

Краткие ответы: экзамен по многомерной статистике, зима 2020

Дейвид Капаца

11 сентября 2021 г.

Содержание

1 Введение	3
2 Краткие ответы	3
2.1 Факторный анализ. Модель данных и модель для ковариационной матрицы	3
2.2 Параметры в факторном анализе, их число, условие на корректность модели	4
2.3 Общность и уникальность	4
2.4 Оценка параметров. Что минимизируется в факторном анализе (отличие от анализа главных компонент)	5
2.5 Вращение в факторном анализе. Для чего нужны, пример. Метод varimax .	6
2.6 Общий подход к классификации через апостериорные вероятности	6
2.7 Какая ошибка минимизируется в подходе через максимизацию апостериорных вероятностей? Каким априорным весам соответствует доля неправильных классификаций в матрице классификации?	7
2.8 Линейный дискриминантный анализ. Модель. Классифицирующие функции.	7
2.9 Канонические переменные, их смысл. Значимость LDA	8
2.10 Почему линейный дискриминантный анализ называется линейным, а квадратичный – квадратичным?	9
2.11 Две группы, граница между двумя классами. Что происходит с границей при изменении априорных вероятностей (в общем случае по смыслу и на примере ЛДА по формулам)?	9
2.12 Как проверяют качество построенной классифицирующей процедуры (cross-validation)?	9
2.13 Что такое ROC-кривая и AUC, для чего используются? Связь с ошибками 1 и 2 рода. Пример построения ROC	10
2.14 Разные модели для классов в ДА, число параметров в моделях и возможный overfitting	10
2.15 Кластерный анализ, пример model-based подхода, вид функции правдоподобия	11
2.16 Кластерный анализ (partitioning): k-means (целевая функция, алгоритм, свойства, какие предположения о кластерах)	11
2.17 Кластерный анализ иерархический. Расстояния между точками и между кластерами. Разница между complete и single linkage	12

1 Введение

Данный документ был создан на основе лекций Голяндиной Н.Э. по многомерной статистике. Здесь будут собраны краткие ответы на вопросы по курсу, не касающиеся части по анализу главных компонент.¹ Отметим, что изложение краткое и в некотором смысле специфически авторское, поэтому используйте на свой страх и риск.

2 Краткие ответы

2.1 Факторный анализ. Модель данных и модель для ковариационной матрицы

В отличие от АГК здесь мы идём от случайных величин. Рассматриваем случайные векторы $\boldsymbol{\xi} = (\xi_1, \dots, \xi_p)^T$ и $\boldsymbol{\eta} = (\eta_1, \dots, \eta_r)^T$. По сути, $\boldsymbol{\xi}$ — старые признаки, $\boldsymbol{\eta}$ — новые признаки. $E\xi_i = 0$, $E\eta_i = 0$, $D\eta_i = 1$, η_i и η_j некоррелированы. Модель для факторного анализа выглядит следующим образом:

$$\boldsymbol{\xi} = \mathbb{F}_r \boldsymbol{\eta} + \boldsymbol{\varepsilon},$$

где $\mathbb{F}_r \in \mathbb{R}^{p \times r}$ — матрица факторных нагрузок ранга r , ковариационная матрица $\text{Cov}(\boldsymbol{\varepsilon}) = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, и при этом $\boldsymbol{\varepsilon}$ и $\boldsymbol{\eta}$ некоррелированы.

Перейдём к выборке:

- $\boldsymbol{\xi}$ соответствует $\mathbf{X}^T \in \mathbb{R}^{p \times n}$;
- $\boldsymbol{\eta}$ соответствует \mathbf{V}_r^T — скрытые признаки (их надо найти);
- \mathbb{E} — остаток.

