МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Государственное бюджетное образовательное учреждение высшего профессионального образования

«ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ СИСТЕМ УПРАВЛЕНИЯ И РАДИОЭЛЕКТРОНИКИ» (ТУСУР)

УТВЕРЖДАЮ

Заведующий кафедрой ЭМИС

_____ И. Г. Боровской

«___» _____ 2012 г.

С.И. КОЛЕСНИКОВА

АНАЛИЗ ДАННЫХ

Методические указания по практическим работам

2012

Составитель: Колесникова С.И., каф.ЭМИС

А Н Н О Т А Ц И Я

Цели настоящих методических указаний: 1) освоение основных понятий и определений раздела знаний «Анализ данных»; 2) приобретение практических навыков в построении алгоритмов анализа данных, поиска закономерностей и распознавания характерных образов, анализа качества алгоритмов. В четырех частях указаний приведены примеры задач и методов их решения (анализа возможного решения) на следующие темы:

1. Методы и средства анализа данных.
2. Методы классификации и кластеризации.
3. Методы построения математических моделей и прогнозирования временных рядов.
4. Основные модели управления данными, многомерный анализ данных.

Теоретический материал приведен *только тот и в том объеме*, который необходим для решения предлагаемых задач. Задачи контрольных заданий являются весьма простыми, они предназначены для усвоения основных начальных понятий и основ современных методов анализа данных. Предполагается, что студенты знают математику в объеме, требуемом в техническом ВУЗе.

Методические указания предназначены для студентов экономического факультета.

Составитель: Колесникова С.И., каф.ЭМИС

СОДЕРЖАНИЕ ПРАКТИЧЕСКИХ ЗАНЯТИЙ
по дисциплине «Анализ данных»
и руководство по выполнению

Обозначения: ИДЗ - индивидуальные домашние задания
СРС - самостоятельная работа студентов
ИнЗ - интерактивное занятие

З-Эл – знания элементарные (определения, понятия, умение приводить иллюстрирующие примеры);

З-Пр – знания продуктивные (умение применить знания элементарные для решения учебных задач);

У-Эл – «умения» элементарные (уметь пользоваться готовыми частными алгоритмами для решения типовых задач), умение решать задачи по шаблону (копировать);

У-Пр – «умения» продуктивные (применять положения и известные частные алгоритмы дисциплины для решения практических задач);

В-Эл – элементарное владение методами дисциплины и уверенное осуществление (построение) основных операций для решения типовых задач;

В-Пр – продуктивно распознавать проблемы, алгоритмизировать их анализ и применять методы дисциплины для решения практических задач;

Составитель: Колесникова С.И., каф.ЭМИС

Краткое содержание тем и планируемых результатов их освоения

Тема практических занятий	Деятельность студента. Решая задачи, студент:	Отрабатываемые компетенции/ ожидаемый уровень освоения
1. Методы и средства анализа данных	• *использует* определения теории вероятностей и математической статистики; • *выбирает* способ построения модели анализируемого объекта; • *устанавливает* роль агента для анализа объекта; • *использует* знания, полученные ранее и самостоятельно решает задачи выбора; • *учится* применять статистические критерии; • совместно с преподавателем *разрабатывает* методику решения сложных (плохоформализуемых) задач. • *использует* понятия Data mining;	**ОК-1, ОК-3, ОК-12**/ З-Эл, У-Эл, В-Эл **ПК-26/** З-Пр, У-Пр, В-Пр
2. Методы классификации и кластеризации	• *учится* ставить задачи классификации; • *определяет* классификационные решающие правила; • *определяет* меры близости; • *изучает* алгоритмы кластеризации; • *применяет* адаптивные методы кластеризации.	**ОК-1, ОК-3, ОК-12**/ З-Эл, У-Эл, В-Эл **ПК-26/** З-Пр, У-Пр, В-Пр
3. Методы построения математических моделей и прогнозирования временных рядов	• *знакомится* с методами анализа временных рядов; • *изучает* свойства хаотических временных рядов; • *решает* задачи прогнозирования стохастических временных рядов на базе известных методов и алгоритмов; • *изучает* модели оценивания параметров модели.	**ОК-1, ОК-3, ОК-12**/ З-Эл, У-Эл, В-Эл **ПК-26/** З-Пр, У-Пр, В-Пр
4. Основные модели управления данными и распределённый анализ данных	• *анализирует* системы анализа распределённых данных и принципы их построения; • *знакомится* с существующими стандартиа Data mining; • *использует* методы инвариантных многообразий для управления сложными системами.	**ОК-1, ОК-3, ОК-12**/ З-Эл, У-Эл, В-Эл **ПК-26/** З-Пр, У-Пр, В-Пр

