверительным интервалом*, его границы - *доверительными границами*, а вероятность p_∂ - *доверительной вероятностью*.

Если число экспериментальных данных N достаточно велико, то погрешность (3) состоятельной оценки $\tilde{\theta}$ можно практически считать распределенной нормально с математическим ожиданием (4), дисперсией $D(\Delta\tilde{\Theta})=D(\tilde{\Theta})=D_\theta$ и средним квадратичным отклонением $\sigma(\Delta\tilde{\Theta})=\sigma(\tilde{\Theta})=\sigma_\theta=\sqrt{D_\theta}$. При этом выражение (6) имеет вид:

$$p_\partial = \Phi\left(\frac{\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}\right) - \Phi\left(\frac{-\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}\right) = 2\Phi\left(\frac{\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}\right) = 2\Phi(t), \qquad (7)$$

где $\Phi(t)$ - функция Лапласа, $t=\dfrac{\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}$.

С помощью этой формулы решается задача определения доверительной вероятности p_∂ по известным данным $\Delta_{\tilde{\theta}}, \sigma_{\tilde{\theta}}$.

Функция Лапласа $\tau=\Phi(t)$ выражает зависимость τ от t . Обратная $t=\Phi^{-1}(\tau)$ выражает зависимость t от τ . При $t=\dfrac{\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}$, $\tau=\dfrac{p_\partial}{2}$ имеем

$$\frac{\Delta_{\tilde{\theta}}}{\sigma_{\tilde{\theta}}}=\Phi^{-1}(\tau) . \qquad (8)$$

С помощью формулы (8) и обратной функции Лапласа решается задача определения доверительного интервала $\Delta_{\tilde{\theta}}$ по известным p_∂ и σ_θ и необходимого числа испытаний по известным p_∂ и $\Delta_{\tilde{\theta}}$.

При решении первой задачи согласно (8) определяется $\Delta_{\tilde{\theta}}$.При решении второй задачи согласно (8) определяется $\sigma_{\tilde{\theta}}$, а затем N.

Для проведения анализа СП обычно приводят к стандартному виду. Для случая двоичной ноль - единичной последовательности это достигается перекодировкой исходной последовательности в симметричную -1,1- ю последовательность в соответствии с правилом $x_i(k)=2x_i'(k)-1$.

Здесь $x_i(k), x_i'(k)$ - элементы стандартной и исходной последовательностей соответственно.

Алгоритмы получения эмпирических оценок числовых характеристик, вероятностей и законов распределения случайных последовательностей и анализ их качества.

Определение математического ожидания

$$\hat{M}_x=\frac{1}{N}\sum_{i=1}^{N}x_i\ ,\ \tilde{M}_x=\frac{1}{N}\sum_{i=1}^{N}X_i\ , \qquad (9)$$

где X_i - независимые случайные величины с одинаковыми $M_{x_i}=M_x$ и $D_{x_i}=D_x$.

Математическое ожидание погрешности оценки среднего равно

$$M\left[\Delta\widetilde{M}_x\right]=M\left[\widetilde{M}_x-M_x\right]=M\left[\frac{1}{N}\sum_{i}^{N}X_i\right]-M_x=\frac{1}{N}NM_x-M_x=0\ . \qquad (10)$$

Дисперсия погрешности оценки среднего равна

$$D\left(\Delta\widetilde{M}_x\right)=D\left(\widetilde{M}_x\right)=D_{\ddot{M}_x}=D\left(\frac{1}{N}\sum_{i}^{N}X_i\right)=\frac{1}{N^2}D\left(\sum_{I}^{N}X_i\right)=\frac{1}{N^2}ND_x=\frac{D_x}{N}=\frac{\sigma_x^2}{N}\ . \quad (11)$$

Среднеквадратическое отклонение оценки математического ожидания (9)

$$\sigma_{\widetilde{M}_x}=\sigma_x\sqrt{N}\approx\widetilde{\sigma}_x\sqrt{N}\ . \qquad (12)$$

Оценка (9) – *несмещенная*, *состоятельная* и *эффективная*.

