

Экзамен по предмету «**Основные методы анализа данных**»
ВШЭ ФКН ПМИ, 3 курс, декабрь 2018г.
Основано на реальных событиях.

ВОПРОСЫ К ЭКЗАМЕНУ

1. Примеры различий между машинным обучением и анализом данных.

Анализ данных: использование данных для улучшения теоретического понимания предметной области.

Машинное обучение: снабжение компьютера методами и правилами для вычисления целевой переменной на основе входных данных.

Пример различий: Neural-Net \in ML - DA. Нейронная сеть подходит для предотвращения взрыва роботом, но не подходит для работы адвоката.

2. Третий закон Кеплера как пример удачного анализа данных.

Третий закон Кеплера состоит в том, что квадрат времени обращения планеты вокруг солнца пропорционален кубу расстояния этой планеты до солнца. Берётся два параметра: среднее время обращения вокруг солнца (в годах) и среднее расстояние до солнца (в средних расстояниях Земли), тогда:

$$P^2 = D^3.$$

Закон был выведен в тот момент, когда было введено понятие логарифма. Кеплер прологарифмировал два параметра:

$$2 \log(P) = 3 \log(D),$$

получим из непонятной зависимости линейную (рис. 1). Из закона Кеплера позднее был выведен Ньютоновский закон всемирного тяготения.

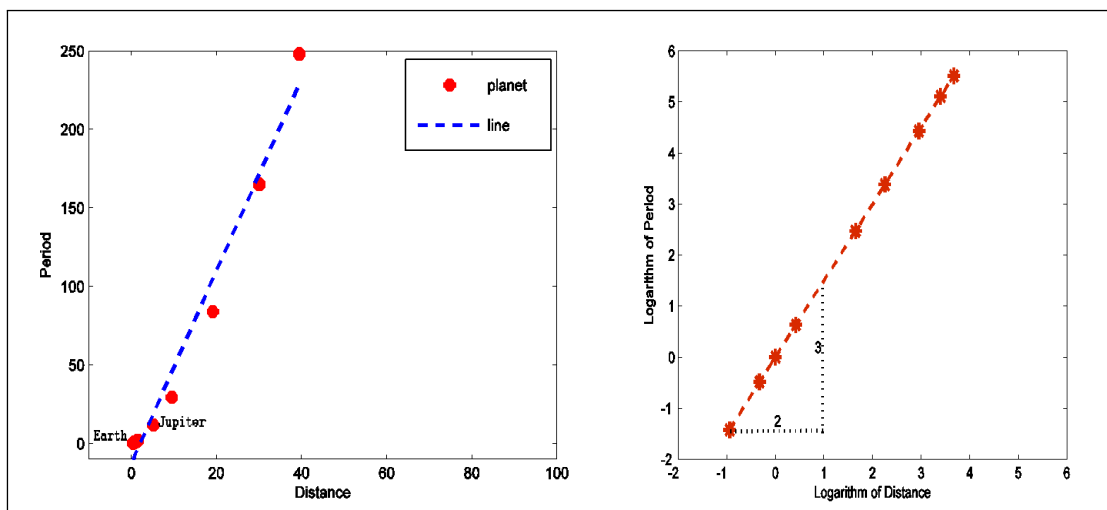


Рис. 1: зависимость в законе Кеплера

3. Пример данных и метаданных.

Пример на ирисах: матрица чисел размера 150x4 – **данные**. Названия признаков, названия таксонов, их номера – **метаданные**.

4. Признак как математическое понятие.

Признак – отображение множества объектов в множество значений признаков.

5. Понятие количественного признака.

Количественный признак – такой признак, который имеет смысл усреднять.

6. Понятие номинального признака. Количественное представление номинального признака.

Номинальный признак имеет малое число значений (которые называются категориями) и эти значения неупорядочены и альтернативны (т.е. нет ни одного объекта, который бы одновременно имел несколько значений). Категориальный признак не является номинальным (номинальный признак может быть

только один единственный для объекта).

Количественное представление: каждое значение изображается как бинарный признак. Вместо одного номинального признака рисуется столько бинарных признаков, сколько значений (их можно называть категориями). Каждой категории соответствует свой бинарный признак. Всё это дело называется **dummy variable**. Количественное представление имеет смысл, потому что для такого признака имеет смысл усреднение (можем посчитать долю данной категории).

7. Понятие гистограммы.

Гистограмма – графическое представление распределения объектов по данному признаку. Соответствует теоретическому понятию плотности распределения.

Построение: выбирается число "бинов" (bins) и весь интервал изменения признака делится на это число равных интервалов. Для каждого "бина" считаем количество объектов, которое в него попало. После этого откладываем столбики по оси ординат, пропорциональные числам в каждом "бине".

8. Метод К-средних.