Составитель: Колесникова С.И., каф.ЭМИС

ХОД ПРАКТИЧЕСКИХ РАБОТ

1. Ознакомиться со справочными интернет-сведениями (СРС)
2. Ознакомиться с указанной темой в основной и дополнительной литературе рабочей программы.
3. Ознакомиться с принципом решения задач аудиторных.
4. Рекомендуется решить задачи домашние (в рамках СРС).
5. Ознакомиться с планом проведения интерактивных занятий в случае их проведения, прилагающегося к каждому разделу, и принципом подготовки к нему.
6. Составить и предоставить преподавателю отчет о работе, если он входит в форму отчетности по данному разделу знаний.

Раздел 1. Практические работы 1-5. Методы и средства анализа данных

ЦЕЛЬ РАБОТЫ

Знакомство с основными математическими дисциплинами, изучающими закономерности массовых случайных явлений и основными этапами исследования любой сложной (плохоформализуемой) проблемы:

1) определение проблемы;

разработка подхода (подходов в случае коллективного принятия решения) к решению проблемы;

2) разработка плана исследования;
3) сбор (определение) данных (обучающих и контрольных)
4) предобработка данных;
5) исследование и решение задачи;
6) подготовка отчета и презентации-доклада.

Примеры типовых аудиторных заданий

Практическое занятие 1 (4 ч.). Статистический анализ данных. Выборочный метод.

Задача 1.1. Путем опроса получены данные (табл. 1.1), n=80.

Таблица 1.1. Исходные данные для задания 1.1

1 4 1 4 3 3 3 1 0 6	1 2 3 5 1 4 3 3 5 1	5 2 4 3 2 2 3 3 1 3
2 3 1 1 4 3 1 4 3 1	6 4 3 4 2 3 2 3 3 1	4 6 1 4 5 3 4 2 4 5
2 6 4 1 3 3 4 1 3 1	0 1 4 6 4 7 4 1 3 5	

Выполнить следующие задания:

а) получить дискретный вариационный ряд и статистическое распределение выборки;
б) построить полигон частот;
в) составить ряд распределения относительных частот;
г) составить эмпирическую функцию распределения;
д) построить график эмпирической функции распределения;
е) найти основные числовые характеристики вариационного ряда (по возможности использовать упрощающие формулы для их нахождения):

1) выборочное среднее $\overline{x}_B$;
2) выборочную дисперсию $D(X)$;
3) выборочное среднее квадратическое отклонение $\sigma(X)$;
4) коэффициент вариации V;
5) интерпретировать полученные результаты.

Составитель: Колесникова С.И., каф.ЭМИС

Задача 1.2. В таблице 1.2 приведены размеры диаметров головок 100 заклепок (в мм), изготовленных станком (который делает их тысячами). Все контролируемые условия, в которых работал станок, оставались неизменными. В тоже время диаметры головок раз от разу несколько изменялись. Характерная черта случайных колебаний: изменения выглядят бессистемными, хаотичными.

Таблица 1.2. Исходные данные для задания 1.2

Диаметры 200 головок заклепок, мм							
17,29	17,42	17,54	17,54	17,40	17,55	17,40	17,25
17,42	17,50	17,22	17,21	17,28	17,52	17,45	17,52
17,28	17,44	17,52	17,52	17,27	17,22	17,24	17,17
17,52	17,52	17,29	17,57	17,51	17,24	17,29	17,47
17,51	17,48	17,52	17,58	17,57	17,22	17,51	17,40
17,20	17,48	17,40	17,57	17,51	17,40	17,52	17,55
17,40	17,24	17,22	17,27	17,48	17,48	17,52	17,25
17,40	17,25	17,45	17,48	17,29	17,58	17,44	17,55
17,28	17,59	17,47	17,45	17,52	17,54	17,20	17,28
17,42	17,25	17,55	17,51	17,47	17,40	17,29	17,20
17,45	17,44	17,42	17,29	17,41	17,29	17,50	17,48
17,52	17,24	17,45	17,42	17,29	17,28	17,45	17,50
17,55	17,22	17,22	17,59	17,45	17,22	17,22	17,48
17,29	17,25	17,44	17,50	17,42	17,51	17,42	17,28
17,24	17,28	17,58	17,21	17,21	17,45	17,42	17,44
17,24	17,49	17,50	17,28	17,48	17,42	17,27	17,29
17,54	17,22	17,25	17,45	17,22	17,44	17,28	17,27
17,55	17,25	17,40	17,52	17,59	17,48	17,45	17,40
17,42	17,25	17,50	17,28	17,42	17,24	17,41	17,24
17,42	17,55	17,27	17,41	17,28	17,14	17,42	17,52
17,28	17,54	17,20	17,18	17,22	17,45	17,29	17,25
17,24	17,27	17,50	17,51	17,42	17,22	17,25	17,40
17,57	17,21	17,40	17,25	17,28	17,58	17,58	17,28
17,25	17,27	17,28	17,29	17,22	17,20	17,42	17,24
17,22	17,22	17,21	17,45	17,29	17,45	17,41	17,45

