§10. Условное распределение вероятностей. Условное математическое ожидание

Лекция 16

Одной из типичных задач теории вероятностей является прогноз возможного значения случайной величины Y по результату наблюдения другой случайной величины X. Примеры таких задач – прогноз метеорологических показателей (температуры воздуха, атмосферного давления, количества осадков) по данным их замеров в прошлом; прогноз стоимости ценных бумаг или обменного курса валюты на ближайшее будущее, причем прогноз осуществляется не столько по значениям этих финансовых показателей на сегодняшний день, сколько по данным информационного центра о состоянии промышленности страны, различных показателей рынка ценных бумаг, их движения, спроса и пр. С аналогичными проблемами мы сталкиваемся в медицине, технике, управлении производством. Легко понять, что во всех этих примерах прогноз носит вероятностный характер, и должен отвечать на вопросы типа: если наблюдается значение x случайной величины X, то какова вероятность того, что Y будет лежать в определенных пределах, каково наиболее вероятное значение Y, или чему равно среднее значение Y, если X приняло значение x?

На все эти вопросы легко ответить в случае дискретных распределений случайного вектора (X, Y). Пусть $f^{X,Y}(x, y) = P(X = x, Y = y)$ – совместная функция плотности X и Y по считающей мере на борелевской плоскости R^2. Тогда решение задачи прогноза дает *условная функция плотности* – условная вероятность

$$f^{Y|X}(y \mid x) = P\{Y = y \mid X = x\} = \frac{P(X = x, Y = y)}{P(X = x)} = \frac{f^{X,Y}(x, y)}{f^X(x)},$$

где $f^X(x)$ – маргинальная функция плотности X. Действительно, вероятность того, что прогнозируемое значение Y будет лежать в пределах $(a; b)$ равно условной вероятности

$$P\{Y \in (a; b) \mid X = x\} = \sum_{y \in (a;b)} f^{Y|X}(y \mid x),$$

наиболее вероятное значение Y – это мода условного условного распределения

$$\mathrm{mod}(Y \mid X) = \arg\sup_y f^{X|Y}(y \mid x),$$

ожидаемое (среднее) значение Y – условное математическое ожидание

$$\mathbf{E}\{Y \mid X = x\} = \sum_y y f^{Y|X}(y \mid x).$$

Но как осуществить аналогичные расчеты в случае непрерывного распределения X, когда $P(X = x) = 0$ и на ноль, как известно, делить нельзя? Конечно, и в этом случае можно предложить некоторое , ,суррогатное", построенное на аналогиях с дискретным случаем, определение условного распределения, но тем не менее отсутствие строгого определения условной вероятности, пригодного для распределений любого типа, долго тормозило развитие теории вероятностей и ее становление как самостоятельной математической дисциплины. Особенно страдала от этого теория случайных процессов, до тех пор пока в 20-х годах нашего столетия Андрей Николаевич Колмогоров ввел строгое определение условного математического ожидания и условного распределения вероятностей. Увы, это слишком сложная теория, которую мы не сможем постичь с нашими скудными , ,Вэ-эМ-Ка-шинскими" познаниями в области теории меры. Замечу только, что в этом определении решающую роль играет теорема Радона–Никодима, о которой мы упоминали в §6, когда вводили функцию плотности распределения случайной величины. Кроме того, условное математическое ожидание и условное распределение трактуются как случайные величины, изменение значений которых на множествах нулевой вероятности не влияет на их распределение, и это обстоятельство позволяет исключить из рассмотрения значения x случайной величины X, в которых $f^X(x) = 0$. Сейчас мы расскажем, как это делается без привлечения теоремы Радона–Никодима, когда совместное распределение X и Y принадлежит непрерывному типу (существует плотность $f^{X,Y}(x, y)$ по мере Лебега на плоскости).

Предположим дополнительно, что совместная функция плотности $f^{X,Y}(x, y)$ непрерывна по обеим переменным, и изучим асимптотическое поведение условной вероятности $P\{Y < y \mid x \le X < x + \Delta x\}$ при $\Delta x \to 0$. Понятно, что таким образом мы пытаемся ввести условную функцию плотности Y, при условии $X = x$. По формуле условной вероятности находим

$$P\{Y < y \mid x \le X < x + \Delta x\} = \frac{P(Y < y,\, x \le X < x + \Delta x)}{P(x \le X < x + \Delta x)} =$$

$$\frac{\int\limits_{-\infty}^{y} ds \int\limits_{x}^{x+\Delta x} f^{X,Y}(s,\, t)dt}{\int\limits_{x}^{x+\Delta x} f^{X}(s)ds}.$$

Применяя теорему о среднем к интегралом в правой части этого равенства, получаем

$$P\{Y < y \,|\, x \le X < x + \Delta x\} = \frac{\int\limits_{-\infty}^{y} f^{X,Y}(s,\, x + \lambda_1 \Delta x)ds\Delta x}{f^{X}(x + \lambda_2 \Delta x)\Delta x},$$

где $0 < \lambda_i < 1$, $i = 0,\, 1$.