Метод К-средних – метод кластерного анализа, который разбивает таблицу данных на заданное число кластеров (K) и каждый кластер представляет своим центром. Для работы метода необходимо, чтобы K начальных центров были заданы. Далее последовательно до тех пор, пока процесс не сойдётся (не выполнится критерий останова) выполняются итерации, каждая из которых состоит из двух шагов:

- обновление кластеров (вокруг центров)
- обновление центров (внутри кластеров)

Алгоритм К-средних:

- Инициализация:** пользователь выбирает число K кластеров и назначает K гипотетических центров (рис.2,а);
- Обновление кластеров:** при заданных K центрах $c_k (k = 1, 2, \dots, K)$, каждый объект $i \in I$ приписывается одному из центров по правилу минимального расстояния: вычисляются расстояния от i до каждого c_k ; объект i приписывается ближайшему центру c_k (рис.2,б). Те объекты, которые приписаны центру c_k , образуют кластер $S_k (k = 1, 2, \dots, K)$ (рис.2,в). В качестве расстояния используется квадрат Евклидова расстояния;
- Обновление центров:** вычисляется арифметический центр (центр масс) каждого кластера S_k , который и назначается новым центром $c'_k (k = 1, 2, \dots, K)$, (рис.2,г). Компоненты центра вычисляются как средние арифметические соответствующих компонент объектов из S_k ;
- Правило останова:** новые центры c'_k сравниваются со старыми. Если $c'_k = c_k$ для каждого $k = 1, 2, \dots, K$, то вычисления останавливаются и выдаются результаты: центр c'_k и кластер S_k для каждого $k = 1, 2, \dots, K$. Если же хотя бы одно из равенств не верно, то каждый центр c_k заменяется вновь полученным центром c'_k , и процесс возвращается к шагу 1 – **Обновление кластеров**;

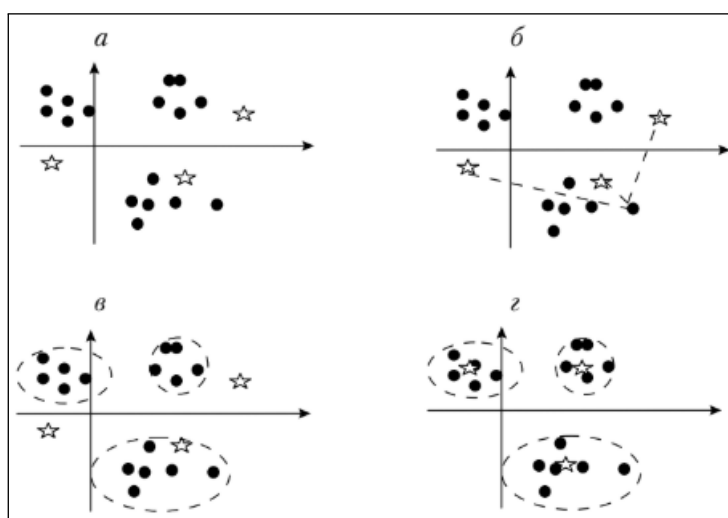


Рис. 2: итерации метода К-средних при K=3

9. Входные данные при использовании метода К-средних.

Для инициализации метода К-средних необходимо задать количество кластеров K и начальные центры $c = (c_1, c_2, \dots, c_K)$.

10. Критерий метода К-средних.

Отдельно выделим, что критерий метода не то же самое, что критерий останова.

Критерий метода К-средних – это критерий, по которому метод ищет разбиение множества объектов на K непересекающихся кластеров $S = \{S_1, S_2, \dots, S_K\}$, представленных в виде списков индексов объектов $i \in S_k$ и центрами этих кластеров (c_1, c_2, \dots, c_k) . Расстояние $d(y_i, c_k)$ высчитывается как:

$$\sum_{v \in V} (y_{iv} - c_{kv})^2.$$

Тогда **критерий метода** выглядит следующим образом:

$$W(S, c) = \sum_{k=1}^K \sum_{i \in S_k} d(y_i, c_k).$$

Полученная величина есть не что иное, как сумма квадратов Евклидовых расстояний между объектами и центрами их кластеров.

11. Разложение Пифагора и дополнительный критерий для метода К-средних.

Пусть имеются следующие данные:

$Y = (y_{iv})$ – матрица данных,

$T(Y) = \sum_{i,v} y_{iv}^2$ – разброс данных (data scatter),

$W(S, c) = \sum_{k=1}^K \sum_{i \in S_k} d(y_i, c_k)$ – критерий метода К-средних,

$F(S, c) = \sum_{k=1}^K |S_k| < c_k, c_k >$, $c_k = (c_{k1}, \dots, c_{kv})$ – сумма квадратов евклидовых расстояний от 0 до c_k , умноженных на количество элементов в k -ом кластере

Тогда **разложением Пифагора** является:

$$T(Y) = W(S, c) + F(S, c).$$

Дополнительный критерий:

$$F(S, c) \longrightarrow \max$$

Его интерпретация следующая – необходимо, чтобы объектов в кластере было как можно больше, а центры были как можно дальше.

12. Метод аномального кластера.