так что получаем (транспонируем)

$$\mathbf{X} = \mathbf{V}_r \mathbb{F}_r^T + \mathbb{E}.$$

Модель можно переписать с помощью ковариаций:

$$\Sigma = \text{Cov}(\boldsymbol{\xi}) = \text{Cov}(\mathbb{F}_r \boldsymbol{\eta}) + \text{Cov}(\boldsymbol{\varepsilon}) = \mathbb{F}_r \mathbb{F}_r^T + \Psi,$$

где $\Psi = \text{diag}(\sigma_1^2, \dots, \sigma_p^2)$, $\text{rank} \mathbb{F}_r \mathbb{F}_r^T \leq r$. Если как данные вместо Σ у нас есть выборочная ковариационная матрица $\frac{\mathbf{X}^T \mathbf{X}}{n}$, то предположение заключается в том, что выборочная ковариационная матрица может быть разложена в сумму симметричной неотрицательно определённой матрицей ранга меньше r и некоторой диагональной матрицы.

(solved) Почему для факторного анализа сохраняется $f_{ij} = \rho(\xi_i, \eta_j)$?

Было в АГК: $\mathbf{X}^T = \sum_{i=1}^d F_i V_i^T$. Здесь всё то же самое. У нас ведь есть ортонормированность у \mathbf{V}_r^T , чтобы всё было так же?

Факторный анализ всегда делается на основе стандартизированных данных, поэтому у нас всегда имеет место модель корреляционной матрицы. В таком случае $f_{ij} = \rho(\xi_i, \eta_j)$. Модель может быть расписана построчно:

$$\begin{aligned}\xi_1 &= f_{11}\eta_1 + \dots + f_{1r}\eta_r + (\varepsilon_1 + 0\varepsilon_2 + \dots + 0\varepsilon_p) \\ \xi_2 &= f_{21}\eta_1 + \dots + f_{2r}\eta_r + (0\varepsilon_1 + \varepsilon_2 + \dots + 0\varepsilon_p) \\ &\dots \\ \xi_p &= f_{p1}\eta_1 + \dots + f_{pr}\eta_r + (0\varepsilon_1 + 0\varepsilon_2 + \dots + \varepsilon_p)\end{aligned}$$

¹для этого смотрите другой конспект

Предполагаем, что в каждом столбце и строчке перед ε_i только один из коэффициентов не равен нулю; это как фактор, специфический для одного признака. В то же время предполагаем, что из f_{ij} в каждом столбце хотя бы два ненулевых.

С первым предположением понятно. Можно ли в предыдущих рассуждениях увидеть предпосылку появления второго предположения? Вроде такого ограничения мы не делали.

2.2 Параметры в факторном анализе, их число, условие на корректность модели

Если предположить, что мы как-то нашли матрицу \mathbb{F}_r , то несложно показать, что любое вращение этой матрицы тоже будет решением. Поэтому возникает проблема с единственностью решения. Чтобы убрать свободу (у ортогональной матрицы $\frac{r(r-1)}{2}$), можно добавить столько же ограничений, чтобы получить однозначную модель.

Для этого предлагается наложить следующее ограничение: матрица $\mathbb{F}_r^T \mathbb{S} \mathbb{F}_r$ — диагональная (\mathbb{S} — выборочная ковариационная матрица). Ясно, что ввиду симметричности мы накладываем как раз необходимое число ограничений — $\frac{r(r-1)}{2}$ (свобода остаётся только в выборе диагонали).

Теперь можем посчитать сколько параметров есть в нашей модели $\Sigma = \mathbb{F}_r \mathbb{F}_r^T + \Psi$.

$$\underbrace{pr}_{\text{для } \mathbb{F}_r} + \underbrace{p}_{\text{для } \Psi} - \underbrace{\frac{r(r-1)}{2}}_{\text{введённые ограничения}}$$

Теперь, ввиду того, что ковариационная матрица задаётся $\frac{p(p+1)}{2}$ элементами, у нас имеется столько же равенств. Для того, чтобы обеспечить разрешимость, хотим чтобы число параметров не превосходило число уравнений, то есть

$$pr + p - \frac{r(r-1)}{2} \leq \frac{p(p+1)}{2},$$

откуда

$$\frac{(p-r)^2 - (p+r)}{2} \geq 0.$$

Это неравенство помогает определить ограничения на количество факторных переменных и даже потом используется при проверке значимости модели.