Определение оценки дисперсии и ее среднеквадратического отклонения

Оценка дисперсии D_x $$D_x=\frac{1}{N}\sum_{i}^{N}\left(x_i-M_x\right)^2 .$$

Так как значение M_x априори неизвестно, то принимают $M_x \approx \widetilde{M}_x$ и тогда

$$\widetilde{D}_x=\frac{1}{N}\sum_{i}^{N}X_i^2-\left(\frac{1}{N}\sum_{i}^{N}X_i\right)^2 . \qquad (13)$$

Математическое ожидание погрешности оценки равно

$$M\left[\Delta\widetilde{D}_x\right]=M\left[\widetilde{D}_x-D_x\right]=-\frac{D_x}{N} , \qquad (14)$$

что означает, что оценка (14) является *смещенной*.

Смещение пропорционально D_x и обратно пропорционально N. Это означает, что оценка D_x, полученная согласно (14), - *состоятельная*.

Смещение устраняется с переходом к $\widetilde{D}_x'=\frac{N}{N-1}\widetilde{D}_x$.

При этом вместо (13) имеем $$\widetilde{D}_x=\frac{1}{N-1}\sum_{i}^{N}x_i^2-\frac{N}{N-1}\widetilde{M}_x^2 \quad . \qquad (15)$$

При больших значениях N результаты расчета по формулам (13) и (15) практически будут одинаковыми.

Зависимость среднего квадратического отклонения $\sigma_{\widetilde{D}_x}$ от его точного значения σ_x определяется выражением

$$\sigma_{\widetilde{D}_x} \approx \sqrt{\frac{2}{N-1}}\sigma_x .$$

Определение корреляционного момента и коэффициента корреляции

Экспериментальное значение корреляционного момента R_{xy} как оценка смешанного центрального момента m_{11} системы двух случайных величин равно

$$R_{xy}=\frac{1}{N}\sum_{1}^{N}\left(x_i-M_x\right)\left(y_i-M_y\right). \quad (16)$$

Так как значения M_x, M_y неизвестны, то принимают $M_x\approx\tilde{M}_x$, $M_y\approx\tilde{M}_y$ и тогда

$$\tilde{R}_{xy}-\frac{1}{N}\sum_{1}^{N}\left(x_i-\tilde{M}_x\right)\left(y_i-\tilde{M}_y\right)=\frac{1}{N}\sum_{1}^{N}x_i y_i-\left(\frac{1}{N}\sum_{1}^{N}x_i\right)\left(\frac{1}{N}\sum_{1}^{N}y_i\right)$$

или

$$\tilde{R}_{xy}=\frac{1}{N}\sum_{1}^{N}X_i Y_i-\left(\frac{1}{N}\sum_{1}^{N}X_i\right)\left(\frac{1}{N}\sum_{1}^{N}Y_i\right). \qquad (17)$$

Погрешность оценки $\tilde{R}_{xy}$

$$\Delta\tilde{R}_{xy}=\tilde{R}_{xy}-R_{xy} \qquad (18)$$

Математическое ожидание погрешности (18)

$$M\left[\Delta\widetilde{R}_{xy}\right]=M\left[\frac{1}{N}\sum_{1}^{N}X_iY_i\right]-M\left[\left(\frac{1}{N}\sum_{1}^{N}X_i\right)\left(\frac{1}{N}\sum_{1}^{N}Y_i\right)\right]=-\frac{R_{xy}}{N}$$

Это означает, что оценка (17) - смещена и равна

$$\widetilde{R}_{xy}=\frac{N-1}{N}R_{xy}\ . \qquad (19)$$

Можно показать, что она является и состоятельной.

Смещение устраняется с переходом от $\widetilde{R}_{xy}$ к $\widetilde{R}'_{xy}=\frac{N-1}{N}\widetilde{R}_{xy}$.