Данный метод находит один кластер, наиболее удалённый от нуля (реперной точки в общем случае), называемый **аномальным кластером**. Принцип действия метода практически аналогичен методу К-средних:

0. **Инициализация аномального центра:** ищется объект c (центр будущего кластера), наиболее отдалённый от 0, т.е. объект, у которого величина $< c_k, c_k >$ наибольшая;
1. **Обновление аномального кластера:** для каждого объекта y_i в цикле проверяется выполнение неравенства $d(y_i, c) < d(y_i, 0)$. Если неравенство выполняется, то мы относим данный объект к аномальному кластеру S , где $S = \{i : d(y_i, c) < d(y_i, 0)\}$;
2. **Обновление аномального центра:** вычисляется центр S : $c' = \frac{\sum_{i \in S} y_i}{|S|}$;
3. **Правило останова:** $c' \stackrel{?}{=} c$. Если нет, то переобозначаем как $c' = c$ и возвращаемся к шагу 1;
4. **Выдача результатов:** список группы S и центр c выдаются как результат работы алгоритма;

Если хочется **блеснуть**, то можно написать: величина $|S| < c, c >$ – вклад данного кластера в разброс данных, что вытекает из пифагорова разложения: $T(Y) = |S| |c|^2 + W(S, c)$. Величина $\frac{|S| < c, c >}{T}$ – относительный вклад.

13. Интеллектуальная версия метода К-средних.

Помимо прочей входной информации для алгоритма К-средних задаётся пороговое число t – минимальный размер кластера. Тогда **алгоритм иК-средних(t)** имеет следующий вид:

1. **Аномальный кластер:** выделяем аномальный кластер S_k и его центр c_k . Выбрасываем его из нашего множества данных;

2. **Условие остановки:** если в множестве ещё остались объекты, то возвращаемся к шагу 1;
3. **Отбрасывание малых кластеров:** из всей получившейся совокупности аномальных кластеров S_k выбрасываем те кластеры, которые удовлетворяют условию $|S_k| \leq t$. Обозначим количество оставшихся кластеров через K , а их центры через c_1, c_2, \dots, c_K ;
4. **Метод К-средних:** применяем метод К-средних, используя c_1, c_2, \dots, c_K в качестве начальных центров;

14. Интерпретация кластера.

Рассчитываем общий центр $c_k = (c_{k1}, c_{k2}, \dots, c_{kv})$ и общее среднее на всём множестве $g = (g_1, g_2, \dots, g_v)$ и сравниваем, чем отличается кластер от среднего по следующей формуле:

$$\frac{|c_{kv} - g_v|}{g_v}.$$

Признак будет являться важным, если для него данное отношение будет > 0.4 . В таком случае мы говорим, что данный кластер определяется признаками, которые достаточно сильно отличаются от общего среднего. Если важные признаки отсутствуют, то кластер не является интересным.

15. Использование метода бутстрэп для сравнения средних.

Бутстрэп – использование случайной выборки для увеличения вариабельности данных.

Пусть имеется 2 кластера. Сгенерируем 5000 раз признак на множестве S_1 , затем на S_2 (т.е. на каждой из 5000 итераций случайным образом возьмём N объектов из текущего признака с повторением) и вычислим средние по этим признакам, получим два вектора $(m_1^1, \dots, m_{5000}^1)$ и $(m_1^2, \dots, m_{5000}^2)$. После этого поэлементно вычтем из первого вектора второй, получив новый вектор разностей средних $(m_i^1 - m_i^2)$. Если все получившиеся значения положительные, то в первом кластере средние больше, чем во втором (и наоборот). Если часть средних положительна, а часть отрицательна, то мы не можем отбросить гипотезу о том, что эти кластеры равны. Тогда возьмём 95-процентный интервал этих разностей и посмотрим, покрывает данный интервал 0 или нет. Если покрывает, то гипотезу о том, что средние данного признака на этих двух кластерах одни и те же, мы отбросить не можем.

16. Чем отличаются методы бутстрэпа с опорой и без опоры.

Метод с опорой (pivot) основан на предположении, что распределение значений средних бутстрэпа является Гауссовым, находит оценку среднего – m_b , оценку стандартного отклонения (сигмы) – s_b и определяет confidence interval как

$$m_b \pm 1.96 \cdot s_b.$$

Метод без опоры (non-pivot): не берёт допущений о характере распределения средних значений бутстрэпа, а берёт распределение как оно есть, сортирует его, отбрасывает по 125 объектов с начала и с конца (5000 - 2.5%), и тогда значение признака на 126-м объекте – левая граница доверительного интервала, а на 4875 – правая граница.

17. Среднее арифметическое и медиана как средние Минковского.

Если задано множество чисел x_1, \dots, x_N , то величина c называется **центром Минковского**, если она минимизирует сумму:

$$\sum_i |x_i - c|^p,$$

при заданном p .

При $p = 1$ центр Минковского называется **медианой**, а при $p = 2$ – **средним**.

18. Таблица сопряженности.

Для анализа связи между двумя номинальными признаками составляют так называемые таблицы сопряженности. Возьмём две системы категорий от номинальных признаков: s и t . Строки таблицы сопряженности соответствуют одной системе категорий, столбцы – другой системе категорий. Для каждой пары (s, t) рассчитывается количество объектов, попавших одновременно и в s , и в t . Таблица этих чисел называется **таблицей сопряженности**.