2.3 Общность и уникальность

Обозначаем $\xi' = \mathbb{F}_r \eta$, откуда из модели $\xi = \xi' + \varepsilon$. По определению $D\xi'_i = 1 - \sigma_i^2$. Также, $\rho_{ij} = \text{Cov}(\xi_i, \xi_j) = \text{Cov}(\xi'_i + \varepsilon_i, \xi'_j + \varepsilon_j) = \text{Cov}(\xi'_i, \xi'_j)$. Таким образом,

$$\text{Cov}(\xi') = \mathbb{F}_r \text{Cov}(\eta) \mathbb{F}_r^T = \mathbb{F}_r \mathbb{F}_r^T.$$

То есть,

$$\mathbb{F}_r \mathbb{F}_r^T = \text{Cov}(\xi') = \begin{pmatrix} 1 - \sigma_1^2 & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & 1 - \sigma_p^2 \end{pmatrix}.$$

Мы получили, что

$$1 = D\xi_i = D\xi'_i + D\varepsilon_i = \underbrace{(1 - \sigma_i^2)}_{\text{общность}} + \underbrace{\sigma_i^2}_{\text{уникальность}}.$$

Также можно убедиться, что общность является множественным коэффициентом корреляции.

Общность — это то, что описывается факторами, а уникальность — то, что не описывается факторами.

2.4 Оценка параметров. Что минимизируется в факторном анализе (отличие от анализа главных компонент)

Уже поняли, что задача состоит в оценивании параметров для нашей модели. Есть несколько способов.

MINRES

По сути, мы решаем задачу МНК:

$$\|\mathbb{S} - \tilde{\mathbb{S}}\|_2^2 \rightarrow \min_{\tilde{\mathbb{S}}: \tilde{\mathbb{S}} = \tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T + \tilde{\Psi}}.$$

Алгоритм OLS для ковариационной матрицы такой:

1. Решаем задачу

$$\sum_{i < j \leq p} (s_{ij} - \tilde{s}_{ij})^2 \rightarrow \min_{\tilde{\mathbb{S}}: \tilde{\mathbb{S}} = \tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T};$$

2. Находим $\hat{\mathbb{F}}_r$;

3. σ_i^2 определяем из диагонали $\hat{\mathbb{F}}_r \hat{\mathbb{F}}_r^T$.

Взвешенный МНК (WLS)

Здесь особенность в том, что чем меньше уникальность, тем больше все слагаемого. Решается задача:

$$\sum_{i,j=1}^p \frac{(s_{ij} - \tilde{s}_{ij})^2}{\hat{\sigma}_i^2 \hat{\sigma}_j^2} \rightarrow \min_{\tilde{\mathbb{S}}: \tilde{\mathbb{S}} = \tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T + \tilde{\Psi}}.$$

Метод максимального правдоподобия (MLE)

Выписывается распределение ковариационной матрицы данных в модели факторного анализа и оптимизацией находятся параметры.

Отличие от АГК

Главное отличие — в наличии модели. Ещё одно — в оптимизационных задачах:

- Задача факторного анализа

$$\sum_{i < j \leq p} (s_{ij} - \tilde{s}_{ij})^2 \rightarrow \min_{\tilde{\mathbb{S}}: \tilde{\mathbb{S}} = \tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T};$$

- Задача АГК

$$\sum_{i,j \leq p} (s_{ij} - \tilde{s}_{ij})^2 \rightarrow \min_{\tilde{\mathbb{S}}: \tilde{\mathbb{S}} = \tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T}.$$

2.5 Вращение в факторном анализе. Для чего нужны, пример. Метод varimax

Уже выяснили, что любое вращение матрицы \mathbb{F}_r тоже будет решением. Пусть \mathbf{W} — некоторая ортогональная матрица, $\tilde{\mathbb{F}}_r = \mathbb{F}_r \mathbf{W}$. Тогда $\tilde{\mathbb{F}}_r \tilde{\mathbb{F}}_r^T = \mathbb{F}_r \mathbb{F}_r^T$.

