Теория вероятностей и математическая статистика

Вебинар 8

Двухфакторный дисперсионный анализ.
Факторный анализ. Логистическая регрессия

Двухфакторный дисперсионный анализ

В *однофакторном дисперсионном анализе* исследуется влияние одной категориальной переменной x, имеющей k уровней, на количественную переменную y. Проверяется нулевая гипотеза о том, что среднее значение переменной y на всех уровнях фактора x совпадает.

Двухфакторный дисперсионный анализ

В *однофакторном дисперсионном анализе* исследуется влияние одной категориальной переменной x, имеющей k уровней, на количественную переменную y. Проверяется нулевая гипотеза о том, что среднее значение переменной y на всех уровнях фактора x совпадает.

В *двухфакторном дисперсионном анализе* имеются два фактора a, b, каждый из которых является категориальным. Проверяются гипотезы о влиянии каждого фактора на значение переменной y.

Замечание. Почему здесь нельзя просто использовать два однофакторных дисперсионных анализа? Потому что в таком случае мы не учитываем тот факт, что два фактора могут зависеть друг от друга.

Если мы уверены, что два фактора независимы, то разницы нет. В противном случае применение двух однофакторных дисперсионных анализов будет менее точным, чем использование одного двухфакторного.

Двухфакторный дисперсионный анализ

Рассмотрим схему двухфакторного дисперсионного анализа *с однократными наблюдениями*. При таком подходе для каждой пары уровней факторов a и b выполняется только одно измерение переменной y.

Пусть фактор a имеет m уровней, а фактор b имеет k уровней. Тогда исходные данные можно представить в виде таблицы

$$Y = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ y_{21} & \cdots & y_{2k} \\ \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mk} \end{pmatrix}$$

где y_{ij} — наблюдение на i-м уровне фактора a и j-м уровне фактора b.

Замечание. В двухфакторном дисперсионном анализе *с многократными наблюдениями* каждый y_{ij} представлял бы собой какой-то массив из значений.

Двухфакторный дисперсионный анализ

По каждому фактору проверяется нулевая гипотеза о равенстве средних значений на каждом уровне. Пусть

- Y_{i*} — i-я строка, т.е. значения переменной y на i-м уровне фактора a и k уровнях фактора b,
- Y_{*j} — j-й столбец, т.е. значения переменной y на m уровнях фактора a и j-м уровне фактора b.

Двухфакторный дисперсионный анализ

По каждому фактору проверяется нулевая гипотеза о равенстве средних значений на каждом уровне. Пусть

- Y_{i*} — i-я строка, т.е. значения переменной y на i-м уровне фактора a и k уровнях фактора b,
- Y_{*j} — j-й столбец, т.е. значения переменной y на m уровнях фактора a и j-м уровне фактора b.

Нулевые гипотезы:

$$H_{0a}: \ \overline{Y_{1*}} = \cdots = \overline{Y_{m*}}, \ \ H_{0b}: \ \overline{Y_{*1}} = \cdots = \overline{Y_{*k}}$$

Двухфакторный дисперсионный анализ

Для вычисления значений статистик нам вновь понадобятся оценки дисперсий. Они вычисляются похожим образом. Суммы квадратов отклонений:

$$S_a^2 = k \cdot \sum_{i=1}^{m} \left(\overline{Y_{i*}} - \overline{Y}\right)^2, \;\; S_b^2 = m \cdot \sum_{j=1}^{k} \left(\overline{Y_{*j}} - \overline{Y}\right)^2, \;\; S_w^2 = \sum_{i=1}^{m} \sum_{j=1}^{k} \left(y_{ij} - \overline{Y_{i*}} - \overline{Y_{*j}} + \overline{Y}\right)^2$$

Замечание. Первая сумма — отклонения между уровнями фактора a, вторая — между уровнями фактора b, третья — внутригрупповые отклонения.

Оценки дисперсий:

$$\sigma_a^2 = \frac{S_a^2}{m-1}, \;\; \sigma_b^2 = \frac{S_b^2}{k-1}, \;\; \sigma_w^2 = \frac{S_w^2}{(k-1)(m-1)}$$

Двухфакторный дисперсионный анализ

Напомним, что в двухфакторном дисперсионном анализе мы проверяем две гипотезы (отдельная гипотеза о влиянии каждого из факторов). Итак, статистика для гипотезы о влиянии фактора a:

$$F_a = \frac{\sigma_a^2}{\sigma_w^2}$$

В предположении верности гипотезы H_{0a} эта статистика имеет распределение Фишера с параметрами $k_{1a} = m - 1$, $k_{2a} = n - m$. Далее, как обычно, строится критическая область (правосторонняя, поскольку распределение Фишера имеет один хвост), и проводится тест.

Двухфакторный дисперсионный анализ

Напомним, что в двухфакторном дисперсионном анализе мы проверяем две гипотезы (отдельная гипотеза о влиянии каждого из факторов). Итак, статистика для гипотезы о влиянии фактора a:

$$F_a = \frac{\sigma_a^2}{\sigma_w^2}$$

В предположении верности гипотезы H_{0a} эта статистика имеет распределение Фишера с параметрами $k_{1a} = m - 1$, $k_{2a} = n - m$. Далее, как обычно, строится критическая область (правосторонняя, поскольку распределение Фишера имеет один хвост), и проводится тест.

