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.

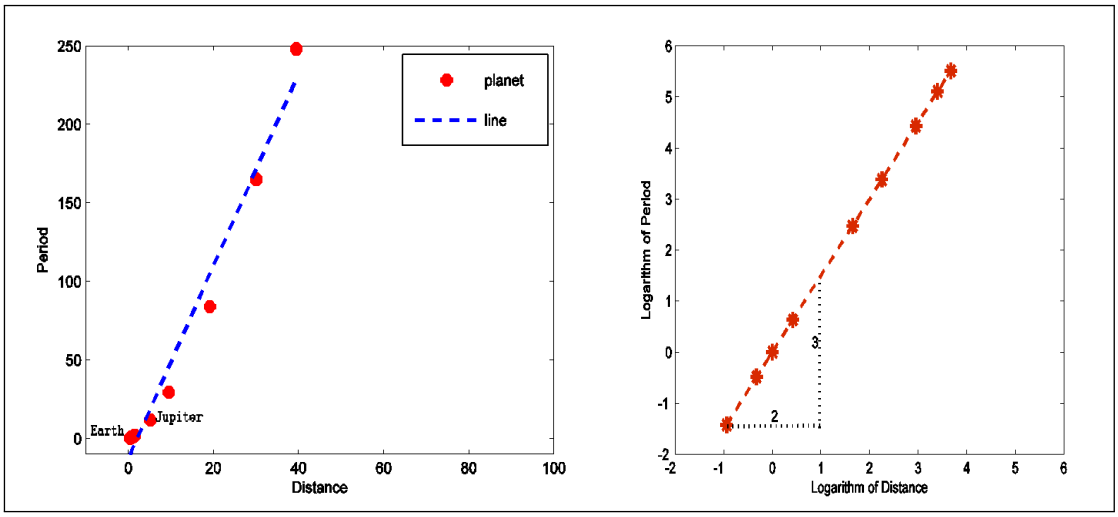

Рис. 1: зависимость в законе Кеплера

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 K-средних.

Метод K-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:

a) обновление кластеров (вокруг центров)
b) обновление центров (внутри кластеров)

Алгоритм K-средних:

0. **Инициализация**: пользователь выбирает число K кластеров и назначает K гипотетических центров (рис.2,*а*);
1. **Обновление кластеров**: при заданных K центрах $c_k(k = 1, 2, \ldots, K)$, 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 $S_k(k = 1, 2, \ldots, K)$ (рис.2,*в*). В качестве расстояния используется квадрат Евклидова расстояния;
2. **Обновление центров:** вычисляется арифметический центр (центр масс) каждого кластера S_k, который и назначается новым центром $c'_k(k = 1, 2, \ldots, K)$, (рис.2,*г*). Компоненты центра вычисляются как средние арифметические соответствующих компонент объектов из S_k;
3. **Правило остановки**: новые центры c'_k сравниваются со старыми. Если $c'_k = c_k$ для каждого $k = 1, 2, \ldots, K$, то вычисления останавливаются и выдаются результаты: центр c'_k и кластер S_k для каждого $k = 1, 2, \ldots, K$. 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**;

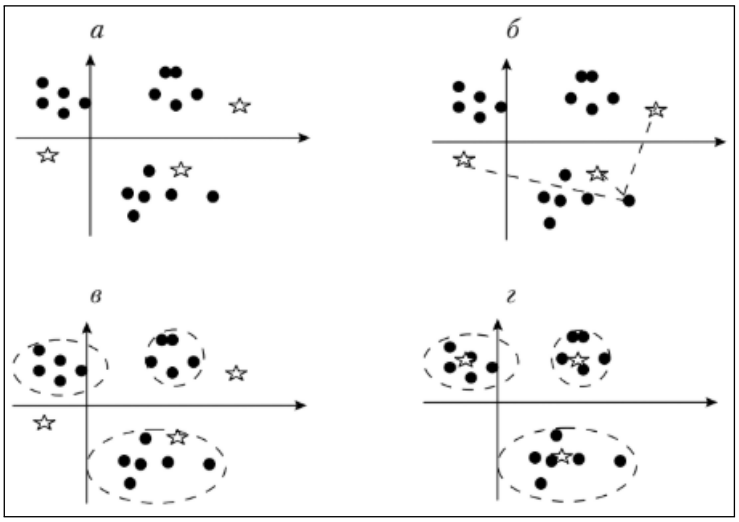

Рис. 2: итерации метода K-средних при K=3

9. Входные данные при использовании метода K-средних.

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

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 К непересекающихся кластеров $S = \{S_1, S_2, \ldots, S_K\}$, представленных в виде списков объектов $y_i \in S_k$ и центрами этих кластеров $(c_1, c_2, \ldots, c_k)$. 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,

$F(S, c) = \sum_{k=1}^{K} |S_k| < c_k, c_k >$, $c_k = (c_{k1}, \ldots, c_{kv})$ – сумма 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: $c' = \dfrac{\sum_{i \in S} y_i}{|S|}$;
3. **Правило остановки**: $c' \stackrel{?}{=} c$. 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: величиина $|S| < c, c >$ – вклад данного кластера в разброс данных, что вытекает из пифагорова разложения: $T(Y) = |S||c||^2 + W(S, c)$. Величина $\dfrac{|S| < c, c >}{T}$ – 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 совокупномти аномальных кластеров S_k выбрасываем те кластеры, которые удовлетворяют условию $|S_k| \leqslant t$. Обозначим количество оставшихся кластеров через K, а их центры через $c_1, c_2, \ldots, c_K$;
4. **Метод K-средних**: применяем метод K-средних, используя $c_1, c_2, \ldots, c_K$ в качестве начальных центров;

14. Интерпретация кластера.