На выборочном языке: пусть \mathbf{X}' — та часть данных, которая описывается факторами; тогда

$$\mathbf{X}' = \mathbf{V} \mathbb{F}_r^T = \underbrace{\mathbf{V} \mathbf{W}}_{\tilde{\mathbf{V}}} \underbrace{\mathbf{W}^T \mathbb{F}_r^T}_{\tilde{\mathbb{F}}_r^T}.$$

Тогда $\tilde{\mathbf{V}}$ — новые факторы, $\tilde{\mathbb{F}}_r$ — новые факторные нагрузки. Хотелось бы найти такую \mathbf{W} , чтобы интерпретация была наиболее простой. Хороший вариант: больше нулей в столбцах матрицы $\tilde{\mathbb{F}}_r$.

Метод Varimax

Задача такая: увеличить разброс между квадратами f_{ij} .

$$\sum_{i=1}^r \left[\frac{1}{p} \sum_{j=1}^p \tilde{f}_{ij}^2 - \left(\frac{1}{p} \sum_{j=1}^p \tilde{f}_{ij}^2 \right)^2 \right] \rightarrow \max_{\mathbf{W}: \tilde{\mathbb{F}}_r = \mathbb{F}_r \mathbf{W}}$$

Косоугольные вращения

TODO: дописать

2.6 Общий подход к классификации через апостериорные вероятности

Вообще общий подход такой: строим классифицирующие функции f_i , которые отражают степень принадлежности к различным характеристикам/группам A_i и классифицируем индивида x по правилу

$$\hat{A} = \arg \max_i f_i(x).$$

Рассмотрим дискретную случайную величину ξ , которая принимает значения $\{A_i\}_{i=1}^k$ и имеет условные распределения $\mathcal{P}_i = \mathcal{P}(\eta \mid \xi = A_i)$ и соответствующие им условные плотности $p_i(x)$ для каждого A_i .

η в нашем случае можно воспринимать как некоторые другие характеристики индивида? Типа распределение роста при условии мужского/женского пола

Эти условные плотности можно оценить параметрическими или непараметрическими способами.

Подход через апостериорные вероятности Нам может помочь некоторый набор наблюдений, для которых заведомо известна их классификация. Введём событие $C_i := \{\xi = A_i\}$. Тогда $P(C_i) =: \pi_i$ будет означать вероятность принадлежности классу A_i для нового наблюдения. Задача классификации нового объекта x в таком случае имеет следующий вид

$$\hat{A} = \arg \max_i P(C_i \mid \eta = x) = \arg \max_i \frac{\pi_i P(x \mid C_i)}{\sum_{j=1}^k \pi_j P(x \mid C_j)} \stackrel{\substack{\text{замена на} \\ \text{плотность } p_i(x) \\ \text{+ знаменатель} \\ \text{одинаковый}}}{=} \arg \max_i \pi_i p_i(x).$$

Априорные вероятности π_i выбираются несколькими способами:

- $\pi_i = 1/k$ для k разных классов;
- $\pi_i = \frac{n_i}{N}$, где n_i — размер класса A_i в обучающей выборке, N — общий размер обучающей выборки;
- некоторые другие данные (заранее известные результаты).

Отметим, что такой метод классификации обладает свойством минимизации средней апостериорной ошибки:

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

Кстати, отсюда видим, что задание вероятностей π_i имеет смысл весов, которые определяют важность ошибочной классификации для разных классов.

2.7 Какая ошибка минимизируется в подходе через максимизацию апостериорных вероятностей? Каким априорным весам соответствует доля неправильных классификаций в матрице классификации?

Минимизируется такая ошибка:

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

Отсюда видим, что задание вероятностей π_i имеет смысл весов, которые определяют важность ошибочной классификации для разных классов.

Каким априорным весам соответствует доля неправильных классификаций в матрице классификации?

(я бы ответил, что если задать π_i маленьким, то доля классификаций i -го класса как других будет увеличиваться)

2.8 Линейный дискриминантный анализ. Модель. Классифицирующие функции.