Аналогично, для гипотезы о влиянии фактора b статистика:

$$F_b = \frac{\sigma_b^2}{\sigma_w^2}$$

Она также имеет распределение Фишера, теперь с параметрами $k_{1b} = k - 1$, $k_{2b} = n - k$.

Факторный анализ

Факторный анализ

Факторный анализ — это способ приведения множества непосредственно наблюдаемых факторов x_j, $j = 1, \ldots, m$, к меньшему числу новых линейно независимых факторов y_j, $j = 1, \ldots, q$, $q < m$.

Рассмотрим *метод главных компонент*. Этот метод заключается в вычислении собственных значений и собственных векторов для ковариационной матрицы:

$$\text{cov} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \ldots & \sigma_{1m} \\ \sigma_{12} & \sigma_2^2 & \ldots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1m} & \sigma_{2m} & \ldots & \sigma_m^2 \end{pmatrix}$$

Схема метода главных компонент

Допустим, имеется матрица объект-признак: $X = (x_{ij})_{n \times m}$ (т.е. n объектов, m признаков).

Метод главных компонент:

1. *Центрировать* матрицу X, т.е. вычесть из каждого столбца среднее по этому столбцу. В результате получится матрица $X^* = \left(x^*_{ij}\right)_{n \times m}$, в которой средние по столбцам равны 0.
2. Вычислить матрицу несмещённых оценок ковариаций $\mathrm{cov} = (\sigma_{ij})_{m \times m}$.
3. Вычислить собственные векторы и собственные значения матрицы cov.
4. Пусть T — матрица, составленная из q собственных векторов (столбцов), соответствующих q наибольшим собственным значениям. Новая матрица объект-признак: $Y = X^* \cdot T$.

Метод главных компонент

Качество метода главных компонент можно оценить, сравнивая дисперсии признаков до и после применения метода.

Пусть σ_X^2 — сумма дисперсий признаков до применения метода, а σ_Y^2 — сумма дисперсий после применения метода. Тогда *доля объяснённой дисперсии* равна отношению

$$\frac{\sigma_Y^2}{\sigma_X^2}$$

Замечание. Поскольку мы в некотором смысле «отсеиваем» признаки, сумма дисперсий после применения метода не может быть больше, чем до.

Долю объяснённой дисперсии можно интерпретировать как процент сохранённой информации.

Логистическая регрессия

Логистическая регрессия

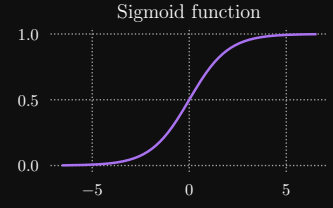
Логистическая регрессия возникает в задачах *бинарной классификации*: исследуется некоторый набор объектов, и каждому объекту приписана бинарная метка (0 или 1).

В модели *логистической регрессии* вероятность объекта $x = (x_0, x_1, \ldots, x_m)$ принадлежать классу 1 моделируется следующим образом:

$$P(y = 1|x) = \sigma(b_0 x_0 + b_1 x_1 + \cdots + b_m x_m) = \sigma(x \cdot b),$$

где $\sigma(z)$ — *логистическая функция* или *сигмоида*:

$$\sigma(z) = \frac{1}{1 + e^{-z}}$$



Сигмоида принимает в качестве аргумента вещественное число, а отдаёт число из промежутка $[0, 1]$.

Замечание. Как и ранее в линейной регрессии, мы под x_0 понимаем «фиктивный» фактор (равный 1 для каждого объекта), который нужен просто чтобы записать выражение в векторном виде $x \cdot b$.

Логистическая регрессия

Для оптимизации параметров модели используется *метод максимального правдоподобия*. Его схему можно изобразить следующим образом:

$$\hat{b} = \underset{b}{\operatorname{argmax}} \prod_{i=1}^{n} P(y = y_i | x = x_i)$$

По сути мы подбираем набор параметров $\hat{b}$ так, чтобы *максимизировать вероятность наблюдать ту выборку, которая у нас есть.*

Логистическая регрессия

Для оптимизации параметров модели используется *метод максимального правдоподобия*. Его схему можно изобразить следующим образом:

$$\hat{b} = \underset{b}{\operatorname{argmax}} \prod_{i=1}^{n} P(y = y_i | x = x_i)$$

По сути мы подбираем набор параметров $\hat{b}$ так, чтобы *максимизировать вероятность наблюдать ту выборку, которая у нас есть.*

Тут кроме формулы для $P(y = 1|x)$ нам понадобится также формула вероятности принадлежности объекта к нулевому классу:

$$P(y = 0|x) = 1 - \sigma(x \cdot b)$$

Отсюда запишем общую вероятность:

$$P(y|x) = \sigma(x \cdot b)^y \cdot (1 - \sigma(x \cdot b))^{1-y}$$

Эти вероятности и используются в методе максимального правдоподобия.