Теперь, используя непрерывность функций $f^{X,Y}$ и f^{X}, мы можем определить условную функцию распределения Y относительно события $X = x$, если устремим в обеих частях последнего неравенства $\Delta x \to 0$:

$$F^{Y|X}(y \,|\, x) = \lim_{\Delta x \to 0} P\{Y < y \,|\, x \le X < x + \Delta x\} = \frac{\int\limits_{-\infty}^{y} f^{X,Y}(s,\, x)ds}{f^{X}(x)}.$$

Условная функция плотности получается дифференцированием условной функции распределения:

$$f^{Y|X}(y \,|\, x) = \frac{d}{dy} F^{Y|X}(y \,|\, x) = \frac{f^{X,Y}(x,\, y)}{f^{X}(x)}.$$

Удивительно, но мы пришли к той же формуле условной плотности, что и в дискретном случае!

Поскольку значения x, в которых $f^{X}(x) = 0$, можно вообще исключить из области возможных значений случайной величины X (эти значения в совокупности дают множество нулевой вероятности), то, подставляя в определении условной плотности вместо x случайную величину X, приходим к следующему определению условного распределения и условного математического ожидания для дискретных и непрерывных распределений случайного вектора $(X,\, Y)$.

Определение 10.1. Если случайный вектор (X, Y) имеет дискретную или непрерывную совместную функцию плотности $f^{X,Y}(x,\, y)$, то *условное распределение случайной величины* Y *относительно случайной величины* X определяется функцией плотности

$$f^{Y|X}(y \,|\, X) = \frac{f^{X,Y}(y,\, X)}{f^{X}(X)},\ y \in \mathrm{R},$$

(2) Если X и Y независимы, то $f^{X,Y}(x,\,y) = f^X(x)f^Y(y)$, откуда

$$\mathbf{E}\{Y \mid X\} = \int_{\mathrm{R}} y\frac{f^X(x)f^Y(y)}{f^X(x)}d\mu(y) = \int_{\mathrm{R}} yf^Y(y)d\mu(y) = \mathbf{E}Y.$$

Теперь мы располагаем некоторой техникой для решения простейших задач прогноза.

Наилучший в среднем квадратическом прогноз ожидаемого значения Y по результату наблюдения X.

Проблема оптимального предсказания ожидаемого значения Y по результату x наблюдения случайной величины X состоит в отыскании функции $g(X)$, которая минимизирует некоторую меру уклонения $g(X)$ от Y. Мы будем решать задачу минимизации *среднего квадратического уклонения* $\mathbf{E}(Y - g(X))^2$, в котором математическое ожидание вычисляется по совместному распределению X и Y. Решение задачи дает

Предложение 10.2. *Минимум функционала* $\mathbf{E}(Y - g(X))^2$ *достигается на функции* $g^*(X) = \mathbf{E}\{Y \mid X\}$.

Доказательство. Используя свойство (1) условного математического ожидания, установленное в предложении 10.1, представим среднее квадратическое уклонение в виде $\mathbf{E}(Y - g(X))^2 = \mathbf{E}^X\mathbf{E}\{(Y - g(X))^2 \mid X\}$ и будем минимизировать момент второго порядка $\mathbf{E}\{(Y - g(X))^2 \mid X\}$ случайной величины Y относительно , ,точки" $g(X)$, когда распределение Y определяется условной функцией плотности $f^{Y|X}(y \mid X)$. В §6 (предложение 6.1, утверждение 4^0) было показано, что минимум достигается на среднем значении Y и равен $\mathbf{D}Y$. Поскольку в данном случае момент второго порядка вычислялся по условному распределению Y относительно X, то минимум достигается на функции (случайной величине) $g^*(X) = \mathbf{E}\{(Y \mid X\}$.