Также можно дописать маргинальный столбец (N_{s+} – сумма по строке s) и маргинальную строчку (N_{+t} – сумма по столбцу t). Относительные формулы в частотах: $p_{s+} = \frac{N_{s+}}{N}$, $p_{+t} = \frac{N_{+t}}{N}$, $p_{st} = \frac{N_{st}}{N}$.

19. Условная вероятность и статистическая независимость.

Условная вероятность выражается формулой:

$$p(s|t) = \frac{N_{st}}{N_{+t}} = \frac{p_{st}}{p_{+t}}.$$

s и t **независимы статистически**, если $p(s|t) = p(s)$ или, если переписать то же самое в терминах таблицы: $\frac{p_{st}}{p_{+t}} = p_{s+} \iff p_{st} = p_{s+} \cdot p_{+t}$.

20. Коэффициент Кетле и его смысл.

Допустим, нас интересует степень связи между s и t . Тогда её можно измерить с помощью **коэффициента Кетле**, который записывается как:

$$q(s|t) = \frac{p(s|t) - p(s)}{p(s)},$$

то же самое можно переписать как

$$q_{st} = \frac{\frac{p_{st}}{p_{+t}} - p_{s+}}{p_{s+}} = \frac{p_{st}}{p_{s+} \cdot p_{+t}} - 1,$$

что есть относительный прирост вероятности категории s при условии, что второй признак принимает категорию t .

21. Суммарный коэффициент Кетле.

Суммарный коэффициент Кетле определяется следующей формулой:

$$Q = \sum_{s,t} p_{st} \cdot q_{st}$$

и показывает, на сколько в среднем изменится вероятность категории s первого признака, если категория t второго признака станет известной.

22. Коэффициент сопряженности Пирсона, его смысл.

Коэффициент сопряженности Пирсона определяется следующей формулой:

$$\chi^2 = N \cdot \sum_{s,t} \left(\frac{(p_{st} - p_{s+} \cdot p_{+t})^2}{p_{s+} \cdot p_{+t}} \right)$$

и показывает меру отклонения наблюдаемого двумерного распределения в таблице сопряженности от условия статистической независимости признаков.

23. Связь между коэффициентами Пирсона и Кетле.

Мы можем выразить коэффициент Пирсона через коэффициенты Кетле:

$$\chi^2 = N \cdot \sum_{s,t} p_{st} q_{st} = N \cdot Q,$$

что, в свою очередь, даёт нам **усредненный коэффициент Кетле**.

Таким образом, усредненный индекс Кетле, а значит и коэффициент Пирсона, показывает, на сколько в среднем относительно увеличится вероятность категории одного признака при условии, что категория другого признака становится известной.

24. Метод главных компонент на основе сингулярного разложения прямоугольных матриц.

Метод главных компонент — один из основных методов суммаризации данных. Он формирует линейные комбинации исходных признаков, вносящие максимальный вклад в разброс данных — они-то и называются главными компонентами.

Пусть $Y = (y_{iv})$ — матрица данных. Сингулярная тройка z , μ , c определяется следующими двумя условиями:
$$\begin{cases} Yc = \mu z, \\ Y^T z = \mu c \end{cases}.$$

Обозначим сингулярные значения как $\mu_1 > \mu_2 > \dots > \mu_r > 0$, где r – ранг матрицы. Им соответствуют векторы (z_1, z_2, \dots, z_r) и (c_1, c_2, \dots, c_r) . Тогда **сингулярным разложением** называется:

$$y_{iv} = \mu_1 \cdot z_{i1} \cdot c_{v1} + \mu_2 \cdot z_{i2} \cdot c_{v2} + \dots + \mu_r \cdot z_{ir} \cdot c_{vr}.$$

Метод главных компонент на основе метода сингулярного разложения представляет матрицу Y как

$$y_{iv} = z_{i1}^* \cdot c_{v1}^* + z_{i2}^* \cdot c_{v2}^* + \dots + z_{ir}^* \cdot c_{vr}^*,$$

где $z_{ij}^* = \sqrt{\mu_j} z_{ij}$ и $c_{vj}^* = \sqrt{\mu_j} c_{vj}$.

25. Вклад главной компоненты в разброс данных.

Вклад равен квадрату соответствующей сингулярной величины μ_i .

26. Визуализация данных на двумерной плоскости с помощью метода главных компонент.

Получившееся разложение на первые две главные компоненты $z_{i1}^* = \sqrt{\mu_1} z_{i1}$, $z_{i2}^* = \sqrt{\mu_2} z_{i2}$ можно отобразить как scatter plot по данным двум осям (рис. 3), даже прикрепив к каждому объекту лейбл из третьего (номинального) признака.