Метод максимального правдоподобия

В практическом смысле удобнее максимизировать не саму функцию, а её логарифм (поскольку в этом случае множители превращаются в слагаемые). Итак, *максимизируется функционал*:

$$Q(b) = \sum_{i=1}^{n} \Big[y_i \cdot \ln\big(\sigma(x_i \cdot b)\big) + (1 - y_i) \cdot \ln\big(1 - \sigma(x_i \cdot b)\big) \Big],$$

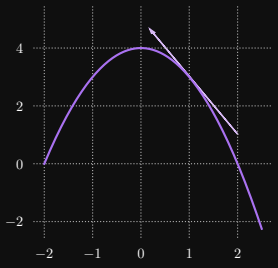
где x_i — набор признаков i-го объекта, y_i — его метка (0 или 1).

Градиентный спуск

Для нахождения оптимального решения используют оптимизационные методы, например, *градиентный спуск*.

Здесь нам понадобится вектор *градиента*, который состоит из частных производных функционала $Q(b)$ по переменным b_j:

$$\nabla Q = \left(\frac{\partial Q}{\partial b_0}, \dots, \frac{\partial Q}{\partial b_m} \right)$$



Результат взятия каждой частной производной вычисляется по формуле:

$$\frac{\partial Q}{\partial b_j} = \sum_{i=1}^{n} \big(y_i - \sigma(b_0 x_{i0} + \dots + b_m x_{im}) \big) x_{ij},$$

где x_{ij} — j-й признак i-го объекта из выборки.

Вектор градиента указывает направление *наискорейшего роста*.

Градиентный спуск

Непосредственно метод градиентного спуска заключается в следующем. Сначала выбираются начальные значения параметров $b_0, \ldots, b_m$, т.е. вектор $b^{[0]}$. Затем итеративно повторяется вычисление:

$$b^{[k+1]} = b^{[k]} + \lambda_k \nabla Q\left(b^{[k]}\right)$$

Замечание. Перед вектором градиента стоит знак «$+$», поскольку мы хотим двигаться в направлении роста функционала.

Параметр λ_k отвечает за скорость спуска.

Описанный выше процесс повторяется, пока соседние векторы $b^{[k+1]}$, $b^{[k]}$ не перестанут сильно отличаться друг от друга.

Логистическая регрессия. Принятие решения

Напомним, что модель логистической регрессии можно записать в следующем виде:

$$P(y = 1|x) = \sigma(x \cdot b)$$

Такая модель на выходе даёт значение из интервала $[0, 1]$, которое интерпретируется как вероятность объекта x принадлежать классу 1. Как правило, дальше по некоторому пороговому значению t принимается решение о том, к какому классу причислять объект:

$$y = \begin{cases} 1, & P(y = 1|x) \geq t, \\ 0 & \text{иначе.} \end{cases}$$

В силу своей линейной природы модель логистической регрессии вместе с пороговым значением t представляет собой *разделяющую гиперплоскость*, т.е. $(m-1)$-мерную плоскость (в m-мерном пространстве признаков). Такие гиперплоскости *делят пространство пополам*, т.е. любой объект оказывается либо с одной, либо с другой стороны от этой плоскости. В зависимости от того, в какую половину пространства попадает объект, ему приписывается метка класса 0 или 1.

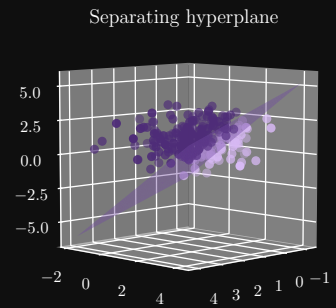
Логистическая регрессия. Разделяющая гиперплоскость

Посчитаем, как именно это делается. Для порогового значения t уравнение разделяющей гиперплоскости имеет вид:

$$t = \sigma(b_0 + b_1 x_1 + \cdots + b_m x_m),$$

откуда, избавляясь от сигмоиды, получаем:

$$b_0 + b_1 x_1 + \cdots + b_m x_m = C,$$



где $C = \ln t - \ln(1-t)$. Такое уравнение и задаёт плоскость. Принятие решения о метке класса для объекта x выглядит теперь следующим образом:

$$y = \begin{cases} 1, & b_0 + b_1 x_1 + \cdots + b_m x_m \geq C, \\ 0 & \text{иначе.} \end{cases}$$

Оценка модели логистической регрессии

Для оценки качества классификации (как бинарной, так и многоклассовой) чаще всего используется метрика *accuracy* (*точность*), которая равна доле верных классификаций.

Пусть y — массив из реальных меток объектов, а z — предсказанные моделью метки. Тогда значение метрики accuracy можно записать следующим образом:

$$accuracy = \frac{1}{n}\sum_{i=1}^{n} I_{y_i=z_i},$$

где I_A — *индикатор*:

$$I_A = \begin{cases} 1, & A \text{ истинно,} \\ 0 & \text{иначе.} \end{cases}$$