При этом вместо (17) имеем $\widetilde{R}_{xy}=\frac{1}{N-1}\sum_{1}^{N}x_i y_i-\frac{N}{N-1}\widetilde{M}_x\widetilde{M}_y$. (20)

Среднеквадратическое значение погрешности (18) равно среднему квадратическому отклонению оценки (20): $\sigma_{\widetilde{R}_{xy}}\approx\sqrt{\left(\widetilde{R}^2_{xy}+D_x D_y\right)/\left(N-1\right)}$. (23)

Оценка коэффициента корреляции определяется согласно

$$\widetilde{r}_{xy}=\widetilde{R}_{xy}/\left(\widetilde{\sigma}_x\widetilde{\sigma}_y\right)\ . \qquad (24)$$

Определение вероятности события через его повторяемость

Экспериментальное значение вероятности P некоторого события - это его повторяемость [1-3] $$W=\frac{n}{N}=\tilde{P}=\frac{1}{N}\sum_{1}^{N} x_i , \qquad (26)$$

причем число n появлений события в серии из N испытаний можно рассматривать как сумму N независимых случайных слагаемых:

$$W=\frac{1}{N}\sum_{1}^{N} X_i , \qquad (27)$$

каждое из которых может принимать только два значения 1 и 0 с вероятностями P и $1-P$.

Математическое ожидание и дисперсия случайной величины X_i:

$$M_{x_i}=P; \quad D_{x_i}=P(1-P) . \qquad (28)$$

Погрешность оценки (26) равна

$$\Delta W=W-P . (29)$$

Математическое ожидание погрешности и ее дисперсия:

$$M[\Delta W]=0; \quad D(\Delta W)=P(1-P)/N=D(W) . \qquad (30)$$

Таким образом, оценка (26) - несмещенная и состоятельная. Среднее квадратическое отклонение оценки (26)

$$\sigma_w=\sqrt{P(1-P)/N} .$$

На практике принимают

$$\sigma_w\approx\sqrt{W(1-W)/N} . \quad (31)$$

Определение законов распределения случайной величины (строим гистограмму).

Если случайная величина X - дискретная, то определяются $\widetilde{M}_x$, $\widetilde{D}_x$ и оценки $\widetilde{P}_{x_i}$ значений функции вероятности $P(x_i)$ или оценки $\widetilde{F}(x_i)$ значений функции распределения $F(x_i)$.

Если случайная величина X - непрерывная, то определяются M_x, D_x и оценки $f_x(x)$, $F_x(x)$ плотности вероятности $f_x(x)$ и функции распределения $F_x(x)$.

При оценивании законов распределения непрерывной случайной величины процесс обработки экспериментальных данных - реализаций $x,...,x_N$,, начинается с выбора границ a и $b > a$ интервала, заключающего возможные значения X, и деления этого интервала на k равных элементарных промежутков $c = (b - a)/ k$.

При расчете c значения a и b следует для удобства округлять, принимая, например, вместо b = 3,341, a = -2,63 значения 3,4 и -2,7. Во всех случаях округление производится в сторону увеличения разности b-a. Значение k выбирается в пределах от 8 до 20. Удобно принять k= 10.

После этого определяют границы $x_j^{'}$ всех элементарных промежутков и составляют таблицу (табл.1), в которой $x'_0=a$, $x'_k=b$. Значение $\widetilde{n}_j$ - это число реализаций X, оказавшихся в пределах j-ого интервала от $x_{j-1}^{'}$, до $x_j^{'}$. Значения $\widetilde{P}_j$ и $\widetilde{F}_j$:

$$\widetilde{P}_j=\widetilde{n}_j/N\approx P\left(x_{j-1}^{'}<X<x_j^{'}\right); \quad (32)$$

$$\widetilde{F}_j=\widetilde{P}_1+\ldots+\widetilde{P}_{j-1}=P\left(X<\widetilde{x}_j\right). \quad (33)$$

При группировке реализаций X по отдельным интервалам может оказаться что некоторые из них придутся точно на границу двух смежных промежутков. В этих случаях необходимо прибавить к числам $\widetilde{n}_j$ и $\widetilde{n}_{j+2}$ смежных интервалов по 1/2.