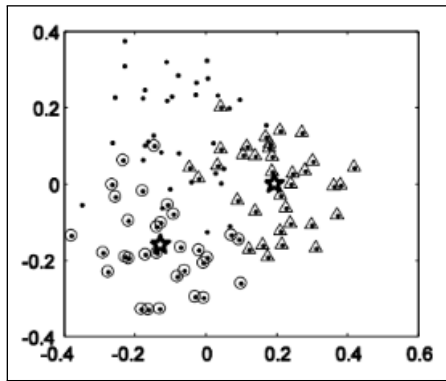


Рис. 3: scatter plot на первые две главные компоненты

27. Традиционный метод главных компонент (на основе ковариационной матрицы).

Матрица ковариации $Cov = \frac{Y^T Y}{N}$, где Y предварительно центрированная матрица данных (т.е. из каждой строки вычтен вектор среднего $Y = X - a$).

Метод заключается в поиске собственных чисел этой матрицы (максимальных) λ_1, λ_2 , соответствующих собственным векторам c_1, c_2 (где $\|c_{1,2}\|^2 = 1$) таких, что $\begin{cases} Cov \cdot c_1 = \lambda_1 c_1, \\ Cov \cdot c_2 = \lambda_2 c_2 \end{cases}$ и определении главных

$$\text{компонент} \begin{cases} z_1 = \frac{Y c_1}{\lambda_1}, \\ z_2 = \frac{Y c_2}{\lambda_2} \end{cases}.$$

28. Связь сингулярных чисел матрицы данных с собственными числами ковариационной матрицы.

Пусть $Y_{N \times V}$ – предварительно центрированная матрица данных с сингулярной тройкой $\begin{matrix} z & \mu & c \\ N \times 1 & 1 & 1 \times V \end{matrix}$,

удовлетворяющей следующим условиям (смотрим для $|\mu| = 1$): $\begin{cases} Y c = \mu z, (*) \\ Y^T z = \mu c (**) \end{cases}$.

Связь с собственными числами этой матрицы: из (*) выразим $z = \frac{Y c}{\mu}$. Подставим её в (**):

$\frac{Y^T Y c}{\mu} = \mu c$. Домножим это дело на μ : $Y^T Y c = \mu^2 c$ и заодно поделим на N :

$$\left(\frac{Y^T Y}{N} \right)_{Cov} c = \left(\frac{\mu^2}{\lambda} \right) c,$$

что следует из условия $Cov \cdot c = \lambda c$, откуда получаем $\lambda = \frac{\mu^2}{N}$, что и требовалось.

29. Отношение Райли (Рэля) и его связь со спектральным анализом.

Техники **спектрального анализа (spectral clustering)** используют собственные значения и собственные векторы для осуществления понижения размерности перед кластеризацией в меньшем количестве пространств.

Пусть $C = \frac{Y^T \cdot Y}{N}$ – матрица ковариации. **Отношением Рэля** называется выражение:

$$q(c) = \frac{c^T Y^T Y c}{c^T c} \rightarrow \min.$$

Максимум в отношении Райли достигается в собственном векторе c матрицы C , соответствующим максимальному собственному значению λ .

30. Матрица корреляции.

Пусть Y – центрированная матрица данных (т.е. $Y = X - a$, где a – среднее). Тогда матрица

$$U = \frac{Y}{\sigma(y)}$$

является нормализацией матрицы Y , а матрица

$$\frac{U^T U}{N},$$

где N – количество элементов, называется **матрицей коэффициентов корреляции**.

31. Коэффициент корреляции в вероятностной перспективе.

Оговорка: про это нигде нет, так что фиг его знает.

Коэффициент корреляции представляет собой безразмерную величину, изменяющуюся в пределах от -1 до 1:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma(x) \cdot \sigma(y)} = \frac{M[(X - M(X)) \cdot (Y - M(Y))]}{\sigma(x) \cdot \sigma(y)}$$

32. Коэффициент корреляции в аппроксимационной перспективе, его свойства.

Имеем N объектов, для каждого из которых заданы целевой признак y и входной признак x следующим образом: $((x_1, y_1), \dots, (x_N, y_N))$. Пусть имеется задача линейной регрессии $y_i = ax_i + b + e_i$, в которой необходимо минимизировать среднюю квадратичную ошибку

$$L(a, b) = \frac{\sum_i e_i^2}{N} = \frac{\sum_i (y_i - ax_i - b)^2}{N},$$

где

$$\begin{cases} a = \rho \frac{\sigma(y)}{\sigma(x)}, \\ b = m_y - am_x \end{cases}$$

следующим образом (см. 34 билет):

$$L_m(a, b) = \sigma^2(y)(1 - \rho^2).$$

Коэффициент корреляции в аппроксимационной перспективе определяется следующей формулой:

$$\rho = \frac{\sum_i (x_i - m_x)(y_i - m_y)}{N\sigma(x)\sigma(y)},$$

где m_x и m_y – средние значения x_i и y_i , а $\sigma(x)$ и $\sigma(y)$ – их стандартные отклонения и имеет следующие **свойства**:

1. Коэффициент корреляции ρ изменяется в интервале от -1 до 1. Чем ближе ρ к 1 или к -1, тем меньше остатки в линейном регрессионном уравнении (то есть признаки связаны линейным уравнением с точностью до малых ошибок). Например, величина $\rho = 0.9$ означает, что необъясненная часть дисперсии y в L_m равна $1 - \rho^2 = 1 - 0.81 = 19\%$ от исходной величины.