Выполнить задания:

1. Для выборки диаметров головок заклепок вычислить *среднее значение, медиану, дисперсию, минимальный и максимальный элементы*.
2. Для выборки диаметров шляпок заклепок построить гистограмму частот с шагом группировки h (например, 0,075мм) на интервале от X_{min} (например, 13мм) до X_{max} (например, 13,75мм) (без учета сильно выделяющегося наблюдения)
3. Используя инструмент <*Описательная статистика*> создать таблицу основных статистических характеристик и разместить ее с соответствующим заголовком справа от исходных данных. Уметь объяснить смысл каждой статистики.
4. Обработать данные с целью выдвижения гипотезы о виде распределения наблюдаемой случайной величины и ее проверки.
5. Проверить выдвинутую гипотезу. Сделать выводы.

Решение. *Алгоритм* выполнения задания по проверке статистической гипотезы о нормальном распределении:

1. Определить размах выборки: R=Max-Min.
2. Назначить число карманов, m=8 (любое число от 7 до 25).

Составитель: Колесникова С.И., каф.ЭМИС

3. Должны быть найдены среднее (M) и станд.отклонение (σ).
4. Найти левые и правые границы для карманов, пронумерованных от 0 до m. При этом для кармана № 0 правая граница равна минимуму, для кармана № 1 правая граница равна минимуму+длина кармана,…
5. Построить гистограмму и выдвинуть гипотезу о виде распределения.
6. Найти значения предполагаемой ФР на границах карманов:

Для норм.распределения существует встроенная функция НОРМРАСПР(), где в качестве последнего аргумента печатаем ИСТИНА.

7. Найти теоретические вероятности попадания в карман (разность ФР по границам карманов).
8. Найти теоретические частоты (произведение теоретических вероятностей попадания в карман на обьем выборки).
9. Вычислить величины: (выборочная частота- теоретическая частота)^2/ теоретическая частота. Сумма этих величин – значение выборочного $\chi^2_{выб}$ критерия.
10. Найти значение теоретического критерия согласия $\chi^2_{теор}$ при заданном уровне значимости (у нас 0,05) можно по формуле ХИ2ОБР(вероятность; число степеней свободы), где число степеней свободы k=m-1-r, r=2 для норм.распр.
11. Сравнить $\chi^2_{выб}$ с $\chi^2_{теор}$, сделать выводы.

Практическое занятие 2 (4 ч.). Трендовый анализ.

Задача 1.3. Несколько человек решили организовать видеокафе на 6 столиков по 4 места за каждым. С каждого посетителя будет взиматься плата за сеанс видеофильма и ужин (всем посетителям будет предлагаться один и тот же набор блюд). Администрация города постановила, что плата за вход не должна превышать 5$. Требуется определить такую входную плату, при которой будет получена наибольшая выручка.

Решение. Предлагается следующая математическая модель данной задачи.

Обозначим входную плату через **X**. Тогда среднее число посетителей видеосалона является функцией от **X**. Обозначим эту функцию через **P(X).** В задаче требуется найти такое значение **X**, при котором выручка **X*P(X)** достигает максимума.

После обобщения опыта работы подобных кафе пришли к следующему виду функции

P(X): $P(X)=ax^2-bx+c$.

Коэффициенты для каждого кафе свои. Коэффициент **c** находится из соображений: **P(0)=c**, т.е. если в видеокафе пускают бесплатно, то свободных мест не будет, и **c** равно числу мест в кафе.

Таблица 1.3. Исходные данные для задания 1.3

Таблица эксперимента

a	b	Входная плата	Кол-во посетителей	Выручка	Кол-во посетителей	Выручка	Отклонение
		X	Эксперим. P(X)	Эксперимент	Теоретич. P(X)	Торетич.	
		1,5	17,5				
		2	16				
		2,5	14				
		3	12,5				
		3,5	11				
		4	9,2				
		5	7				
					Погрешность:		

2. Подберите a, b. Вычислите теоретическое количество посетителей и теоретическую выручку.
3. Вычислите отклонение между экспериментальной и теоретической выручкой и погрешность.
4. Подберите a, b, минимизировав погрешность.
5. Постройте графики экспериментальной и теоретической зависимости количества посетителей от входной платы и оформите их по своему усмотрению.

