Лекции, начиная с MANOVA

Собрано 8 декабря 2017 г. в 02:14

1 Лекция 23.11.2017

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

Этот билет во многом аналогичен текста 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 (**??**) 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 (**??**) и так же 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 (**??**).

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

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

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

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 так же, чтоб* $\nu_{\mathbf{H}} < \nu_{\mathbf{E}}$ *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 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.

1.4 MANOVA и значимость LDA. Разные критерии, чем отличаются.

Использование MANOVA для проверки значимости дискриминантного анализа совершенно естественно. Один из критериев 1.1 уже был построен — это критерий Лямбда Уилкса.

[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.

На самом деле это далеко не единственный критерий, который может быть построен, в частности, для проверки значимости дискриминантного анализа. Как уже отмечалось, главными объектами, с которыми приходится работать, являются матрицы $\mathbf{H}$ и $\mathbf{E}$. На основе разных способов «совмещения» этих объектов могут получаться разные критерии. Рассмотрим несколько способов. Напомним, что $s = \min\{\nu_{\mathbf{H}}, p\}$.

Проверяется гипотеза о равенстве средних[2]

$$\mathrm{H}_0 : \mathbb{E}\boldsymbol{\eta}_1 = \mathbb{E}\boldsymbol{\eta}_2 = \ldots = \mathbb{E}\boldsymbol{\eta}_k.$$

1. Lambda Wilks's
$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \frac{1}{|1+\mathbf{E}^{-1}\mathbf{H}|} = \prod_{i=1}^{s}\frac{1}{1+\lambda_i} \sim \Lambda_p(\nu_{\mathbf{H}}, \nu_{\mathbf{E}}).$$

2. Roy's largest root
$$Q = \frac{\lambda_1}{1+\lambda_1} = r_1^2.$$
Этот тест использует только первое собственное число матрицы $\mathbf{E}^{-1}\mathbf{H}$. Напомним, что r_i^2 — называют i-тым каноническим корнем. На самом деле r_i — так называемые канонические корреляции [3]. Потому они на самом деле имеют интерпретируемый смысл.

3. Pillai's:
$$V^{(s)} = \mathrm{tr}\left(\mathbf{H}(\mathbf{E}+\mathbf{H})^{-1}\right) = \sum_{i=1}^{s} r_i^2.$$

4. Hotelling:
$$U^{(s)} = \mathrm{tr}\left(\mathbf{E}^{-1}\mathbf{H}\right) = \sum_{i=1}^{s}\lambda_i.$$

Как всегда, когда мы видим много критериев для проверки одной гипотезы, нужно научиться их сравнивать. Сравнивают критерии по мощности. Но, чтобы говорить о мощности нужно фиксировать альтернативу. Таким образом, вопрос можно поставить следующим образом: «Для каких альтернатив предложенные критериев более мощные?».

Посмотрим на это на примере первых двух критериев. Первый критерий включает в статистику все (!) направления, по которым разброс максимален (иными словами, все λ_i). Второй же критерий включает в себя лишь первое направление, с максимальной степенью разброса, измеряемой λ_1. Из этого становится ясно, что если у нас на самом деле лишь одно направление определяет разброс (например, k шариков лежат на одной прямой), то остальные λ_i при $i > 2$ уже не отображают различие между группами и потому в такой ситуации следует ожидать, что второй критерий окажется мощнее первого. В обратном случае, когда все направления описывают различие между группами, следует ожидать, что первый критерий окажется мощнее второго.

Про третий и четвертый критерий в литературе утверждается, что они «где-то по середине» между первым и вторым.

В заключение заметим, что у всех критериев кроме первого критическая область находится «слева» и около 0 (все статистики неотрицательны), у первого критерия все наоборот: носитель статистики от 0 до 1 и критическая область расположена около 1.

[2]На самом деле о равенстве распределений в рамках нормальной модели при условии гомоскедастичности.

[3] Позднее, если успею, расскажу про то, между чем это корреляции.

1.5 Частный случай — ошибки сферические

Пусть $\mathbf{Y} \in \mathbb{R}^{n\times p}$ — матрица данных (строки $\mathbf{y}_k$, $k \in 1:n$ (или $\mathbf{y}_{ij}$ в другой нумерации по группам) – наблюдения, столбцы $Y_j, j \in 1:p$ – признаки), наблюдения принадлежат к одной из k групп, в каждой группе n_i наблюдений, $n = \sum_{i=1}^{k} n_i$. Матрица межгрупповых отклонений $\mathbf{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbf{E} = \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)^{\mathrm{T}}$.

В случае, если $\mathbf{E} = \sigma^2\mathbf{I}$ (все группы сферические и одинакового размера), дискриминантные функции данных, т.е. $A_1, \ldots, A_s$ – собственные вектора $\mathbf{E}^{-1}\mathbf{H}$, становятся собственными векторами матрицы $\sigma^2\mathbf{H}$. То есть дискриминантные функции являются собственными векторами матрицы межгрупповых ковариаций. Так они совпадают с главными направлениями в терминах анализа главных компонент, если каждую группу рассматривать как одного индивида (в АГК главные направления — это собственные вектора ковариационной матрицы), а дискриминантные переменные $\mathbf{Y}A_i$ при такой постановке являются главными компонентами.