2. Наклон a пропорционален ρ согласно уравнению:

$$a = \rho \frac{\sigma(y)}{\sigma(x)};$$

a положительно или отрицательно в зависимости от знака ρ . Если $\rho = 0$, то наклон нулевой: в этом случае y и x называются некоррелированными;

3. $\rho = \pm 1 \iff L_m(a, b)^2 = 0$, что означает, что все точки лежат на прямой \Rightarrow чисто линейная связь;

33. Коэффициент детерминации, его смысл.

Имеем N объектов, для каждого из которых заданы целевой признак y и входной признак x следующим образом: $((x_1, y_1), \dots, (x_N, y_N))$. Пусть имеется задача линейной регрессии, в которой необходимо минимизировать среднюю квадратичную ошибку $L(a, b)$.

Коэффициент детерминации ρ^2 характеризует долю дисперсии признака y , учтенную в построенной линейной регрессии y по x , что следует из уравнения (см. 34 билет):

$$L_m(a, b) = \sigma^2(y)(1 - \rho^2).$$

34. Задача линейной регрессии и ее решение.

Задача: Имеем N объектов, для каждого из которых заданы целевой признак y и входной признак x следующим образом: $((x_1, y_1), \dots, (x_N, y_N))$. Хотим найти линейное уравнение, которое бы их связывало:

$$y = ax + b.$$

Очевидно, что наши данные не обязаны выстраиваться в одну прямую на двумерной плоскости (x, y) , поэтому каждый элемент будет иметь некоторую ошибку e_i , которая будет подгонять его под такую прямую (это будет называться уравнением линейной регрессии y по x):

$$y_i = ax_i + b + e_i, (i = 1, 2, \dots, N).$$

Явный учёт таких ошибок позволяет поставить задачу поиска коэффициентов a и b так, чтобы остатки могли быть минимизированы по критерию наименьших квадратов. Минимизируем среднюю квадратичную ошибку:

$$L(a, b) = \frac{\sum_i e_i^2}{N} = \frac{\sum_i (y_i - ax_i - b)^2}{N} \rightarrow \min,$$

где минимум будет достигаться для всех возможных a и b , а также данных x_i и y_i , ($i = 1, 2, \dots, N$).

Решение: так как $L(a, b)$ – параболическая функция с ветвями вверх, то её минимум будет находиться в точке, где частные производные по a и b равны нулю:

$$\frac{\partial L}{\partial a} = 0, \quad \frac{\partial L}{\partial b} = 0.$$

Решение можно выразить формулами для a и для b :

$$a = \rho \frac{\sigma(y)}{\sigma(x)},$$

где

$$\rho = \frac{\sum_i (x_i - m_x)(y_i - m_y)}{N\sigma(x)\sigma(y)}$$

так называемый коэффициент корреляции, m_x и m_y – средние значения x_i и y_i , а $\sigma(x)$ и $\sigma(y)$ – их стандартные отклонения.

$$b = m_y - am_x.$$

Подставив a и b , получим выражение для минимального значения критерия:

$$L_m(a, b) = \sigma^2(y)(1 - \rho^2).$$

35. Привести пример взаимосвязанных признаков с нулевым коэффициентом корреляции.

Совсем низкое или нулевое значение коэффициента корреляции не всегда означает отсутствие взаимосвязи. Речь идет об отсутствии именно линейной связи. Нулевой коэффициент корреляции может соответствовать другому, более тонкому, типу функциональной зависимости. На рис. 4 представлены три различных поля рассеяния при нулевой корреляции в данных. Только один из них, тот, что слева, на самом деле свидетельствует о том, что между x и y нет связи, т.е. знание значения одного признака никак не помогает в прогнозе значения другого. Каждый из двух других случаев показывает довольно высокую степень связи x и y . В частности, в центре — график квадратичной зависимости ($y = (x - 2)^2 + 5$), а справа — случай, когда совокупность объектов разнородна — она состоит из двух частей, таких что в каждой признаки связаны линейно, но связи взаимно противоположны ($y = 2x - 5$ и $y = -2x + 3$).

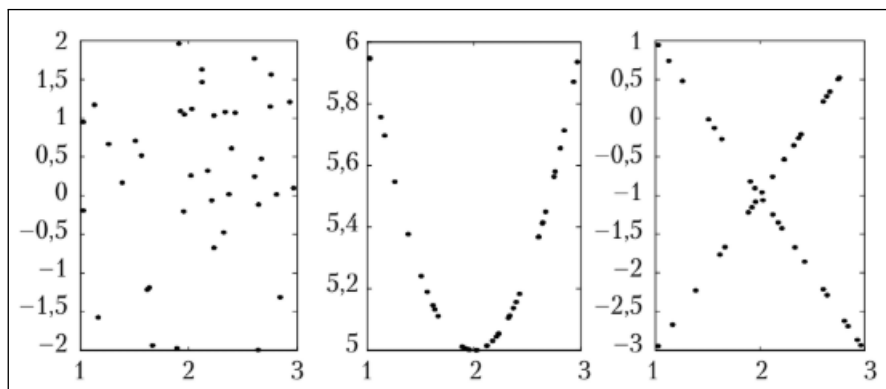


Рис. 4: поля рассеяния, соответствующие нулевому или почти нулевому значения корреляции

36. Что можно сказать о коэффициенте корреляции между ростом и весом группы мужчин из одной местности?