Рассчитываем общий центр $c_k = (c_{k1}, c_{k2}, \ldots, c_{kv})$ и общее среднее на всём множестве $g_k = (g_1, g_2, \ldots, g_v)$ 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 и вычислим средние по этим признакам, получим два вектора $(m_1^1, \ldots, m_{5000}^1)$ и $(m_1^2, \ldots, m_{5000}^2)$. 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, накрывает данный интервал 0 или нет. Если накрывает, 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 (pivotal) 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-pivotal): 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.

Среднее Минковского – если задано множество чисел $x_1, \ldots, x_N$, 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 насколько s связано с t. Тогда **коэффициент Кетле** записывается как:

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

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:

$$Q = \sum p_{st} \cdot q_{st}$$

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. (дописать)

Коэффициент сопряженности Пирсона определяется следующей формулой:

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

23. Связь между коэффициентами Пирсона и Кетле. (написать)

текст

24. Метод главных компонент на основе сингулярного разложения прямоугольных матриц. (дописать)

Пусть $Y = (y_{iv})$ – матрица данных. Сингулярная тройка z, μ, c определяется следующими двумя условиями: $Y_c = \mu z$, $Y_z^T = \mu c$.

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

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

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

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

где $z_1^* = \sqrt{\mu_1} z_1$ и $z_2^* = \sqrt{\mu_2} z_2$, а также $c_1^* = \sqrt{\mu_1} c_1$ и $c_2^* = \sqrt{\mu_2} c_2$.

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.

текст

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

Матрица ковариации $C = \frac{Y^T \cdot Y}{N}$, где Y предварительно центрирована (т.е. из каждой строки вычтен вектор среднего).

Метод заключается в поиске собственных чисел этой матрицы λ_1, λ_2, соответствующих собственных векторов c_1, c_2 таких, что $Y \cdot c_1 = \lambda_1 \cdot c_1$, $Y \cdot c_2 = \lambda_2 \cdot c_2$ и определении главных компонент $z1 = \frac{Y \cdot c_1}{\sqrt{N \cdot \lambda_1}}$ и $z2 = \frac{Y \cdot c_2}{\sqrt{N \cdot \lambda_2}}$.

28. Связь сингулярных чисел матрицы данных с собственными числами ковариационной матрицы. (дописать)

Связь выглядит следующим образом:

$$\lambda_1 = \mu_1^2,\ \lambda_2 = \mu_2^2,$$

где μ_1, μ_2 рассчитываются для центрированной матрицы Y.

29. Отношение Райли (Рэлея) 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 Райли** называется выражение:

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

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

30. Матрица корреляции.

текст

31. Коэффициент корреляции в вероятностной перспективе.

текст

32. Коэффициент корреляции в аппроксимационной перспективе, его свойства.

текст

33. Коэффициент детерминации, его смысл.

текст

34. Задача линейной регрессии и ее решение.

текст

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. 3 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 $y = -2x + 3$).

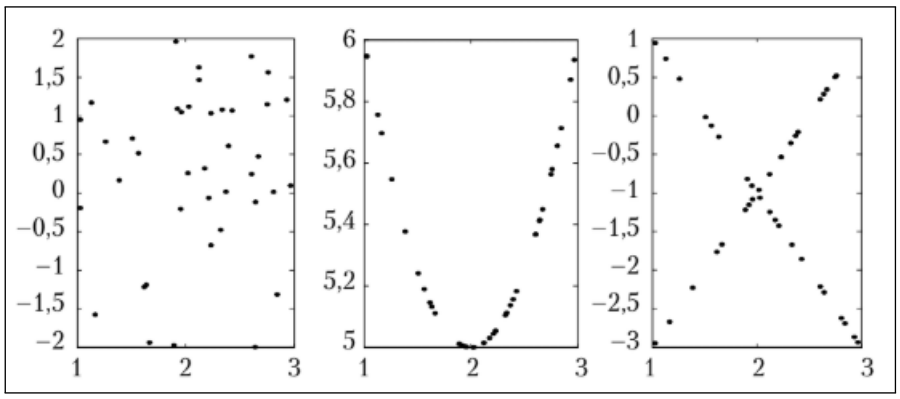

Рис. 3: поля рассеяния, соответствующие нулевому или почти нулевому значения корелляции

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

текст

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

текст

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

текст

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 $S = \{S_1, S_2, \ldots, S_K\}$. $d(c_f, c_g)$ – Евклидово расстояние между двумя центрами кластеров. Тогда **расстояние Уарда между двумя кластерами** определяется как:

$$dw(S_f, S_g) = \frac{|S_f| \cdot |S_g|}{|S_f| + |S_g|} \cdot d(c_f, c_g).$$

В агломеративных методах расстояние Уарда между кластерами, которые должны быть объединены, должно быть наименьшим – это будет удовлетворять объединению больших и маленьких кластеров. В дивизимных методах, в процессе деления, расстояние Уарда должно быть наибольшим, что позволит поделить большие кластеры на относительно соразмерные меньшие.

41. Расстояние ближайшего соседа между двумя кластерами.

текст

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

текст

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

текст

44. Задача максимизации полусредней связи в разбиении; ее связь с методом к-средних.

текст

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

текст

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

текст

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

текст

48. Метод спектрального кластерного анализа.

текст

Ченжлоги

v0.0 (11.12.2018) – исходное, надо дописать 22, 24, 28, написать 23, 26, 30–34, 36–38, 41–48;