2 Лекция 30.11.2017

2.1 Значимое число дискриминантных функций (размерность пространства, где группы различаются).

Модель дискриминантного анализа (на генеральном языке): ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^{k}$, распределение $\boldsymbol{\eta}$ при условии $\xi = A_i$ совпадает с распределением $\boldsymbol{\eta}_i$. Определены матрицы

$$\mathbf{H}_{th} = \mathbb{E}\left(\left(\mathbb{E}\left(\boldsymbol{\eta}|\xi\right) - \mathbb{E}\boldsymbol{\eta}\right)\left(\mathbb{E}\left(\boldsymbol{\eta}|\xi\right) - \ E\boldsymbol{\eta}\right)^{\mathrm{T}}\right) \in \mathbb{R}^{p\times p},$$

$$\mathbf{E}_{th} = \mathbb{E}\left(\left(\boldsymbol{\eta} - \mathbb{E}\left(\boldsymbol{\eta}|\xi\right)\right)\left(\boldsymbol{\eta} - \mathbb{E}\left(\boldsymbol{\eta}|\xi\right)\right)^{\mathrm{T}}\right) \in \mathbb{R}^{p\times p}.$$

Дискриминантные функции $A_1, \ldots, A_s$ — это с.в. матрицы $\mathbf{E}^{-1}\mathbf{H}$, $s = \min\{p, \nu_H\}$.

Для того, чтобы узнать, сколько функций являются значимыми (сколько новых признаков интерпретировать), мы последовательно проверяем гипотезы H_0: «дискриминантные функции с номерами $m, \ldots, s$ не описывают различия в данных» для возрастающих $m \in 1:s$. Формально, гипотеза эквивалентна тому, $\mathrm{rk}\mathbf{H}_{th} \leq m-1$, т.е. все собственные числа матрицы, начиная с m-го равны нулю. Мы можем позволить себе проверять её последовательно, так как знаем, что дискриминантные функции, будучи отсортированы по убыванию соответствующих собственных чисел, оказываются в порядке убывания качества объяснения различий.

Если H_0 верна, то может быть показано, что

$$\Lambda'_m = \prod_{i=m}^{s} \frac{1}{1+\lambda_i} \sim \Lambda_p\left(\nu_H + (m-1), \nu_E - (m-1)\right).$$

Здесь уже матрицы, полученные по выборке.

С помощью статистики Λ'_m (lambda prime) мы для каждого m можем проверить, незначимы ли эта и все последующие дискриминантные функции. Если они значимы, то все перед ними тоже значимы и m нужно увеличить.

2.2 Интерпретация разделения: стандартизованные дискриминантные функции и факторная структура

Пусть $\mathbf{Y}$ — матрица данных (строки $\mathbf{y}_i, i \in 1:n$ – наблюдения, столбцы $Y_j, j \in 1:p$ – признаки), наблюдения принадлежат к одной из k групп, в каждой группе n_i наблюдений, $n = \sum_{i=1}^{k} n_i$. Матрица межгрупповых отклонений $\mathbf{H} = \sum_{i=1}^{k} n_i \left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)\left(\overline{\mathbf{y}}_i - \overline{\mathbf{y}}\right)^{\mathrm{T}}$, матрица внутригрупповых отклонений $\mathbf{E} = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i\right)\left(\mathbf{y}_{ij} - \overline{\mathbf{y}}_i^{\mathrm{T}}\right)$.

$A_1, \ldots, A_s$ — собственные вектора $\mathbf{E}^{-1}\mathbf{H}$, называемые дискриминантными функциями, λ_i — упорядоченные по убыванию собственные числа. Канонические переменные — вектора данных в новых координатах $Z_i = \mathbf{Y}A_i$.

Задача: проинтепретировать канонические переменные (новые признаки).

2.2.1 Стандартизованные дискриминантные функции

Первый способ проинтерпретировать разложение по дискриминантным переменным — посмотреть на коэффициенты, с которыми исходные переменные входят в дискриминантные. Если исходные переменные измерены в различных шкалах, то коэффициенты в векторе A_i одновременно ещё и приводят показатели к нужной шкале. Чтобы избежать этого эффекта, можно посмотреть на стандартизованные дискриминантные функции.

Пусть $\mathbf{S}$ – матрица взвешенных ковариаций (pooled covariance matrix), $s_1^2, \ldots, s_p^2$ – элементы на её диагонали (взвешенные дисперсии признаков, pooled variance), $A_i = \begin{pmatrix} a_{i1} \\ \vdots \\ a_{ip} \end{pmatrix}$.

Тогда

$$Z_i = \mathbf{Y}A_i = \sum_{j=1}^{p} Y_j a_{ij} = \sum_{j=1}^{p} \frac{Y_j}{s_j} \cdot s_j a_{ij},$$

и $\widetilde{A}_i = (\mathrm{diag}\mathbf{S})^{\frac{1}{2}} A_i$ — i-я стандартизованная дискриминантная функция (стд. д. ф.). Коэффициенты стд. д. ф. показывают вклады исходных признаков в дискриминантные переменные.