—

37. Корреляция между длиной и шириной чашелистика в данных Ирис отрицательна. Почему?

—

38. Можно ли значительно увеличить значение коэффициента корреляции в группе наблюдений, добавив одно-два наблюдения?

—

39. Что общего и что различного у методов дивизимного и агломеративного кластер-анализа?

Иерархическая кластеризация — совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров.

Дивизимный подход — кластерная иерархия выстраивается сверху вниз, новые кластеры создаются путём деления более крупных кластеров на более мелкие.

Агломеративный подход — кластерная иерархия выстраивается снизу вверх, новые кластеры создаются путём объединения более мелких кластеров (обычно начиная с синглтонов).

40. Расстояние Уарда между двумя кластерами и его смысл.

Пусть $S = \{S_1, S_2, \dots, S_K\}$, $d(c_f, c_g)$ — квадрат Евклидова расстояния между двумя центрами кластеров. Тогда **расстояние Уарда между двумя кластерами** определяется как:

$$dw(S_i, S_j) = \frac{|S_i| \cdot |S_j|}{|S_i| + |S_j|} \cdot d(c_i, c_j).$$

Смысл: это разница между значениями критерия К-средних $W(S, c)$ при объединении двух кластеров. Например, при соединении двух малых кластеров в один большой расстояние до центра увеличится, т.к. два старых центра объединятся в один общий.

41. Расстояние ближайшего соседа между двумя кластерами.

Расстояние ближайшего соседа между двумя кластерами есть расстояние между двумя их крайними ближайшими друг к другу точками.

42. Минимальное (максимальное) покрывающее дерево и алгоритм Прима для его построения.

Дерево – граф без циклов. Имеем заданную матрицу связей. Её столбцы/строки представляются в виде вершин графа, а эти расстояния выписываются как веса рёбер. **Покрывающее дерево** – дерево, не пропускающее ни одного объекта. **Длина дерева** – сумма весов рёбер, которые в него входят. Соответственно, **минимальное (максимальное) покрывающее дерево** – задача минимизации (максимизации) суммы весов рёбер с помощью жадного алгоритма.

Алгоритм Прима: начинаем с пустого дерева. Запускаем туда любую вершину. После этого добавляем в него по одной вершине с самой короткой (длинной) связью. За N шагов строится дерево.

43. Дивизивный метод ближайшего соседа на основе минимального (максимального) покрывающего дерева.

По построенному покрывающему дереву строим дивизивное (сверху-вниз) дерево иерархии по методу ближайшего соседа (рис. 5): берётся самая слабая связь (ребро с наименьшим значением) и по ней разрезается дерево на два множества. Затем каждое из получившихся деревьев снова делится на два и так до тех пор, пока не останутся только листья.

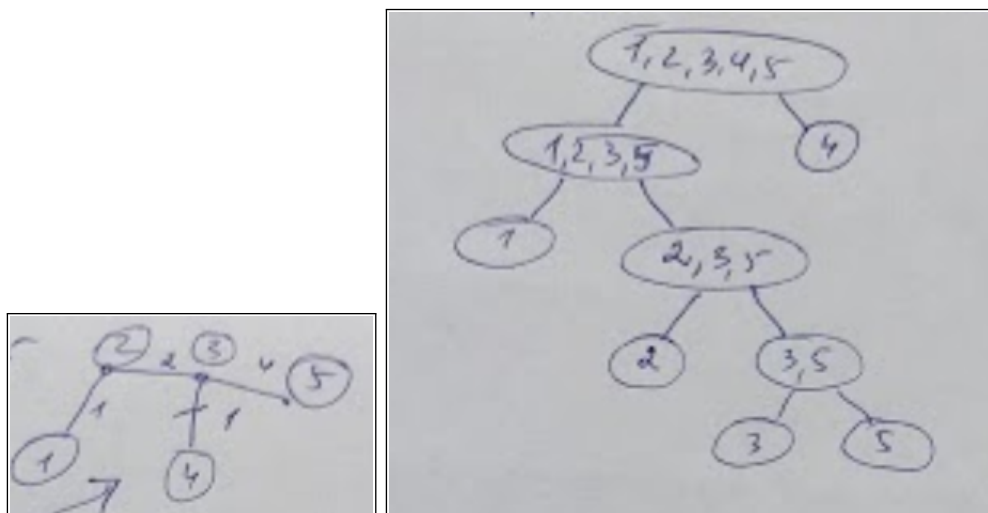


Рис. 5: дерево и как выстраиваем из него иерархию

44. Задача максимизации полусредней связи в разбиении, ее связь с методом К-средних.

