

Многомерный анализ данных, 5/9, (2015/2016)

Голяндина Н.Э.,

Очень подробный список вопросов

Часть I.

1. Многомерное нормальное распределение. Вектор мат.ож. и ковар.матрица при лин. преобразовании (умножении на матрицу).
2. Оценки вектора средних и ковар.матрицы. Несмещенная оценка ковар. матрицы
3. Распределение вектора средних.
4. Переход к новым признакам с помощью ортогональной матрицы. Пример про способности по математике и физике (выписать матрицу вращения).
5. Разложение матрицы данных при переходе к новым признакам в виде суммы и в матричном виде.
6. Как определяется вклад новых признаков.
7. Сингулярное разложение, как строится.
8. Сингулярное разложение. В каком смысле оно единственно.
9. Разложение Шмидта.
10. Выборочный анализ главных компонент и сингулярное разложение, общее и различия.
11. Анализ главных компонент на генеральном языке как частный случай разложения Шмидта
12. Почему главные компоненты так называются, в каком смысле они главные.
13. Оптимальность сингулярного разложения в смысле аппроксимации матрицей ранга r
14. Оптимальность сингулярного разложения в смысле аппроксимации подпространством размерности r
15. Оптимальность в анализе главных компонент в статистической терминологии (через дисперсии).
16. Оптимизация в АГК в терминах ковариационных матриц.
17. В двух статистических пакетах получились разные главные компоненты. Отчего так могло получиться?
18. Смысл первой ГК, если все ковариации (корреляции) исходных признаков положительны.
19. Разница между АГК по корреляционной и по ковариационной матрице на примере двух признаков. Когда что использовать.
20. Способы выбора числа главных компонент.
21. Почему доля собственного числа по отношению к сумме собственных чисел называется объясненной долей общей дисперсии?
22. На основе каких элементов сингулярного разложения интерпретируются главные компоненты как линейные комбинации исходных признаков? Привести формулу и пример.
23. АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков. Координаты в новых базисах.
24. Как выявить индивидов, которые плохо описываются плоскостью первых двух главных компонент?
25. Как вычислить значения главных компонент для индивида, которого не было в исходной выборке. А как вычислить значения факторных значений?
26. В каком случае координаты в ортонормированном базисе можно назвать корреляциями?
27. Чему равны суммы по строкам и по столбцам в матрице, составленной из собственных векторов в АГК?
28. Чему равны суммы по строкам и по столбцам в матрице факторных нагрузок в АГК?
29. Как интерпретировать скалярное произведение строк в матрице факторных нагрузок в АГК?
30. Как нарисовать исходные орты в плоскости первых двух главных компонент?
31. Зачем и когда первые две координаты факторных нагрузок рисуются в единичном круге?
32. Чему равна норма i -го вектора из главных компонент?
33. Как формализовать веса для признаков и для индивидов в АГК?
34. Какова модель в факторном анализе?
35. Что делает АГК в модели факторного анализа при равных общностях?
36. Какая разница между АГК и факторным анализом?
37. Связь между числом факторов и числом признаков для корректности задачи.
38. Что минимизируется в методе MINRES? В чем разница с тем, что минимизируется в АГК?
39. Какой вид имеет функция правдоподобия в ФА?
40. Проверка значимости модели ФА.
41. Критерий сферичности Бартлетта, для чего нужен.
42. Что такое общность и уникальность признака? Какие факторы не находит факторный анализ?
43. Общность как множественный коэффициент корреляции.
44. Как интерпретируются признаки в ФА?
45. Зачем нужны вращения в ФА? Как устроены ортогональные вращения?
46. Вращение по методу varimax.
47. Методы нахождения факторных значений, LS и WLS (метод Бартлетта).
48. Факторная структура (корреляции исходных признаков с факторами) и факторный паттерн (коэффициенты лин. комбинации, с которыми исходные признаки выражаются через факторы) в случае ортогональных и не-ортогональных факторов.