2.2.2 Факторная структура

Факторная структура — матрица корреляций между исходными и каноническими переменными. Так как дискриминантные функции, вообще говоря, не ортогональны, то получатся другие числа. Что правильнее - сложный вопрос.

2.3 Уменьшение числа признаков (feature selection)

«Плохие» признаки, это признаки, которые:

1. Являются линейной комбинацией других признаков, т.е. имеют большой коэффициент множественной корреляции $R^2 = R^2\left(\eta^{(i)}; \left\{\eta^{(i)} \middle| j \in 1:p \ \{i\}\right\}\right)$. Соответствующая характеристика – tolerance $= 1 - R^2$.

2. При удалении из модели не влияют на качество разделения. Соответствующая гипотеза H_0: «удаление признака i не влияет на качество разделения», т.е. от удаления i-го признака теоретическая Λ не меняется. Статистика:

$$(\text{Partial } \Lambda)_i = \frac{\Lambda\left(Y_1, \ldots, Y_p\right)}{\Lambda\left(Y_1, \ldots, Y_{i-1}, Y_{i+1}, \ldots, Y_p\right)} \sim \Lambda_1\left(\nu_H, \nu_E - p + 1\right) \tag{9}$$

2.3.1 Пошаговый дискриминантный анализ

Пошаговый дискриминантный анализ подбирает тот набор признаков, который лучше всего будет разделять переменные (аналогично пошаговой множественной регрессии) жадным образом. На каждом шаге добавляется одна переменная, которая максимально увеличивает качество разделения групп (качество меряется статистикой partial lambda из (9), точнее, из-за наличия точного преобразования из распределения Λ_1 в F-распределение, эквивалентной статистикой с распределением Фишера). После этого набранная модель пересматривается на предмет наличия избыточных переменных. Процедура останавливается, когда максимальное значение F для вновь добавляемых переменных не превосходит наперёд заданного порога.

Первой переменной в модели становится та, для которой F-статистика из ANOVA оказалась наибольшей.

Выше описан вариант forward с добавлением признаков. Аналогично можно рассмотреть backward версию.

2.4 Общий подход к классификации через апостериорные вероятности

Общая подход к классификации: строятся классифицирующие функции f_i, такие что классификация проводится так: индивид с признаками x относится к группе с максимальным значением на нем классифицирующей функции: $\arg\max \mathrm{f}_i(x)$.

Откуда берутся эти классифицирующие функции? Естественная идея взять в качестве f_i вероятность (ее оценку) принадлежности к i-му классу. Пусть ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\eta \sim \mathcal{P}_i$ и имеет плотность $p_i(x)$, если $\xi = A_i$. Тогда было бы логично взять $f_i = p_i$. Для практического применения надо было бы оценить плотности, либо непараметрически (например, по числу точек, попавших в дельта-окрестность — типа метода ближайших соседей), либо параметрически (если известно, что распределение нормальное, тогда просто оцениваем векторы средних и ковариационные матрицы).

Более сложный подход — через апостериорные вероятности. Если у нас есть априорное знание вероятности того, что индивид из того или иного класса, то мы можем его учесть. Введем понятие класса $C_i = \{\xi = A_i\}$. Чтобы классифицировать наблюдение x, необходимо найти

$$\arg\max \mathrm{P}\left(\xi \in A_i | \eta = x\right) = \arg\max \mathrm{P}\left(C_i | x\right).$$

Пусть известны априорные вероятности принадлежности нового наблюдения к i-му классу $\pi_i = \mathrm{P}\left(C_i\right)$. Тогда апостериорные вероятности по формуле Байеса будут иметь вид

$$\mathrm{P}\left(C_i | x\right) = \frac{\mathrm{P}\left(x | C_i\right)\pi_i}{\sum_{j=1}^k \mathrm{P}\left(x | C_j\right)\pi_j}.$$

Поэтому в качестве классифицирующих функций берут

$$f_i\left(x\right) = \frac{p_i(x)\pi_i}{\sum_{j=1}^k p_j(x)\pi_j}.$$

Так как знаменатель у всех f_i одинаковый, его можно отбросить, и итоговые классифицирующие функции будут выглядеть как $f_i\left(x\right) = \mathrm{P}\left(x | C_i\right)\pi_i$.

Как выбрать априорные вероятности?

1. Равномерно, $\forall i \in 1 : k\ \pi_i = 1 \,/\, k$.
2. По соотношениям в обучающей выборке: $\pi_i = n_i \,/\, \sum_{j=1}^k n_j$.

3. На основе другой дополнительной информации о данных (результаты предыдущих исследований, etc.)

Свойство. *Построенный метод классификации* $\text{predict}(x) = \arg\max_i \pi_i p_i(x)$ *минимизирует среднюю апостериорную ошибку:*

$$\sum_{i=1}^{k} \pi_i P(\text{predict}(x)! = i \mid C_i).$$

Видно, что можно с помощью априорных вероятностей формально задавать важность ошибочных классификаций для разных классов.