Дана матрица связей $A = (a_{ij})$, где $i, j = 1, \dots, N$. Необходимо найти такое разбиение S на кластеры так, чтобы **полусредние связи** ($\frac{1}{|S_k|} \sum_{i,j \in S_k} a_{ij}$ полусреднее, так как делим не на $|S_k|^2$, хотя элементов a_{ij} берём как раз-таки квадрат) в кластере были максимальными. Тогда **задачей максимизации полусредней связи** является максимизация критерия полусредней связи (т.е. поиск такого разбиения, которое его максимизирует):

$$G(S) = \sum_{k=1}^K \frac{1}{|S_k|} \sum_{i,j \in S_k} a_{ij} \longrightarrow \max.$$

Связь с K-means: если взять $a_{ij} = \langle y_i, y_j \rangle$, то a_{ij} – дополнительный критерий K-means, согласно Пифагорову разложению, где было:

$$F(S, c) = \sum_{k=1}^K |S_k| \langle c_k, c_k \rangle \longrightarrow \max.$$

45. Задача максимизации суммарной связи в разбиении, ее версии: модулярность и постоянный сдвиг (порог).

Суммарная связь в разбиении:

$$F(S) = \sum_{k=1}^K \sum_{i,j \in S_k} a_{ij} \longrightarrow \max.$$

Эту задачу можно решить только за счёт преобразования чисел a_{ij} , а именно двумя способами:

1. **Постоянный сдвиг:** $a'_{ij} = a_{ij} - \pi$, где в качестве порога π можно взять, например, среднюю связь;
2. **Модулярность:** $k_{ij} = a_{ij} - \frac{a_{i+}a_{+j}}{a_{++}}$, где a_{i+} – сумма по строке, a_{+j} – сумма по столбцу, a_{++} – сумма по всем элементам;

46. Задача о минимизации нормализованного разреза/максимизации нормализованной внутренней связи (НР).

Достаточно рассказать для случая двух кластеров, т.е. рассмотрим ситуацию, когда хотим разбить множество на два кластера S_1, S_2 . Тогда **нормализованный разрез**:

$$NC = \frac{A(S_1, S_2)}{A(S_1, I)} + \frac{A(S_2, S_1)}{A(S_2, I)} \longrightarrow \min,$$

где I – множество всех объектов, A – матрица связей, $A(*, *)$ – сумма связей.

47. Сведение НР к задаче о минимизации отношения Райли (Рэлея) для Лапласового преобразования матрицы связи.

Вводится матрица Лапласа (англ. *лапласиан*):

$$L = l_{ij} = \delta_{ij} - \frac{a_{ij}}{\sqrt{a_{i+}a_{+j}}},$$

где $\delta_{ij} = \begin{cases} 1 & \text{на диагонали,} \\ 0 & \text{вне} \end{cases}$. Этот лапласиан обладает массой полезных свойств, и оказывается, что **задача о минимальном нормализованном разрезе эквивалентна задаче минимизации отношения Рэлея**:

$$\frac{s^T L s}{s^T s} \longrightarrow \min,$$

где S – специальные вектора, определяемые данным разбиением на два кластера S_1, S_2 следующим образом: $\begin{cases} S_1 = \{i : s_i > 0\}, \\ S_2 = \{j : s_j \leq 0\} \end{cases}$.

48. Метод спектрального кластерного анализа.

Задача о минимизации отношения Рэлея есть в точности задача об отыскании минимального собственного числа ($\lambda \neq 0$) и соответствующего собственного вектора.

Как же работает **спектральный кластерный анализ**:

1. $A \longrightarrow L$, т.е. преобразуем матрицу связей в матрицу Лапласа по формуле

$$L = l_{ij} = \delta_{ij} - \frac{a_{ij}}{\sqrt{a_{i+}a_{+j}}};$$

2. Найдём минимальное $\lambda \neq 0$ и соответствующий ему собственный вектор у матрицы L ;
3. Определим $\begin{cases} S_1 = \{i : s_i > 0\}, \\ S_2 = \{j : s_j \leq 0\} \end{cases}$. Вместо того, чтобы оптимизировать комбинаторные критерии, просто решаем задачу на собственные числа и достаём всё из собственных векторов.

ЧЕНЖЛОГИ

- v0.0 (11.12.2018) – исходное, надо дописать 22, 24, 28, написать 23, 26, 30–34, 36–38, 41–48;
- v0.5 (13.12.2018) – поправил 10, 15, 17, 20, 21, дописал 22, 24 (спасибо Даше), написал 23, 26, 30–34, 38 а всё что ≥ 41 узнаем на консультации, туда же 36–37;
- v1.0 (14.12.2018) – поправил (спасибо Даше) 24, 27, 29, 31(!), 32, переписал 28, 30, 40, убрал 36–38 (их не будет), добавил 41–48 (без опоры на учебник)