1 Лекция 15.11.2018

1.1 One way MANOVA. Модель, запись через мат. ожидания по группам и условные мат.ожидания. Разложение ковариационной матрицы. Критерия Лямбда Уилкса.

Этот раздел во многом аналогичен тексту про ANOVA. Основная разница заключается лишь в том, что здесь η является многомерным вектором.

Общая постановка задачи. Пусть есть k групп и p признаков, мы пытаемся проверить, что группы друг от друга не отличаются. Формально есть k (многомерных) случайных векторов $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_k \in \mathbb{R}^p$ и гипотеза формулируется так:

$$\mathrm{H}_0 : \mathcal{P}(\boldsymbol{\eta}_1) = \mathcal{P}(\boldsymbol{\eta}_2) = \ldots = P(\boldsymbol{\eta}_k). \tag{1}$$

Эту задачу можно переформулировать следующим образом. Пусть есть дискретный (м.б. качественный) признак ξ, принимающий ровно k значений: $A_1, A_2, \ldots, A_k$. Рассмотрим случайный вектор $(\boldsymbol{\eta}, \xi)^{\mathrm{T}} \in \mathbb{R}^k \times \{A_1, A_2, \ldots, A_k\}$, такой что $\mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_i) = \mathcal{P}(\eta_i)$ для всех $i \in 1 : k$. Тогда гипотеза (1) переписывается в виде:

$$\mathrm{H}_0^* : \mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_i) = \mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_j) \text{ для всех } i, j. \tag{2}$$

В MANOVA рассматривается частный случай (модель), когда $\mathcal{P}(\boldsymbol{\eta}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma)$. В рамках этой модели гипотезам (1) и (2) соответствуют следующие две равносильные гипотезы:

$$\begin{gathered} \mathrm{H}_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_k. \\ \mathrm{H}_0^* : \mathbb{E}(\boldsymbol{\eta} \,|\, \xi = A_1) = \ldots = \mathbb{E}(\boldsymbol{\eta} \,|\, \xi = A_k). \end{gathered} \tag{3}$$

Прежде чем сформулировать гипотезу в рамках Основного Дисперсионного Тождества, запишем его на генеральном языке в текущей модели.

$$\mathrm{Cov}\boldsymbol{\eta} = \mathbb{E}\big[(\boldsymbol{\eta} - \mathbb{E}\boldsymbol{\eta})(\boldsymbol{\eta} - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big] = \tag{4}$$

$$= \mathbb{E}\big[(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))^{\mathrm{T}}\big] + \mathbb{E}\big[(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big]. \tag{5}$$

Теперь уже можно легко заметить, что в рамках Основного Дисперсионного Тождества гипотезу можно сформулировать следующим образом:

$$\mathrm{H}_0 : \mathbb{E}\big[(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big] = 0. \tag{6}$$

Эта запись является многомерным обобщением одномерной формулировки задачи ANOVA и также означает, что средние внутри групп не отличаются от общего среднего.

Выборка задается так же, как в ANOVA, но теперь индивиды многомерные. Имеется n_i индивидов из i-той группы ($i \in 1 : k$). Обозначим $\mathbf{y}_{ij} \in \mathbb{R}^p$— j-того индивида из i-той группы ($i \in 1 : k$ и $j \in 1 : n_i$). Обозначим $\overline{\mathbf{y}}$ — выборочное среднее по всем индивидам, а $\overline{\mathbf{y}_i}$ — выборочное среднее индивидов i-той группы.

Запишем (4) на выборочном языке:

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}})(\mathbf{y}_{ij} - \overline{\mathbf{y}})^{\mathrm{T}} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})^{\mathrm{T}} + \sum_{i=1}^{k} n_i(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})^{\mathrm{T}}. \tag{7}$$

Должно быть ясно, что это лишь многомерное обобщение разложения суммы квадратов.