3 Лекция 07.12.2017

3.1 Линейный и квадратичный дискриминантный анализ для классификации

3.1.1 LDA

Модель: ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\eta \sim \mathcal{N}(\mu_i, \Sigma)$, если $\xi = A_i$. Тогда плотность x

$$p(x|\xi = A_i) = \frac{1}{(2\pi)^{p/2} |\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{\mathrm{T}} \Sigma^{-1} (x-\mu_i)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p(x|\xi = A_i)$, где π_i — априорная вероятность наблюдения попасть в i-ю группу. Для упрощения вычислений можно переписать классифицирующую функцию как

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x-\mu_i)^{\mathrm{T}} \Sigma^{-1} (x-\mu_i).$$

Сократив часть, не зависящую от номера класса, получаем линейные классифицирующие функции

$$h_i(x) = \frac{1}{2}\mu_i^{\mathrm{T}} \Sigma^{-1} \mu_i + \mu_i \Sigma^{-1} x + \log \pi_i.$$

3.1.2 QDA

Модель: ξ – дискретная с.в., принимающая значения $\{A_i\}_{i=1}^k$, $\eta \sim \mathcal{N}(\mu_i, \Sigma_i)$, если $\xi = A_i$. Тогда плотность x

$$p(x|\xi = A_i) = \frac{1}{(2\pi)^{p/2} |\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_i)^{\mathrm{T}} \Sigma_i^{-1} (x-\mu_i)\right),$$

и классифицирующая функция $f_i(x) = \pi_i p(x|\xi = A_i)$. Оставляем в классифицирующей функции только монотонность и члены, отличающиеся в разных группах:

$$g_i(x) = \log f_i(x) = \log \pi_i - \frac{1}{2}\log|\Sigma_i| - \frac{1}{2}(x-\mu_i)^{\mathrm{T}} \Sigma_i^{-1} (x-\mu_i),$$

получаем квадратично зависящую от x классифицирующую функцию.

3.2 Классификация в случае двух классов

Если всего два класса, то можно построить границу между классами, приравняв классифицирующие функции.

3.2.1 LDA

Приравняв $h_1(x) = h_2()$, получим разделяющую гиперплоскость Разделяющая два класса гиперплоскость имеет вид

$$\{\mathbf{x} : h_1(\mathbf{x}) = h_2(\mathbf{x})\} = \\ = \{\mathbf{x} : -\frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mu_1 + \mu_2) + (\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x} + \log(\pi_1/\pi_2) = 0\}.$$

От соотношения между априорными вероятностями зависит положение границы относительно классов (к какому она ближе). Видно, что априорные вероятности влияют только на сдвиг разделяющей гиперплоскости.

Заметим, что классификацию можно записать как сравнение $-\frac{1}{2}(\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mu_1 + \mu_2) + (\mu_1 - \mu_2)^{\mathrm{T}}\mathbf{\Sigma}^{-1}\mathbf{x}$ с некоторым порогом $(-\log(\pi_1/\pi_2))$, который зависит от априорных вероятностей (или весов ошибок для разных классов, смотря как на это смотреть).

3.2.2 QDA

В данном случае, разделяющая поверхность имеет вид квадратичной поверхности, может состоять из двух гиперболоидом, может иметь форму эллипса.

3.2.3 Картинки

Здесь мы обсуждали число параметров в моделях, возможный overfitting (переподгонку). Использовали слова — обобщающая способность алгоритма.