В данном случае появляется модель. Пусть как и было, ξ — с.в., которая принимает значения из $\{A_i\}_{i=1}^k$, однако теперь нам известна модель для условного распределения:

$$\mathcal{P}\{\eta \mid \xi = A_i\} = N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}),$$

то есть предполагаем, что распределения для групп нормальные, причём для всех групп вид ковариационных матриц у распределений такой же, есть только отличие в средних. Вспоминая плотность многомерного нормального распределения

$$p_i(\mathbf{x}) = p(\mathbf{x} \mid \xi = A_i) = \frac{1}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2}} \exp \left(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) \right),$$

можем записать классифицирующую функцию $f_i(\mathbf{x}) = \pi_i p_i(\mathbf{x})$, предварительно её прологарифмировав;

$$g_i(\mathbf{x}) = \log f_i(\mathbf{x}) = \log \pi_i - \frac{p}{2} \log 2\pi - \frac{1}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} (\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}_i).$$

Вроде как ошибку нашёл в выводе при логарифмировании в оригинале ($\log 2\pi$)

Можем убрать те слагаемые, которые не зависят от номера класса; получим линейные классифицирующие функции:

$$h_i(\mathbf{x}) = \log \pi_i - \frac{1}{2} \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_i + \boldsymbol{\mu}_i^T \boldsymbol{\Sigma}^{-1} \mathbf{x}.$$

Осмысленность LDA Очевидно, что классификация имеет смысл только тогда, когда многомерные средние отличаются. Поэтому LDA применима, когда отвергается гипотеза

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2$$

в случае двух групп. Для такого применяется критерий Хотеллинга. Если же групп несколько, применяются критерии Wilk's lambda и Roy's greatest root, как в MANOVA.

2.9 Канонические переменные, их смысл. Значимость LDA

Словесно так: строим новые признаки как линейную комбинацию старых, только таким образом, что первая каноническая переменная является такой линейной комбинацией признаков, по которой группы максимально отличаются, вторая строится как ортогональная первой и опять же обеспечивает максимальное отличие, и т.д.

Теперь с формулами: имеет место многомерный вариант разложения суммы квадратов (разложение дисперсии)

$$\sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}})(\mathbf{y}_{ij} - \bar{\mathbf{y}})^T = \underbrace{\sum_{i=1}^k n_i (\bar{\mathbf{y}}_i - \bar{\mathbf{y}})(\bar{\mathbf{y}}_i - \bar{\mathbf{y}})^T}_{\mathbf{H}} + \underbrace{\sum_{i=1}^k \sum_{j=1}^{n_i} (\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)(\mathbf{y}_{ij} - \bar{\mathbf{y}}_i)^T}_{\mathbf{E}}$$

\mathbf{E} есть сумма ковариационных матриц отдельных групп, по сути, сигма.

Почему говорим, что это разложение выборочной матрицы? Уточняя: \mathbf{y}_{ij} — это получается j -й вектор i -й группы, $\bar{\mathbf{y}}_i$ — вектор средних для i -й группы, $\bar{\mathbf{y}}$ — вектор средних для всех групп?

Канонические переменные являются собственными векторами матрицы $\mathbf{E}^{-1}\mathbf{H}$ а собственные числа этой матрицы отражают степень разделимости групп по соответствующей переменной. Число ненулевых собственных чисел $\leq \min(n, k-1)$.

Почему $\mathbf{E}^{-1}\mathbf{H}$?

Два критерия для проверки гипотезы о том, что группы не разделимы:

- **Критерий Вилкса**

Статистика критерия такая:

$$\Lambda = \prod_{i=1}^s \frac{1}{1 + \lambda_i}.$$

Неразделимость означает, что λ_i все маленькие, а тогда Λ будет большим числом. То есть критическая область будет справа, 1 есть идеальное значение. Данный критерий будет мощнее против такого расположения группы, где центры облаков точек не лежат на одной прямой.

- Критерий Roy's greatest root

Статистика критерия такая:

$$r_1^2 = \frac{\lambda_1}{1 + \lambda_1}.$$

Неразделимость означает, что λ_1 маленькая, а тогда r_1^2 будет малым числом. То есть критическая область будет слева, 0 есть идеальное значение. Данный критерий будет мощнее против такого расположения группы, где центры облаков точек лежат на одной прямой.

Для случая двух групп они совпадают.

2.10 Почему линейный дискриминантный анализ называется линейным, а квадратичный – квадратичным?

Ответ простой или что-то нужно ещё сказать?