Составитель: Колесникова С.И., каф.ЭМИС

6. Определите, при какой входной плате выручка будет максимальна. (Каково среднее число посетителей сеанса при найденной оптимальной входной плате).

Практическое занятие 3 (4 ч.). Регрессионный анализ.

Задача 1.4. Дана выборка биржевых ставок относительно времени совершения сделки и цены сделки в рублях за один день работы биржи. Подобрать функциональную зависимость для набора наблюдений с помощью метода наименьших квадратов. Предсказать цены следующих сделок.

Таблица 1.4. Исходные данные для задания 1.4

Время	Цена сделки в рублях
11:16:45	99,45
11:21:53	99,4
11:23:09	99,31
11:23:37	99,31
11:24:49	99
11:24:57	99
11:48:40	98,61
11:49:45	98,99
11:53:51	98,66
11:55:05	98,65
11:55:24	98,7
11:58:18	98,8
11:58:18	98,8
11:58:24	98,65
11:58:35	98,8

Решение. Возможный вариант решения (цена сделки в рублях).

Наблюдение	Предсказанная цена сделки в рублях	Остатки
1	72,22015	27,22985
2	72,76796	26,63204
3	72,90313	26,40687
4	72,95293	26,35707
5	73,08099	25,91901
6	73,09522	25,90478
7	75,62617	22,98383
8	75,74178	23,24822
9	76,17932	22,48068
10	76,31094	22,33906
11	76,34473	22,35527
12	76,65421	22,14579
13	76,65421	22,14579
14	76,66488	21,98512
15	76,68444	22,11556

Интерактивные занятие №4-5 (№И1) по теме: «Анализ реальной проблемы и выбор методов и средства анализа данных. Основы непараметрической статистики»

Цель занятия: активное воспроизведение ранее полученных знаний по разделу 1 «Методы и средства анализа данных» в «незнакомых» условиях (применение основных понятий темы раздела 1 для решения задачи: построение статистических моделей для практически важных текстовых задач).

Форма текущего контроля освоения компетенций ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр (см. табл.1.5): *отчет* по решению следующих практических текстовых задач:

Задача И1.1. *Исследование математической модели метода главных компонент* - ядра факторного анализа и метода «Гусеница». Пусть дан исходный набор векторов линейного пространства R^n. Это входные данные: X - матрица данных (число столбцов равно числу признаков, число строк равно числу образцов).

Задача И1.2. Основы непараметрической статистики. Цели применения непараметрических методов. Работа с малыми выборками. Непараметрические критерии: критерии и-тест Манна-Уитни, W-тест Уилкоксона и др. Условия применимости. Апробировать критерии на указанной выборке.

Дополнительная литература.

Сошникова Л.А., Тамашевич В.Н. и др. Многомерный статистический анализ в экономике: Учеб. Пособие для вузов/Под ред. проф. Тамашевича. – М.: ЮНИТИ-ДАНА, 1999. –598с.
http://ru.wikipedia.org/wiki/%D1%F2%E0%F2%E8%F1%F2%E8%F7%E5%F1%EA%E8%E9_%EA%F0%E8%F2%E5%F0%E8%E9

Ряды данных: результаты пулковских и международных наблюдений международной службы вращения Земли (http: //hpiers.obspm.fr/), в дальнейшем С01 и С02;

Подготовка *занятия №1*. Выбор ведущего студента, ответственного за выбор и подачу необходимой информации и обсуждение с ним алгоритма занятия.

Таблица 1.5

№	№ задачи	Вид (совмещение нескольких видов) интерактивной работы	Трудо-емкость (час)	Отрабатываемые компетенции/ ожидаемый уровень освоения	Оценка личностных качеств	Контроль выполнения работы (участие в полемике, индивидуальные групповые задания (ИГЗ) и т.д)
1.	И1.1	Работа в команде. Решение ситуационных задач.	4	ОК-1, ОК-3, ОК-12/ З-Эл, У-Эл, В-Эл ПК-26/ З-Пр, У-Пр, В-Пр	Качество работы; своевременность сдачи отчета по решению ИГЗ	ИГЗ. Критерии оценивания поведения на занятии: активность, инициативность, грамотность, обоснованность защищаемой позиции.
Всего			4			

Составитель: Колесникова С.И., каф.ЭМИС

Вступление. **С**ообщение темы и обоснование ее актуальности через вышеуказанные задачи.