Следуя обобщенным обозначениям, предложенным в предварительных материалах про ANOVA и регрессию, можно ввести следующие:

$$\mathbf{E} \stackrel{\text{def}}{=} \mathbb{E}\big[\widehat{(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))}(\boldsymbol{\eta} - \mathbb{E}(\boldsymbol{\eta} \,|\, \xi))^{\mathrm{T}}\big] = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \overline{\mathbf{y}_i})(\mathbf{y}_{ij} - \overline{\mathbf{y}_i})^{\mathrm{T}};$$

$$\mathbf{H} \stackrel{\text{def}}{=} \mathbb{E}\big[\widehat{(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})}(\mathbb{E}(\boldsymbol{\eta} \,|\, \xi) - \mathbb{E}\boldsymbol{\eta})^{\mathrm{T}}\big] = \sum_{i=1}^{k} n_i (\overline{\mathbf{y}_i} - \overline{\mathbf{y}})(\overline{\mathbf{y}_i} - \overline{\mathbf{y}})^{\mathrm{T}}.$$

Стоит сразу отметить, что, когда в дальнейшем мы будем строить критерии для проверки значимости, мы всегда будем их строить, как некие преобразования над парой матриц $(\mathbf{H}, \mathbf{E})$.

Теорема. *Пусть матрица* $\mathbf{H} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbf{H}})$ *и* $\mathbf{E} \sim \mathcal{W}_p(\Sigma, \nu_{\mathbf{E}})$*, причем* $\mathbf{H}$ *и* $\mathbf{E}$ *независимы. Потребуем также, чтобы* $\nu_{\mathbf{H}} < \nu_{\mathbf{E}}$ *и* $\nu_{\mathbf{E}} > p$*. Обозначим* λ_i *— i-тое по упорядоченности по невозрастанию собственное число матрицы* $\mathbf{E}^{-1}\mathbf{H}$*. Тогда*

$$t = \frac{|\mathbf{E}|}{|\mathbf{H} + \mathbf{E}|} = \frac{1}{|1 + \mathbf{E}^{-1}\mathbf{H}|} = \prod_{i=1}^{s} \frac{1}{1 + \lambda_i} \sim \Lambda_p(\nu_{\mathbf{H}}, \nu_{\mathbf{E}}), \tag{8}$$

где $s = \min(\nu_{\mathbf{H}}, p)$.

Таким образом, (8) дает нам статистику критерия, где в случае дискриминантного анализа $\nu_{\mathbf{H}} = k - 1$, $\nu_{\mathbf{E}} = n - k$. В случае многомерной множественной регрессии $\nu_{\mathbf{H}} = k$, $\nu_{\mathbf{E}} = n - k - 1$.

Значение s — это граница сверху для числа ненулевых чисел, которое определяется рангом матрицы $\mathbf{H}$. Во-первых, ранг не превосходит размерности p матрицы $\mathbf{H}$. Во-вторых, для дисперсионного анализа матрица $\mathbf{H}$ соответствует ковариационной матрице, построенной на основе k индивидов с весами, пропорциональными размерам групп. Поэтому ее ранг не может быть больше k. А так как данные исходно центрированы, то ранг не превосходит $k - 1$.

Заметим, что построенный критерий для проверки соответствующей гипотезы является далеко не единственным. О построении других критериев речь пойдет дальше.

1.2 Про дискриминантный анализ в целом

Постановка задачи, отличие от кластерного анализа.

Задачи — построить классифицирующее правило, объяснить его (интерпретация), feature selection, feature extraction.

Первый вопрос — а вообще группы отличаются?

Рассмотрим модель MANOVA, т.е. распределения в группах нормальные и отличаются только векторами средних. Тогда MANOVA используется для проверки значимости отличия. (гипотеза о том, что группы не отличаются.) В этом случае получаем модель линейного дискриминантного анализ (LDA).

Посмотрим на MANOVA с точки зрения отличия групп в LDA и того, как увидеть это отличие.

1.3 Канонические дискриминантные функции (коэффициенты) и переменные (feature extraction)

Напомним, что имеется p признаков и k групп: $\boldsymbol{\eta} \in \mathbb{R}^p$, причем $\mathcal{P}(\boldsymbol{\eta} \,|\, \xi = A_i) = \mathcal{N}(\boldsymbol{\mu}_i, \Sigma)$. Попытаемся на основе имеющихся p признаках построить новые, по которым

группы «наиболее бы отличались», причем было бы удобно, чтобы эти новые признаки были ортогональны. Эта неформальная задача на самом деле очень просто формализуется, как мы увидем несколько далее.

