Многомерный анализ данных, 5/9, (2018/2019)

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. Почему главные компоненты так называются, в каком смысле они главные.
12. Оптимальность сингулярного разложения в смысле аппроксимации матрицей ранга r
13. Оптимальность сингулярного разложения в смысле аппроксимации подпространством размерности r
14. Оптимальность в анализе главных компонент в статистической терминологии (через дисперсии).
15. Оптимизация в АГК в терминах ковариационных матриц.
16. В двух статистических пакетах получились разные главные компоненты. Отчего так могло получиться?
17. Смысл первой ГК, если все ковариации (корреляции) исходных признаков положительны.
18. АГК по корреляционной и по ковариационной матрице. Когда что использовать.
19. Способы выбора числа главных компонент.
20. Почему доля собственного числа по отношению к сумме собственных чисел называется объясненной долей общей дисперсии?
21. На основе каких элементов сингулярного разложения интерпретируются главные компоненты как линейные комбинации исходных признаков? Привести формулу и пример.
22. АГК с точки зрения построения базиса в пространстве индивидов и в пространстве признаков. Координаты в новых базисах.
23. Как выявить индивидов, которые плохо описываются плоскостью первых двух главных компонент?
24. Как вычислить значения главных компонент для индивида, которого не было в исходной выборке. А как вычислить значения факторных значений?
25. В каком случае координаты в ортонормированном базисе можно назвать корреляциями?
26. Чему равны суммы по строкам и по столбцам в матрице, составленной из собственных векторов в АГК?
27. Чему равны суммы по строкам и по столбцам в матрице факторных нагрузок в АГК?
28. Как интерпретировать скалярное произведение строк в матрице факторных нагрузок в АГК?
29. Как нарисовать исходные орты в плоскости первых двух главных компонент?
30. Зачем и когда первые две координаты факторных нагрузок рисуются в единичном круге?
31. Чему равна норма i-го вектора из главных компонент?
32. Как формализовать веса для признаков и для индивидов в АГК?
33. Какова модель в факторном анализе?
34. Какая разница между АГК и факторным анализом?
35. Связь между числом факторов и числом признаков для корректности задачи.
36. Что минимизируется в методе MINRES? В чем разница с тем, что минимизируется в АГК?
37. Проверка соответствия модели ФА с r факторами.
38. Критерий сферичности Бартлетта, для чего нужен.
39. Что такое общность и уникальность признака? Какие факторы не находит факторный анализ?
40. Общность как множественный коэффициент корреляции.
41. Как интерпретируются признаки в ФА?
42. Зачем нужны вращения в ФА? Как устроены ортогональные вращения?
43. Вращение по методу varimax.
44. Методы нахождения факторных значений, LS и WLS (метод Бартлетта).
45. Факторная структура (корреляции исходных признаков с факторами) и факторный паттерн (коэффициенты лин. комбинации, с которыми исходные признаки выражаются через факторы) в случае ортогональных и неортогональных факторов.

Часть II.

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