Основная часть:

I. **С**ообщение в виде доклада-презентации ответственными двумя студентами за проведение занятия 1, в котором излагается суть обсуждаемых положений:
   1) Схема исследования сложных динамических объектов с вероятностно-статистической точки зрения;
   2) Методы факторного анализа.
   3) Метод «Гусеница».

II. Выяснение позиций участников с зафиксированными точками зрения на решение вышеизложенных задач.

Итог II-го этапа: формирование целевых групп по общности позиций каждой из групп.

III. Организация коммуникации между группами: 1) выяснение позиции-варианта решения выявленных групп и защита занятой позиции; 2) формирование нового набора вариантов решений на основании общего обсуждения; 3) выбор одного решения голосованием;

IV. Повторная защита позиций-вариантов групп после проведения расчетов с целью оценки отклонения от «истинного» решения (попарное оценивание).

Выводы: реализован самостоятельный поиск учащимися путей и вариантов решения поставленной учебной задачи (выбор одного из предложенных вариантов или нахождение собственного варианта и обоснование решения на базе коллективной интерактивной работы).

Итог занятия №И1: Оценивание компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) по результатам работы на занятиях (активность, инициативность, грамотность, обоснованность защищаемой позиции) и своевременности сдачи отчета по решению практических задач И1.1-И1.2.

ВАРИАНТЫ ДОМАШНИХ ЗАДАНИЙ К РАЗДЕЛУ 1

Задача Д1.1. Дан ранжированный ряд: 23 23 24 24 25 25 25 27 28. Найти: частоты, относительные частоты, накопленные частоты.

Задача Д1.2. Распределение относительных частот появления признака задано таблицей 1.6.

Таблица 1.6

x_i	0	1	2	1	4	5	6	7
n_i	0.05	0.161	0.175	0.1	0.2	0.05	0.018	0.025

Построить эмпирическую функцию распределения, используя накопленные частоты; найти моду, медиану и выборочные среднее и дисперсию.

Задача Д1.3. Пусть X – непрерывная случайная величина подчинена показательному (экспоненциальному) закону, плотность распределения которого зависит от одного неизвестного параметра λ: $f(x, \lambda)=\lambda\exp(-\lambda x)$, $x\geq 0$.
Используя полученные экспериментальные данные $x1$, $x2$, ... , xn, получить оценку параметра λ. Исследовать ее свойства.

Задача Д1.4. Чему равен коэффициент корреляции величин ax+b и ch+d, где a, b, c, d - детерминированные константы, а x и h имеют коэффициент корреляции r?

Составитель: Колесникова С.И., каф.ЭМИС

ВАРИАНТЫ КОНТРОЛЬНЫХ ЗАДАНИЙ К РАЗДЕЛУ 1

I вариант. Дана совокупность значений наблюдае6мой характеристики:

Таблица 1.7

№ единицы совокупности	1	2	3	4	5	6	7	8	9	10
Значение признака	19,2	17,8	20,1	16,9	20,4	18,7	18,3	19,6	19,8	17,5

Определить *показатели вариации*: размах вариации; отклонение признаков x_i от типического уровня, свободного от случайных колебаний; общий объем вариации; средний размер отклонений в расчете на единицу совокупности; среднее линейное отклонение; дисперсию; среднее квадратическое отклонение; коэффициент вариации.

II вариант

Обобщите первичные результаты, полученные в результате наблюдений и экспериментов (описательная статистика). Необходимо определить также наличие линейной связи между признаками x и y. Осуществить регрессионный анализ зависимости,

Таблица 1.8

x	3	2	4	5	6	7	8	9	10	11
y	9	7	12	15	17	19	21	23	25	27

III вариант

Обобщите первичные результаты, полученные в результате наблюдений и экспериментов (описательная статистика). Необходимо определить наличие и характер связи между признаками x и y. Осуществить дисперсионный анализ зависимости,

Таблица 1.9

x	1	2	4	5	6	7	8	9	10	11
y	8	7	11	13	15	17	19	19	11	11

Контрольные вопросы к разделу 1

1. Типы данных.
2. Перечислите основные задачи статистического анализа.
3. Дать краткую характеристику.основным статистическим методам:
   корреляционный анализ, дисперсионный анализ, факторный анализ, кластерный анализ, дискриминантный анализ, регрессионный анализ, многомерное шкалирование.

Раздел 2. Практические работы 6-10
Методы классификации и кластеризации

Цель работы: Изучение и закрепление материала: по второму разделу: Методы классификации и кластеризации, в частности, знакомство с основными из них:

- классификация с помощью деревьев решений;
- байесовская классификация;
- классификация при помощи искусственных нейронных сетей;
- классификация методом опорных векторов;
- статистические методы, в частности, линейная регрессия;
- классификация при помощи метода ближайшего соседа;
- классификация CBR-методом (Case Based Reasoning, метод рассуждения по аналогии);

Составитель: Колесникова С.И., каф.ЭМИС

• классификация при помощи генетических алгоритмов.