Итак, наилучший в среднем квадратическом прогноз ожидаемого значения Y по результату x наблюдения случайной величины X доставляет реализация в точке $X = x$ условного математического ожидания $\mathbf{E}\{Y \mid X\}$. Допуская некоторую вольность, запишем эту реализацию в виде функции $y = g^*(x) = \mathbf{E}\{Y \mid X = x\}$. Функция $y = g^*(x)$

называется *кривой средней квадратической регрессии* Y на X, а минимальное среднее квадратическое уклонение $\sigma^2_{Y|X} = \mathbf{E}(Y - g^*(X))^2$ – *остаточной дисперсией.*

Найдем кривую $g^*(x)$ и вычислим остаточную дисперсию в случае совместного нормального распределения X и Y.

Предложение 10.3. *Пусть* $(X, Y) \sim \mathcal{N}_2(\mu, \Lambda)$. *Тогда кривая средней квадратической регрессии* Y *на* X *есть прямая*

$$y = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Остаточная дисперсия $\sigma^2_{Y|X} = \sigma_2(1 - \rho^2)$.

Доказательство. В случае двумерного нормального распределения

$$f^{X,Y}(x, y) = f^{Y|X}(y \mid x) \cdot f^X(x) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \cdot$$

$$\exp\left\{-\frac{1}{2(1-\rho^2)}\left(\frac{(x-\mu_1)^2}{\sigma_1^2} - \frac{2\rho(x-\mu_1)(y-\mu_2)}{\sigma_1\sigma_2} + \frac{(y-\mu_2)^2}{\sigma_2^2}\right)\right\},$$

и остается только убедиться, что

$$f^X(x) = \frac{1}{\sqrt{2\pi}\sigma_1}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right\},$$

а

$$f^{Y|X}(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{(y-g(x)}{2\sigma_2^2(1-\rho^2)}\right\}, \tag{1}$$

где

$$g(x) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Маргинальную функцию плотности $f^X(x)$ компоненты X мы вычисляли в конце предыдущего параграфа. Перемножая $f^X(x)$ и $f^{Y|X}(y \mid x)$ и производя несложные алгебраические преобразования, убеждаемся, что это произведение равно плотности $f^{X,Y}(x, y)$ двумерного нормального распределения.

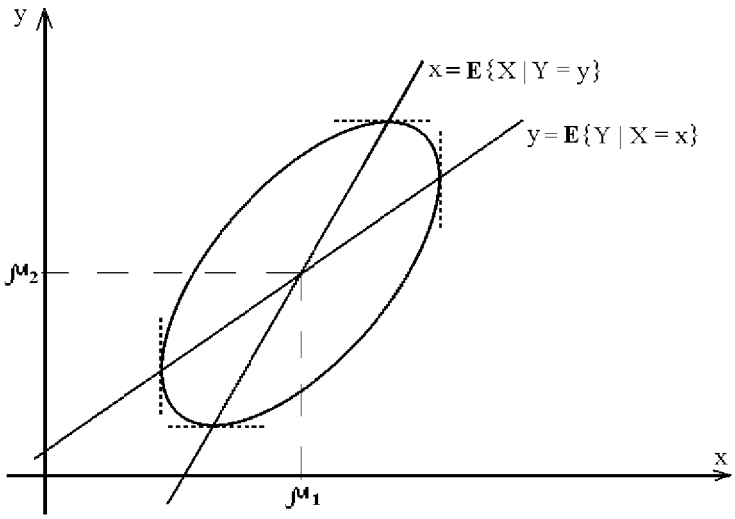
Из формулы (1) находим, что условная плотность имеет среднее $\mathbf{E}\{Y \mid X = x\} = g(x)$ и дисперсию $\sigma_2^2(1 - \rho^2)$, которые совпадают с формулами, приведенными в формулировке предложения.

Легко понять, что если проблема состоит в предсказании значения x случайной величины X по наблюдению y случайной величины Y, то наилучший в среднем квадратическом прогноз x имеет вид

$$x = \mu_1 + \rho\frac{\sigma_1}{\sigma_2}(y - \mu_2),$$

что не совпадает с функцией, обратной к

$$y = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(x - \mu_1).$$

Это вполне естественно, поскольку мы минимизируем в каждом из этих случаев уклонения от прямой регрессии в различных (перпендикулярных друг к другу) направлениях: одно – вдоль оси OY, второе – вдоль оси OX.

Ниже приводится рисунок, иллюстрирующий расположение прямых регрессии Y на X и X на Y относительно главной оси эллипса рассеивания, которая носит название *прямой ортогональной регрессии* и дает наилучший в среднем квадратическом прогноз, когда ошибки измеряются не вдоль координатных осей, а по кратчайшему расстоянию к кривой прогноза $g(x)$.