3.3 Качество классификации

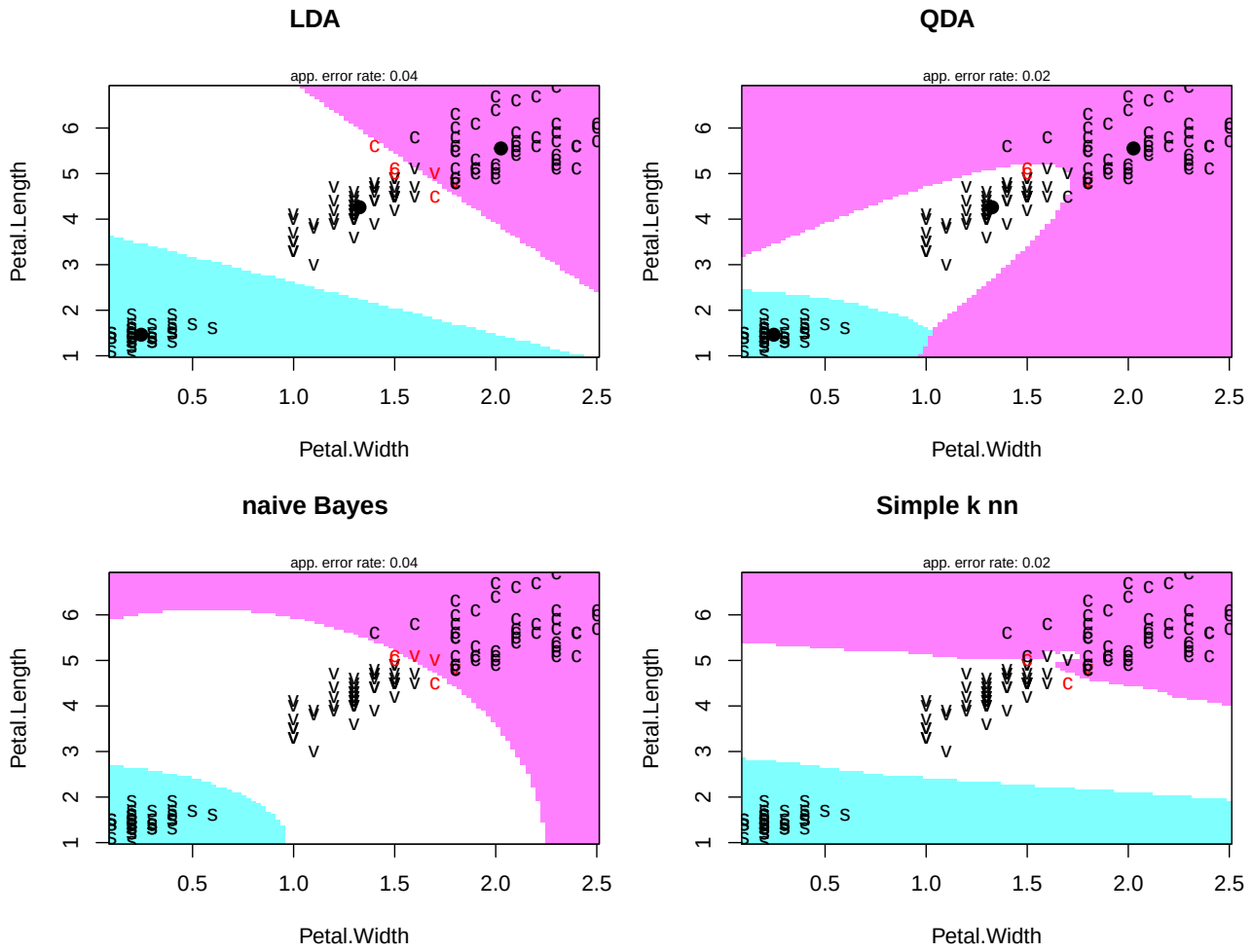
3.3.1 Ошибки класификации

Качество классификации измеряется ошибками классификации (доля неправильно классифицированных объектов). n_{ij} — число объектов из класса i, отнесенных к классу j. В соответствующей матрице классификации на диагонали стоят правильно классифицированные объекты, вне диагонали — ошибки.

На самом деле, нельзя проверять качество предсказания на тех данных, на которых это предсказание строилось. Поэтому используют кросс-валидацию (скользящий контроль). Например, каждое наблюдение по очереди исключается из выборки, классифицирующее правило строится без него и с помощью этого правила индивид классифицируется. Строится аналогичная таблица из n_{ij}. В ней ошибок будет, вообще говоря, больше.

Здесь обсуждали, что имеет смысл смотреть на ошибки без кросс-валидации и с ней. Если разница существенная, то это говорит о переподгонке используемой модели. Вероятно, она не очень хорошая; например, слишком много параметров.

Замечание. Нельзя путать классификацию и различие групп. Группы могут значимо различаться, классификация может быть при этом бессмысленной (ошибок чуть меньше 50%).



3.3.2 ROC и AUC

wikipedia ROC-кривая (англ. receiver operating characteristic, рабочая характеристика приёмника) — график, позволяющий оценить качество бинарной классификации, отображает соотношение между долей объектов от общего количества носителей признака, верно классифицированных как несущих признак, (англ. true positive rate, TPR, называемой чувствительностью алгоритма классификации) и долей объектов от общего количества объектов, не несущих признака, ошибочно классифицированных как несущих признак (англ. false positive rate, FPR, величина 1-FPR называется специфичностью алгоритма классификации) при варьировании порога решающего правила.

Также известна как кривая ошибок. Анализ классификаций с применением ROC-кривых называется ROC-анализом.

Количественную интерпретацию ROC даёт показатель AUC (англ. area under ROC curve, площадь под ROC-кривой) — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций. Чем выше показатель AUC, тем качественнее классификатор, при этом значение 0,5 демонстрирует непригодность выбранного метода классификации (соответствует случайному гаданию). Значение менее 0,5 говорит, что классификатор действует с точностью до наоборот: если положительные назвать отрицательными и наоборот, классификатор будет работать лучше.

мои комментарии Если кто-то хорошо представляет себе, как выглядит график зависимости мощности от ошибки первого рода, то это именно такой график. Меняется уровень значимости (как порог отвергнуть - не отвергнуть) и по оси x откладывается

ошибка первого рода, она же false positive, а по оси y откладывается мощность, она же true positive (слово positive означает, что нулевая гипотеза отвергнута в пользу второй, альтернативной, гипотезы, а в случае классификации, что элемент классифицируется как относящийся ко второму классу).

Таким образом, меняем порог для метода классификации и по оси x откладываем долю неправильно классифицированных элементов из первого класса, а по оси y — долю правильно классифицированных элементов из второго класса.