Форма текущего контроля освоения компетенций ОК-12, ОК-13, ПК-14, ПК-15 уровни З-Пр, У-Пр, В-Пр (см. табл.2.1): отчет по решению следующих практических текстовых задач:

Примеры типовых аудиторных заданий

Практическое занятие 6 (2 ч). Задачи классификации и классификационные решающие правила.

Задача 2.1. Привести краткую характеристику подходов к кластеризации.
Решение. Условно существуют следующие классы подходов к кластеризации.
• Алгоритмы, основанные на разделении данных.
- разделение объектов на k кластеров;
- итеративное перераспределение объектов для улучшения кластеризации.
• Иерархические алгоритмы:
- агломерация: каждый объект первоначально является кластером, кластеры, соединяясь друг с другом, формируют больший кластер и т.д.
• Методы, основанные на концентрации объектов:
- основаны на возможности соединения объектов;
- игнорируют шумы, нахождение кластеров произвольной формы.
• Грид-методы:
- квантование объектов в грид-структуры.
• Модельные методы:
- использование модели для нахождения кластеров, наиболее соответствующих данным.

Задача 2.2. Проанализировать пример дерева решений, задача которого - ответить на вопрос: «Играть ли в гольф?» и ответить на вопрос: в чем суть принятия решения на базе метода «Дерево решений» (дихотомическая классификационная модель бинарных деревьев).

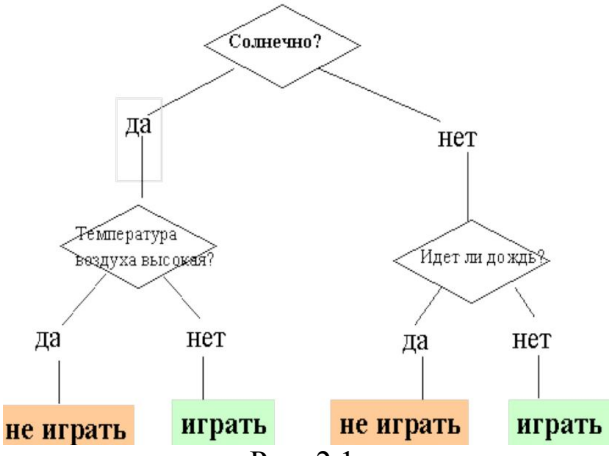

Рис. 2.1.

Задача 2.3. Представить набор данных в декартовых и параллельных координатах. Сделать выводы об условиях применения указанных систем координат.
Справка. Переменные кодируются по горизонтали, вертикальная линия определяет значение переменной. Метод представления многомерных данных был изобретен Альфредом Инселбергом (Alfred Inselberg) в 1985 году.
Решение.

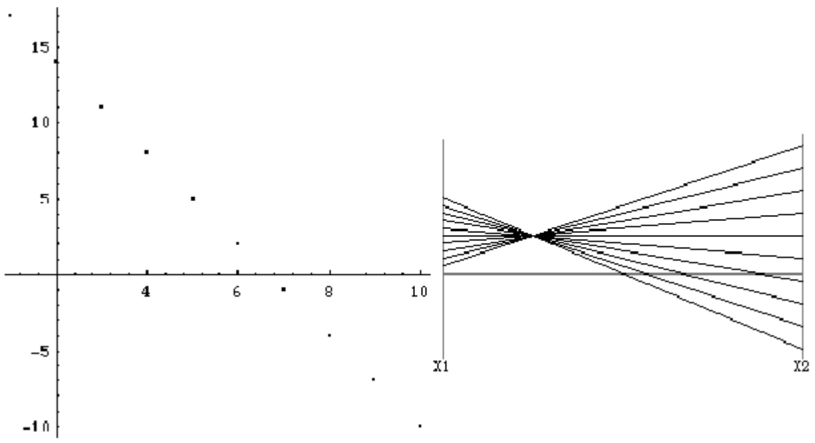

Рис. 2.2.

Практическое занятие 7 (2 ч.). Типы данных. Виды метрик.

Задача 2.4. Перечислите меры расстояния и укажите условия применимости.
Евклидово расстояние. Это вероятно наиболее часто используемый тип расстояния. Оно является простым геометрическим расстоянием в многомерном пространстве и вычисляется как: $d_2(x, y) = \left(\sum_i (x_i - y_i)^2 \right)^{1/2}$.