В LDA границы задаются линейными функциями, в случае QDA разделяющая поверхность задаётся квадратичными функциями (эллипсоид, гиперboloид и т.п.).

2.11 Две группы, граница между двумя классами. Что происходит с границей при изменении априорных вероятностей (в общем случае по смыслу и на примере ЛДА по формулам)?

В общем смысле при задании априорных вероятностей происходит сдвиг границ классификации.

LDA Если есть только два класса, можем построить разделяющую гиперплоскость. Приравняем h_1 и h_2 , посмотрим какое уравнение будет задавать гиперплоскость:

$$\begin{aligned} h_1(\mathbf{x}) &= \log \pi_1 - \frac{1}{2} \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_1 + \boldsymbol{\mu}_1^T \boldsymbol{\Sigma}^{-1} \mathbf{x}; \\ h_2(\mathbf{x}) &= \log \pi_2 - \frac{1}{2} \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}_2 + \boldsymbol{\mu}_2^T \boldsymbol{\Sigma}^{-1} \mathbf{x}; \\ \{\mathbf{x} : h_1(\mathbf{x}) - h_2(\mathbf{x})\} &= \left\{ \mathbf{x} : \log \frac{\pi_1}{\pi_2} - \frac{1}{2} (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_2) + (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)^T \boldsymbol{\Sigma}^{-1} \mathbf{x} = 0 \right\}. \end{aligned}$$

Имеем линейную функцию от \mathbf{x} . Заданием π_1 и π_2 можно сдвигать эту разделяющую гиперплоскость в зависимости от наших приоритетов/дополнительных данных.

2.12 Как проверяют качество построенной классифицирующей процедуры (cross-validation)?

Нужно немного сказать про оценку качества проводимой классификации. Удобный способ — построение матрицы классификации, в которой элемент $[i, j]$ есть число объектов класса i , отнесённых к классу j .

Существуют различные варианты кросс-валидации.

LOOCV (leave one out CV) Для каждого индивида делается так: классифицирующая модель строится без него, затем индивид классифицируется по соответствующему правилу.

Разделение на валидационную и тренировочную выборки Делим выборки на две части случайным образом (пропорции разбиения выбираются в зависимости от обстоятельств). Обычно на большем множестве происходит тренировка классификационной модели, а на малом «валидация».

Проверка на переобучение: имеет смысл сравнить матрицы классификации по всей выборке и по результатам кросс-валидации.

2.13 Что такое ROC-кривая и AUC, для чего используются? Связь с ошибками 1 и 2 рода. Пример построения ROC

Вспомним из прошлого (проверка гипотез):

- **True positive** H_0 отвергается верно («что-то обнаружили» — *positive*);
- **True negative** H_0 не отвергается верно;
- **False positive** H_0 отвергается неверно;
- **False negative** H_0 не отвергается неверно.

Для бинарной классификации ситуация похожая, true/false отвечает за верность классификации, positive/negative отвечает за наличие/отсутствие некоторого признака (классификация же бинарная, можем закодировать 0 или 1)

TPR и FPR Существуют такие оценки качества *бинарной* классификации:

- **TPR, true positive rate** (*чувствительность* классификации)

$$\text{TPR} = \frac{\# \text{TP}}{\# \text{TP} + \# \text{FN}} = \frac{\# \text{ верно классифицированных признаком}}{\# \text{ всех с признаком}}$$

- **FPR, false positive rate**

$$\text{FPR} = \frac{\# \text{FP}}{\# \text{FP} + \# \text{TN}} = \frac{\# \text{ неверно классифицированных признаком}}{\# \text{ всех без признака}}$$

При этом величина $1 - \text{FPR}$ называется *специфичностью*.

ROC-кривая и AUC Если по оси x отложить **FPR**, а по оси y отложить **TPR**, то при варьировании параметров классификации, например π_1 и π_2 , получим ROC-кривую. Площадь подграфика этой кривой называется AUC.

Если продолжать линию с аналогией про проверку гипотез, то **FPR** есть аналог ошибки первого рода, а **TPR** есть аналог мощности. Изменение параметров классификации аналогично изменению уровня значимости.

2.14 Разные модели для классов в ДА, число параметров в моделях и возможный overfitting

Разные модели для классов в ДА

Это про то, что есть LDA, а есть QDA?