3.4 Канонические корреляции

Пусть у нас имеется два набора случайных величин:

$$\eta = (\eta_1, \ldots, \eta_q)^{\mathrm{T}} \tag{10}$$

$$\xi = (\xi_1, \ldots, \xi_p)^{\mathrm{T}} \tag{11}$$

Тогда

Определение. *Первой канонической корреляцией называется*

$$r_1^2 = \max_{A,B} \rho^2(A^{\mathrm{T}}\xi, B^{\mathrm{T}}\eta), \tag{12}$$

где $A \in \mathbb{R}^p, B \in \mathbb{R}^q$. i-ой каноническая корреляция определяется аналогично с условием, что максимум берется по некоррелированным с предыдущими случайным величинам:

$$r_i^2 = \max_{\substack{A,B \\ \rho(A\xi, A_j\xi)=0 \\ \rho(B\eta, B_j\eta)=0 \\ 1 \le j < i}} \rho^2(A^{\mathrm{T}}\xi, B^{\mathrm{T}}\eta) \tag{13}$$

Канонических корреляций будет, очевидно, $s = \min(p, q)$.

Так как считаются корреляции, это не важно, случайные величины центрированы или нет, стандартизованы или нет.

На выборочном языке, первая каноническая корреляция — это квадрат косинуса угла между подпространствами, натянутыми на первый и второй наборы признаков (центрированные). Собственно, угол между подпространствами так и определяется как минимальный угол между (ненулевыми) векторами из подпространств.

Множественная корреляция — частный случай, когда один из наборов переменных состоит только из одного признака. Это следует из того, что множественный коэффициент корреляции в квадрате — это тоже квадрат косинуса угла между вектором и подпространством.

3.5 Канонические переменные

Смотрим на (13). У нас каждая каноническая корреляция определяется с помощью двух новых признаков, $\widetilde{\eta}_i = B_i^{\mathrm{T}}\eta$ и $\widetilde{\xi}_i = A_i^{\mathrm{T}}\xi$. Они определяют новые ортогональные системы признаков в пространстве, натянутом на признаки из одного набора, и в пространстве, натянутом на признаки из другого набора. Новые признаки $\widetilde{\eta}_i = B_i^{\mathrm{T}}\eta$ и $\widetilde{\xi}_i = A_i^{\mathrm{T}}\xi$ называются правыми и левыми каноническими переменными (или X- и Y- каноническими переменными).

Скаттерплот, на котором лучше всего видна линейная зависимость между двумя наборами признаков — это скаттеплот, построенный по первым левой и правой канонической переменной. Корреляция между ними равна первой канонической корреляции.

Интерпретация канонических переменных через канонические функции такая же, как интерпретация канонических переменных через дискриминантные функции, так как это просто коэффициенты линейной комбинации, показывающей, как новые признаки выражаются через старые.

Как обычно, чтобы избавиться от масштаба, нужно стандартизовать признаки, разделив их на выборочное стандартное отклонение, тем самым, сами коэффициенты умножатся на то же число. Получатся стандартизованные канонические функции.

3.6 Канонический корреляционный анализ и многомерная множественная линейная регрессия

Рассмотрим многомерную множественную регрессию

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \Xi,$$

здесь столбцы матрицы $\mathbf{X} \in \mathbb{R}^{n\times p}$ — регрессоры (то, что поступает на вход), а столбцы $\mathbf{Y} \in \mathbb{R}^{n\times q}$ — это отклик (зависимые переменные, результат измерений). Матрица из неизвестных коэффициентов $\mathbf{B} \in \mathbb{R}^{p\times q}$. Матрица $\Xi \in \mathbb{R}^{n\times q}$ состоит из независимых случайных величин с нулевым мат.ожиданием и одинаковой дисперсией.

Предположим, что столбцы $\mathbf{Y}$ и $\mathbf{X}$ — центрированы. Решение данной задачи $\widehat{\mathbf{B}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$, предсказание имеет вид $\widehat{\mathbf{Y}} = \mathbf{X}\widehat{\mathbf{B}}$. Этот результат отдельно доказывать не надо, так как многомерная множественная регрессия - это просто много (q штук) обычных множественных линейных регрессий.

Значимость регрессии получаем, если отвергается гипотеза

$$H_0 : \mathbf{B} = 0.$$

Для проверки этой гипотезы надо построить критерий, его у нас еще не было. Положим $\mathbf{E} = (\widehat{\mathbf{Y}} - \mathbf{Y})(\widehat{\mathbf{Y}} - \mathbf{Y})^{\mathrm{T}}$, $\mathbf{H} = \widehat{\mathbf{Y}}\widehat{\mathbf{Y}}^{\mathrm{T}}$ и пусть, как обычно, статистика критерия имеет вид

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E}+\mathbf{H}|} = \prod_{i=1}^{s}\left(\frac{1}{1+\lambda_i}\right) = \prod_{i=1}^{s}(1-r_i^2), \tag{14}$$

где λ_i — собственные числа матрицы $\mathbf{E}^{-1}\mathbf{H}$, $s = \min(p, q)$. Если ошибки Ξ кроме указанных выше свойств имеют еще и совместное нормальное распределение, то статистика критерия имеет, как и раньше, распределение $\Lambda_p(\nu_H, \nu_E)$, где
$\nu_H = q$, $\nu_E = n - \nu_H - 1$. Точно так же можно рассмотреть и другие критерии, как в MANOVA.