Запишем на генеральном языке, что значит «новый признак». Есть $A \in \mathbb{R}^p$ и новый признак $\zeta = A^{\mathrm{T}}\boldsymbol{\eta}$: $\mathcal{P}(\zeta \,|\, \xi = A_i) = \mathcal{N}(A^{\mathrm{T}}\boldsymbol{\mu}_i, A^{\mathrm{T}}\Sigma A)$. На выборочном языке — $Z = \mathbf{Y}A$ и выборочная ковариационная матрица (с точностью до коэффициента имеет вид): $A^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}A = A^{\mathrm{T}}(\mathbf{H}+\mathbf{E})A$. Из этого следует,что «аналогом» $\mathbf{H}$ для нового признака является $A^{\mathrm{T}}\mathbf{H}A$, а аналогом $\mathbf{E}$ является $A^{\mathrm{T}}\mathbf{E}A$. [1]

Посмотрим теперь на F-статистику, следуя обобщенному определению статистики, данному в ANOVA:

$$F = F(A) = C\frac{\mathrm{SSH}_\zeta}{\mathrm{SSE}_\zeta} = C\frac{A^{\mathrm{T}}\mathbf{H}A}{A^{\mathrm{T}}\mathbf{E}A} \sim \mathrm{F}_{\nu_{\mathbf{H}},\nu_{\mathbf{E}}}.$$

где $C = \nu_{\mathbf{E}}/\nu_{\mathbf{H}}$ — коэффициент, не зависящий от A.

Исходно ставилась задача — найти такие признаки, по которым группы «наиболее бы отличались», причем желательно, чтобы признаки были ортогональны. В терминах максимизации статистики F — это означает, что решается обобщенная задача на собственные числа и собственные вектора.

$$\frac{A^{\mathrm{T}}\mathbf{H}A}{A^{\mathrm{T}}\mathbf{E}A} \to \max_A.$$

Обозначим собственные числа матрицы $\mathbf{E}^{-1}\mathbf{H}$ в порядке невозрастания: $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_s$, где $s = \min\{p, \nu_H\}$, и собственные вектора той же матрицы — A_i для $i \in 1:s$. Причем $A_i^{\mathrm{T}}\mathbf{E}A_j = 0$ для $j \neq i$. Вектора A_i — канонические коэффициенты (дискриминантные функции). Новые признаки $Z_i = \mathbf{Y}A_i$ — канонические переменные. Можно показать, новые признаки являются ортогональными: $Z_i^{\mathrm{T}}Z_j = 0$ для $j \neq i$.

Теорема.

$$\forall i \neq j \in 1:s \; Z_i \perp Z_j$$

Доказательство. $\langle Z_i, Z_j \rangle = \langle \mathbf{Y}A_i, \mathbf{Y}A_j \rangle = A_i^{\mathrm{T}}\mathbf{Y}^{\mathrm{T}}\mathbf{Y}A_j = A_i^{\mathrm{T}}(\mathbf{H}+\mathbf{E})A_j = A_i^{\mathrm{T}}\mathbf{E}(\mathbf{I}+\mathbf{E}^{-1}\mathbf{H})A_j$. Так как $\forall \mathbf{C} \geq 0$ с с. ч. $\lambda_1, \ldots, \lambda_s$ с. ч. матрицы $\mathbf{I}+\mathbf{C}$ — это $1+\lambda_1, \ldots, 1+\lambda_s$, а собственные вектора те же самые, $(\mathbf{I}+\mathbf{E}^{-1}\mathbf{H})A_j = (1+\lambda_j)A_j$ по определению собственного вектора. Так как собственные вектора разных с.ч. $\mathbf{E}$-ортогональны, $\langle Z_i, Z_j \rangle = (1+\lambda_j)A_i^{\mathrm{T}}\mathbf{E}A_j = 0$. □

Исходя из описания, смысл λ_i — степень различия групп по i-тому направлению, задаваемом i-той дискриминантной функцией. A_i — коэффициенты с которыми нужно взять исходные признаки, чтобы получить новый признак, соответствующий наибольшей разнице групп и ортогональный предыдущим. А сами канонические переменные тем самым — ортогональные новые признаки, по которым соответствующий наибольшее различие равен λ_i.

[1] Чтобы в этом убедиться, нужно написать оценкой чего является $\mathbf{H}$ и $\mathbf{E}$. Дальше домножить это на A нужным образом и ввести обозначение для нового признака. Дальше все станет ясно.