Используется и *квадрат евклидова расстояния,* если мы хотим придать прогрессивно возрастающий вес объектам, которые являются более удаленными. Это расстояние вычисляется как: $d'_2(x, y) = \sum_i (x_i - y_i)^2$.

Метрика Хемминга (покоординатное расстояние, городских кварталов, манхэттенское расстояние). Это расстояние в некотором смысле усредняет разницу между различными компонентами векторов. В большинстве случаев, эта мера расстояния дает результаты, подобные простому евклидову расстоянию. Однако, отметим, что при данной мере, эффект привносимый отдельными большими компонентами демпфируется (так как они не возводятся в квадрат). Покоординатное расстояние вычисляется так: $d_1(x, y) = \sum_i |x_i - y_i|$.

Расстояние Чебышева. Эта мера расстояния может подойти в случае, когда нам потребуется определить два объекта как различные, если они различны хотя бы по одному измерению: $d_\infty(x,y) = \max_i |x_i - y_i|$.

Степенное расстояние. Иногда может потребоваться увеличить или уменьшить вес увеличения расстояний по измерениям. Это может быть достигнуто путем использования степенного расстояния. Расстояние это вычисляется как: $d_{pr}(x, y) = \left(\sum_i (x_i - y_i)^p \right)^{1/r}$ где r и p - определяемые пользователем параметры. Поведение данной меры выглядит следующим образом: Параметр p контролирует вес разностей по отдельным компонентам, параметр r контролирует вес придаваемый расстоянию между объектами в целом. Если r и p равны 2, то это расстояние равно Евклидову расстоянию, при r =p – мера Минковского.
Мера «доли рассогласования». Мера целесообразна для номинальных признаков. Это расстояние вычисляется как: $d(x,y) = \frac{quantiy\ x \neq y}{i}$.

Обобщённая мера расстояний Минковского $d_p(x, y) = \left(\sum_i (x_i - y_i)^p \right)^{1/p}$ (при p=1 – метрика Хемминга, при p=2 – метрика Евклида, при p=∞ – метрика Чебышева).

Существуют и другие виды расстояний (Махаланобиса) и функции близости-различия объектов (FRiS-функция (Загоруйко Н.Г.)), не являющиеся расстоянием в общепринятом смысле (симметричность $d(x,y)= d(y,x)$, $d(x,y)=0$ при $x=y$, неотрицательность $d(x,y)\geq 0$). Ознакомиться с другими подходами по оценке расстояний самостоятельно.

Задача 2.5. Типы данных и признаков: перечислите и приведите примеры.
Решение.
Описательный признак – признак, который может быть выражен только словесно.
Количественный признак – признак, который может быть выражен численно.
Прямой признак – свойство непосредственно присуще характерному объекту.
Косвенный признак – свойства не самого характеризуемого объекта, а объекта связанного с ним либо входящих в него.
Первичный признак – абсолютная величина, может быть измерен.
Вторичный признак – результат сопоставления первичных признаков, он измеряется непосредственно.
Натуральный признак – измеряется в штуках, кг, тоннах, литрах и т.д.
Трудовой признак – измеряется в человеко-днях, человеко-часах.
Стоимостной признак - измеряется в рублях, $, €, £.
Безразмерный признак – измерение в долях, %
Альтернативный признак – признак, который принимает только одно значение из нескольких возможных.
Дискретный признак – принимает только целое значение, без промежуточного.
Непрерывный признак – признак, принимающий любые значения в определенном диапазоне.
Факторный признак – признак, под действием которого изменяется другой признак.
Результативный признак – признак, который изменяется под признаком другого
Моментный признак – признак, измеренный на определенный момент времени.
Интервальный признак – признак за определенный интервал времени.

Практическое занятие 8 (2 ч.). Использование нейро-алгоритмов для классификации и кластеризации.

Задача 2.8. [1] Рассмотреть пример конструирования нейронной сети в пакете Matlab для следующей задачи. Пусть имеется 15 независимых переменных - показателей деятельности фирмы и одна зависимая переменная - объем продаж. Имеем базу данных за прошедший год. Необходимо построить понедельный прогноз объемов продаж на месяц.

Решение. Предлагается использовать трехслойную сеть обратного распространения, включающую 15 нейронов во входом слое (по количеству входных переменных), 8 нейронов во втором слое и 1 нейрон в выходном слое (по количеству выходных переменных).
Для каждого слоя выберем передаточную функцию: первый слой - logsig, второй - logsig, третий - purelin. В среде Matlab синтаксис такой нейронной сети выглядит следующим образом:
Net=netff(PR, [S1,S2, : , Sn],{TF1,TF2, : , TFn},btf, blf, pf), где PR - массив минимальных и максимальных значений для R векторов входа; Si - количество нейронов в i-м слое; TFi - функция активации слоя i; btf - обучающая функция, реализующая метод обратного распространения; blf - функция настройки, реализующая метод обратного распространения; pf - критерий качества обучения.