Оказывается, что $r_i^2 = \frac{\lambda_i}{1+\lambda_i}$ — это i-ая каноническая корреляция (в квадрате). Доказательство техническое, не привожу. Поэтому все критерии можно переписать через канонические корреляции. Для регрессии такой их вид более естествен.

Замечание. *Если рассмотреть дискриминантный анализ и вместо категоризующей переменной с k значениями взять $k-1$ фиктивных (dummy) переменных, состоящих из нулей и единиц, то дискриминантный анализ можно свести к многомерной множественной линейной регрессии с этими $k-1$ переменными в качестве отклика.*

==========

Как и в дисперсионном анализе, встает вопрос о значимых корреляционных переменных, или, что то же самое, о числе значимых (ненулевых) канонических корреляциях.

Мы не будем отдельно рассматривать этот вопрос, так как делается это точно так же, как это делали с помощью Λ_{prime} в LDA.

3.7 Кластерный анализ

Цель кластерного анализа — разбить индивиды на кластеры, т.е., на группы, между которыми, в некотором смысле, расстояние больше, чем между точками внутри. Задача не формализована и, можно сказать, плохо поставлены, поэтому решается плохо.

Вообще, кластерный анализ — это 'обучение без учителя'. Это означает, что вы не сможете формально проверить правильность результата.

Единственный вариант поставить задачу четко — это предположить какую-то статистическую модель данных и в ней находить параметры, например, по методу максимального правдоподобия (model-based clustering).

Все остальные методы — эвристические с плохо определенным (хорошо-плохо) результатом.

3.8 Кластерный анализ, пример model-based подхода

Предположим, что многомерная выборка — неоднородная. Но в отличие от дискриминантного анализа у нас нет признака, объясняющего эту неоднородность, и задачей является ее выявить. Тип классификации, когда есть модель, называется model-based clustering. Например, пусть наша выборка из смеси k нормальных распределений. Таким образом ее плотность имеет вид

$$p(x) = \pi_1 p(x, \mu_1, \Sigma_1) + \ldots + \pi_k p(x, \mu_k, \Sigma_k), \tag{15}$$

где

$$p(x, \mu_i, \Sigma_i) = \frac{1}{(2\pi)^{p/2}\sqrt{|\Sigma_i|}} \exp\left(-\frac{1}{2}(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)^T\right) \tag{16}$$

Эта задача решается методом максимального правдоподобия. Можно выписать функцию правдоподобия (выпишите), но она имеет сложный вид и искать ее максимум по такому большому числу параметров очень непросто. Для нахождения этого максимума используется так называемый EM-алгоритма (Expectation - Maximization). Мы не будем здесь его обсуждать.

3.9 Кластерный анализ: k-means, k-means++

Хотим искать кластеры $C_1, \ldots, C_k$ минимизируя следующий функционал

$$\sum_{i=1}^{k} \sum_{j \in C_i} ||x_j - \mu_i||^2 \tag{17}$$

по разбиению всего пространства индивидов на C_j и по всем μ_i. Можно делать это по следующему алгоритму:

1. Выбираем случайно $\mu_1, \ldots, \mu_k$.
2. C_j – кластер, содержащий точки, которые лежат к μ_j ближе, чем к остальным μ_i.
3. Для каждого C_j пересчитываем центр μ_j как выборочное среднее элементов из этого кластера.
4. Делаем 2 и 3 пока алгоритм не сойдется.

Проблема метода в том, что у такого функционала много локальных минимумов, и алгоритм может сойтись в значение, далекое от истинного. Метод k-means++ повторяет алгоритм, приведенный выше, но начальные значения выбираются не случайно, а следующим образом

1. Выбираем случайным образом первый центр μ_1.
2. Считаем расстояние от всех точек до ближайшего центра $\{\rho_i\}$. После чего выбираем x_i как новый центр с вероятностью, пропорциональной ρ_i.
3. Пока количество центров меньше, чем k, повторяем процедуру.

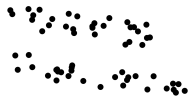
Результат функционала в k-means для данной процедуры выбора начальных центров запишем как $J(\{C_j\}, \{\mu_j\})$. Известно, что при некоторых условиях на форму кластеров

$$\frac{\mathbb{E}(J(\{C_j\}, \{\mu_j\}))}{J_{min}} = O(\ln k),$$

т.е. результат, в среднем, довольно близко к настоящему минимуму.

Замечание. *Есть результаты, что если к данным применить анализ главных компонент, то пространство, натянутое на первые $k-1$ главных векторов, при некоторых условиях будет близко к пространству, проходящему, через центры кластеров. Поэтому часто с помощью АГК уменьшают число признаков и потом применяют процедуру кластерного анализа.*

3.10 «Плохие» кластерные структуры

1.  ленточные кластеры. Внутрикластерные расстояния могут быть больше межкластерных;
2.  перекрывающиеся кластеры;
3.  кластеры, соединяющиеся перемычками и накладывающиеся на фон из редко расположенных объектов;
4.  кластеры могу отсутствовать.

Здесь мы обсуждали, что практически невозможно придумать определение кластера (не статистическое), при котором все эти кластеры будут ему удовлетворять. Вариант смеси нормальных распределений, возможно, подойдет во всех случаях.