Часть II.

1. Распределение Уишарта, свойства
2. Pooled covariance matrix.
3. Распределение Hotelling'a, свойства.
4. Проверка гипотезы о значении вектора средних, одномерный и многомерный случай.
5. Разница между доверительной областью в виде квадрата (прямоугольника) и эллипсоида.
6. Проверка гипотезы о сравнении многомерных мат. ожиданий, независимые выборки.
7. Для чего используется статистика Box's M?
8. Проверка гипотезы о равенстве нескольких средних (Repeated ANOVA). Контрасты, как их выбирать.
9. Для T^2 критериев: предположения; что происходит и что делать, если они не выполняются.
10. Единый подход к множественной регрессии и одномерному однофакторному дисперсионному анализу: ANOVA
11. Представление одномерного однофакторного дисперсионного анализа в виде множественной регрессии с фиктивными переменными.
12. Корреляционное отношение с дискретным одномерным признаком и множеств. коэффициент корреляции.
13. Обобщенная задача на собственные значения.
14. Распределение Лямбда Уилкса. Частный случай $p=1$.
15. MANOVA: модель, запись через условные мат. ожидания η и мат. ожидания η_k . Разложение ковариационной матрицы.
16. MANOVA для дискр. анализа и для многомерной множественной регрессии, общее и различие.
17. Какой смысл у канонических дискриминантных функций (коэффициентов) и переменных?
18. Как вычисляются канонические дискриминантные функции (коэффициенты)?
19. Значимость LDA. Разные критерии, чем отличаются.
20. Максимальное число дискр. функций, почему такое?
21. С чем совпадают дискриминантные функции и переменные, если ошибки сферические?
22. Как определить значимое число дискриминантных функций или, что то же самое, размерность пространства, где группы различаются.
23. Почему канонические дискриминантные переменные получаются ортогональными?
24. Интерпретация разделения: стандартизованные дискр. функции и факторная структура.
25. Свойства исходных признаков, по которым можно понять, какие признаки лишние.
26. Пошаговый дискриминантный анализ.
27. Что уменьшается с помощью Lambda-prime и что с помощью Partial Lambda?
28. Как происходит объяснение различия между группами и классификация в рамках LDA?
29. Почему линейный дискриминантный анализ называется линейным, а квадратичный – квадратичным?
30. Общий подход к классификации через апостериорные вероятности.
31. Две группы. Что происходит с границей при изменении априорных вероятностей?
32. Как проверяют качество построенной классифицирующей процедуры?
33. Как через представление средне-квадратического отклонения через дисперсию и смещение объяснить, как так бывает – модель не верна, а метод работает лучше?
34. Что такое канонические корреляции, сколько их?
35. Значимость корреляции между множествами признаков и значимость многомерной множеств. регрессии.
36. Множественная корреляция как каноническая корреляция, если число признаков с одной стороны равно 1.
37. Что такое канонические переменные, как находятся?
38. Интерпретация канонических переменных через стандартизованные канонические функции (коэффициенты) и через факторную структуру.
39. Как найти число значимых корреляционных переменных (=размерность пространства, содержащего зависимость между множествами).
40. Корреляции внутри множества канонических переменных, левых и правых.
41. Объясненные каноническими переменными доли дисперсии. Избыточность.
42. Что общего между дискриминантным анализом в многомерной множественной регрессией?
43. Две группы, использование множественной линейной регрессии для классификации.
44. Кластерный анализ, пример model-based подхода.
45. Кластерный анализ (partitioning): k-means (целевая функция, алгоритм, свойства, какие предположения о кластерах), k-means++ (начальный выбор центров).
46. Запись задачи, решаемой k-means, как задачи low-rank approximation с ограничениями. Использование PCA.
47. Кластерный анализ иерархический. Расстояния между точками и между кластерами. Разница между complete и single linkage.
48. Анализ соответствий: как устроены данные, к каким данным применяется SVD, как интерпретируется результат?