Таблица 1

x_i	x'_0	x'_1	x'_2	$\ldots$ x'_{k-1}	x'_k	
n_j		$\widetilde{n}_1$	$\widetilde{n}_2$	$\ldots$	$\widetilde{n}_k$	
$\widetilde{P}_j$		$\widetilde{P}_1$	$\widetilde{P}_2$	$\ldots$	$\widetilde{P}_k$	
$\widetilde{F}_j$		$\widetilde{F}_1$	$\widetilde{F}_2$	$\ldots$	$\widetilde{F}_k$	

По данным таблицы могут быть построены эмпирические гистограмма и график функции распределения.

Затем возникает весьма сложная задача подбора аналитического закона распределения, достаточно хорошо согласующегося с результатами эксперимента. Основанием для выбора аналитического выражения плотности вероятности $f_x(x)$ могут служить соображения о том, чтобы простейшие числовые характеристики теоретической случайной величины были равны экспериментальным значениям этих характеристик. Если, например, теоретический закон определяется двумя параметрами, то их выбирают так, чтобы совпали два момента (m_1, μ_2).

Для оценки существенности или несущественности расхождения между теоретическим и эмпирическим распределениями используют различные критерии

Критерий интервальных оценок

Располагая результатами эксперимента согласно (31) рассчитывают средние квадратические отклонения: $\sigma_{\tilde{P}_i}=\sqrt{\tilde{P}_j(1-\tilde{P}_j)N}$; $\sigma_{\tilde{F}_j}=\sqrt{\tilde{F}_j(1-\tilde{F}_j)N}$. (34)

Согласно (8) рассчитываются доверительные интервалы $\Delta_{\tilde{P}_j}=3\sigma_{\tilde{P}_j};\ \Delta_{\tilde{F}_j}=3\sigma_{\tilde{F}_j}$

и границы изменения ВВХ $\tilde{P}_j\pm\Delta_{\tilde{P}_j};\tilde{F}_j\pm\Delta_{\tilde{F}_j}$, (35)

соответствующие доверительной вероятности $p_\partial=0{,}9972$ и $\Phi^{-1}(0{,}4986)=3$.

Располагая выбранным аналитическим выражением плотности вероятности

$f_x(x)$, рассчитываются теоретические значения: $$P_j=P(x'_{j=1}<X<x'_j)=\int_{x'_{j=1}}^{x'_j} f_x(x)\,dx\,;\qquad F_j=P(x'_{j=1}<X<x'_j)=\int_{x'_{j=1}}^{x'_j} f_x(x)\,dx$$ (36)

Критерием согласия теоретического и экспериментального распределения

является соблюдение неравенств:
$$\begin{aligned} &\tilde{P}_j - \Delta_{\tilde{P}_j} < P_j < \tilde{P}_j + \Delta_{\tilde{P}_j}; \\ &\tilde{F}_j - \Delta_{\tilde{F}_j} < F_j < \tilde{F}_j + \Delta_{\tilde{F}_j}. \end{aligned} \qquad (37)$$

Критерий χ^2

Рассчитав P_j согласно (35), находят значения $$n_j = NP_j \qquad (38)$$

и рассчитывают $$\chi^2 = \sum_{1}^{k} (\tilde{n}_j - n_j)^2 / n_j\ . \quad (39)$$

Если расхождение между экспериментальным и теоретическим распределением несущественно, то распределение случайной величины (39) близко к нормальному с математическим ожиданием $M_{\chi^2} = s$ и средним квадратическим отклонением $\sigma_{\chi^2} = \sqrt{2s}$, где s - так называемое *число степеней свободы* и согласно (8) с доверительной вероятностью p_∂ = 0,997 справедливо неравенство

$$(\chi^2 - s) / \sqrt{2s} < 3\ . \quad (40)$$

Число степеней свободы $s = k - u$ - это разность между числом интервалов k, выбираемых произвольно, и числом условий u, которым должно удовлетворять эмпирическое распределение случайной величины. Этих условий обычно три: сумма всех $\widetilde{P}_j$ равна единице, математическое ожидание равно $\widetilde{M}_j$ дисперсия равна $\widetilde{D}_x$.

Заключение.

В заключение отметим, что все используемые в настоящее время методы анализа случайных последовательностей не выходят за рамки представленного общего подхода, однако в некоторых случаях позволяют существенно упростить процедуры их классификации если учитывают специфику анализируемых последовательностей.