Еще обсуждали вопрос, что для данных, где реально обособленных кластеров может и не быть (например, последняя картинка), часто кластеризацией называют сегментацию — просто нарезку на части с описанием каждого сегмента на основе значений признаков.

3.11 Иерархический кластерный анализ

3.11.1 Расстояние между точками ρ

Сначала нужно задать, как мы будем измерять расстояние между точками.
Самое стандартное — евклидово расстояние: $\rho(x,y) = (\sum_i (x_i - y_i)^2)^{1/2}$.
Расстояние городских кварталов (манхэттенское расстояние): $\rho(x,y) = \sum_i |x_i - y_i|$.
Расстояние Чебышёва: $\rho(x,y) = \max_i |x_i - y_i|$.
Процент несогласия (эта мера используется в тех случаях, когда данные являются категориальными): $\rho(x,y) = (\#\{i : x_i \neq y_i\}/i$.

Особый случай, если кластеризуются признаки, а не индивиды (а какая разница — такой кластерный анализ не статистическая процедура, ему все равно), то логично в качестве расстояния рассматривать корреляции. Например, 1 минус модуль корреляции или 1 минус просто корреляция, что правильнее по смыслу для задачи.

Замечание. *Важно либо исходно стандартизовать признаки, либо измерять расстояние специальным образом. Например, использовать расстояние Махаланобиса вместо обычного евклидового, если есть предположения о форме распределения точек внутри кластера.*

3.11.2 Примеры межкластерных расстояний

Правила слияния кластеров (linkage rule) основывается на расстояниях между кластерами.

Расстояние ближнего соседа (single linkage, кластеры в виде цепочек):

$$R^n(U,V) = \min_{u \in U, v \in V} \rho(u,v),\ U,V \subset X;$$

расстояние дальнего соседа (complete linkage, кластеры ближе к шарикам):

$$R^l(U,V) = \max_{u \in U, v \in V} \rho(u,v);$$

групповое среднее расстояние:

$$R^g(U,V) = \frac{1}{|U||V|} \sum_{u \in U} \sum_{v \in V} \rho(u,v);$$

расстояние между центрами:

$$R^c(U,V) = \rho^2 \left(\sum_{u \in U} \frac{u}{|U|}, \sum_{v \in V} \frac{v}{|V|} \right);$$

расстояние Уорда:

$$R^w(U,V) = \frac{|U||V|}{|U| + |V|} R^c(U,V).$$

3.11.3 Алгоритм агломеративной иерархической кластеризации

1. Сначала все кластеры одноэлементные: $C_1 = \{\{x_1\}, \ldots, \{x_l\}\}$; $R_1 = 0$;
   $\forall i \neq j$ вычислить $R(\{x_i\},\{x_j\})$;

2. для всех $t = 2, \ldots, l$ (t – номер итерации)

3. найти в C_{t-1} два ближайших кластера:
$(U,V) = \arg\min_{U \neq V} R(U,V)$;
$R_t = R(U,V)$;

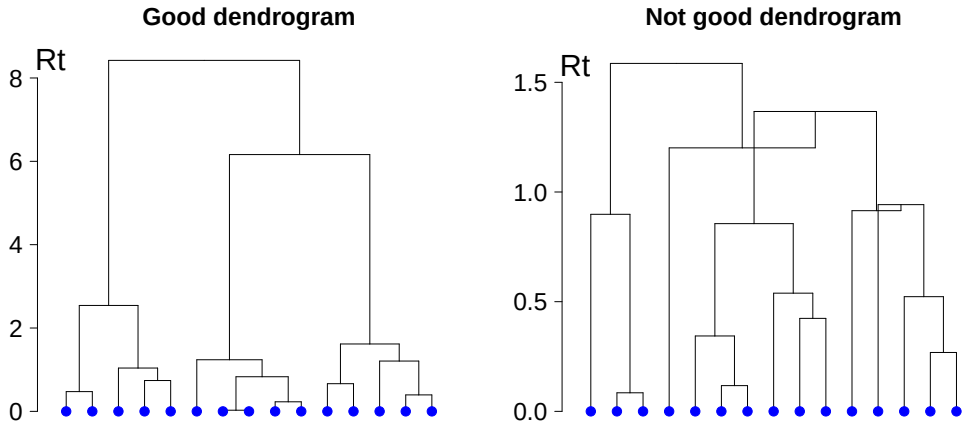
4. слить их в один кластер:
$W = U \cup V$;
$C_t = C_{t-1} \cup W \setminus \{U,V\}$;

5. для всех $S \in C_t \setminus W$

6. вычислить $R(W,S)$;

3.11.4 Визуализация кластерной структуры

Определение. *Дендрограмма — деревоподобный график, отражающий процесс последовательных слияний и структуру кластеров.*



После построения дерева можно его разрезать на поддеревья по заданному расстоянию между кластерами и получить сами кластеры. Разрез делается там, где долго не было объединения кластеров (длинная ветка у дерева).

Но долго-недолго — это субъективно и зависит от выбранного расстояния. Если расстояние в квадрате, то дальние ветки искусственно удлиняются.