Число параметров в моделях

Мы учитываем в подсчётах априорные параметры? - не помню

Для примера предположим, что у нас p признаков и модель предполагает k различных групп.

LDA В случае LDA нужно оценить только векторы средних (pk параметров) и одну матрицу ковариаций ($\frac{p(p+1)}{2}$ параметров).

QDA Для модели QDA помимо средних (pk параметров) нужно оценить k ковариационных матриц ($k \cdot \frac{p(p+1)}{2}$ параметров). Это уже достаточно много параметров для оценки, что может негативно повлиять на дисперсию оценок.²

Возможный overfitting

Это про что?

2.15 Кластерный анализ, пример model-based подхода, вид функции правдоподобия

Основные идеи:

- Цель: разбить индивидов на такие группы, что между группами расстояние больше, чем внутри;
- «обучение без учителя» — невозможность формальной проверки правильности результата;
- Методы кластеризации можно разбить на две группы:
 1. model-based clustering (предполагаем модель, можно использовать метод максимального правдоподобия для нахождения параметров);
 2. остальные (эвристические) методы.

Относительно model-based подхода: например, предположили, что наша выборка есть смесь k нормальных распределений. Тогда плотность имеет вид

$$p(x; \boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \pi_1 p(x, \mu_1, \Sigma_1) + \dots + \pi_k p(x, \mu_k, \Sigma_k).$$

Функция правдоподобия для такой плотности имеет сложный вид для нахождения оптимальных параметров, поэтому поиск максимума происходит с помощью ЕМ-алгоритма.

2.16 Кластерный анализ (partitioning): k-means (целевая функция, алгоритм, свойства, какие предположения о кластерах)

Пусть \mathbf{x}_j , $j = 1 : N$ — индивиды. Делаем предположение о количестве классов k . Обозначим за C_k непересекающиеся множества индексов $1 : N$. Будем говорить, что индивид с индексом j принадлежит классу C_i , если в нём найдётся элемент j . По определению μ_i — средние для класса C_i .

²в особенности при небольших объёмах выборки

Целевая функция для алгоритма k-means имеет следующий вид:

$$J(\{C_j\}, \{\mu_j\}) = \sum_{i=1}^k \sum_{j \in C_i} \|x_j - \mu_i\|^2.$$

Ясно, что хотим его минимизировать. Имеется следующий вариант реализации алгоритма **k-means**:

1. Выбор μ_i случайным образом;
2. Составление классов C_i на основе близости x_i к μ_i (наиближайшие m_i по выбранной метрике);
3. Переподсчёт μ_i как средних для построенных классов;
4. Повторение п.2 и п.3 до сходимости алгоритма.

Случайный выбор μ_i может привести к неудовлетворительным результатам: у целевой функции может быть много локальных минимумов. Поэтому есть алгоритм инициализации μ_i , модифицирующий п.1 алгоритма k-means:

1. Случайным образом выбираем μ_1 ;
2. Считаем расстояния от всех точек до ближайших центров $\mu_i, \{\rho_j\}$. Выбираем x_j как следующий центр с вероятностью, пропорциональной ρ_j ;
3. Повторяем, пока не получим k центров.

Некоторые свойства

- Пусть \hat{J} — значение функционала в результате исполнения алгоритма. Известно, что при некоторых условиях на форму кластеров выполняется следующее

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

то есть результат в среднем должен быть достаточно близким к настоящему минимуму.

- Если применить к данным АГК, то при некоторых условиях пространство, натянутое на первые $k - 1$ главных компонент, будет близким к пространству, которое проходит через центры кластеров. Поэтому иногда делают АГК и потом применяют кластеризацию уже на новых признаках.

2.17 Кластерный анализ иерархический. Расстояния между точками и между кластерами. Разница между complete и single linkage

В отличие от других методов, здесь изначально не предполагается никакая структура. Не нужно, например, задавать количество кластеров. (Дальше некоторый разговор про различные метрики между точками в пространстве, а также про примеры межкластерных расстояний)

Примеры межкластерных расстояний

Алгоритм

Дендрограмма