[1] http://www.arshinov74.ru/files/files/10.pdf

Активационной функцией может выступать любая дифференцируемая функция, например, tansig, logsig, purelin.
Net=netff(minmax (P), [n,m, l],{ logsig, logsig, purelin },trainpr), где P - множество входных векторов; 132n - количество входов НС; m - количество нейронов в скрытом слое; l - количество выходов НС.
Необходимо также установить метод расчета значения ошибки. Например, если выбран метод наименьших квадратов, то эта функция будет выглядеть так: Net.performFcn='SSE'.
Для установления максимального количества эпох равным 10000 воспользуемся функцией: net.trainParam.epochs=10000.
Запустить процесс обучения можно таким образом: [net,tr]=train(net,P,T);
После окончания обучения сети ее можно сохранить в файле, например, с именем nn1.mat. Для этого необходимо выполнить команду: save nn1 net;
Для более детального изучения конструирования нейронных сетей в Neural Network Toolbox можно порекомендовать [49, 50].

Практические занятия 9-10
Интерактивное занятие №2.1-2.2 (№И2) по теме: **«Классификация на базе конструирования нейросети в специализированном нейропакете»**

Цель занятия: активное воспроизведение ранее полученных знаний по разделу 2 и освоение темы «Классификация на базе конструирования нейросети в специализированном нейропакете».

Форма текущего контроля освоения компетенций ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-П, *отчет* по решению указанных практических задач:

Задача И2.1. Рассмотреть решение задачи «Выдавать ли кредит клиенту» в аналитическом специализированном нейропакете (например, Deductor, BaseGroup). В качестве обучающего набора данных выступает база данных, содержащая информацию о клиентах, в частности: Сумма кредита, Срок кредита, Цель кредитования, Возраст, Пол, Образование, Частная собственность, Квартира, Площадь квартиры. На основе этих данных необходимо построить модель, которая сможет дать ответ, входит ли клиент, желающий получить кредит, в группу риска невозврата кредита, т.е. пользователь должен получить ответ на вопрос «Выдавать ли кредит клиенту?». Задача относится к группе задач классификации, т.е. обучения с учителем.

Подготовка *занятия №И2.* Выбор ведущего студента, ответственного за выбор и подачу необходимой информации, и обсуждение с ним алгоритма занятия.

Таблица 2.1

№	№ задачи	Вид интерактивной работы (совмещение нескольких видов)	Трудоемкость (час.)	Отрабатываемые компетенции/ ожидаемый уровень освоения	Оценка личностных качеств	Контроль выполнения работы (участие в полемике, индивидуальные групповые задания (ИГЗ) на базе выбранного программного продукта и т.д)
1.	И2.1	Работа в команде. Решение	4	ОК-1, ОК-3, ОК-12/ З-Эл, У-Эл,	Качество работы; своевременнос	ИГЗ. Критерии оценивания поведения на занятии: активность,

		ситуационных задач.		В-Эл ПК-26/ З-Пр, У-Пр, В-Пр	ть сдачи отчета по решению ИГЗ	инициативность, грамотность, обоснованность защищаемой позиции.
Всего			4			

Вступление. **С**ообщение темы и обоснование актуальности вероятностных распределений: нормального, показательного, равномерного в практических задачах (в экономике, в частности).

Основная часть:

I. **С**ообщение в виде доклада-презентации ответственным (студентом, двумя студентами) за проведение занятия И2, в котором излагается суть обсуждаемых положений:

1) Персептрон Розенблатта.
2) Классификационные правила и их реализация в нейросети;
3) Возможности нейро-алгоритмов.
4) Примеры успешного решения нелинейных проблем на базе нейро-алгоритмов.

II. Выяснение позиций участников с зафиксированными точками зрения на решение основной задачи И2.1, решаемой на занятии.

Итог II-го этапа: формирование целевых групп по общности позиций каждой из групп.

III. Организация коммуникации между группами: 1) выяснение позиции-варианта решения выявленных групп и защита занятой позиции; 2) формирование нового набора вариантов решений на основании общего обсуждения; 3) выбор одного решения голосованием;

IV. Повторная защита позиций-вариантов групп после проведения расчетов с целью оценки отклонения от «истинного» решения.

Выводы: реализован самостоятельный поиск учащимися путей и вариантов решения поставленной учебной задачи (выбор одного из предложенных вариантов или нахождение собственного варианта и обоснование решения на базе коллективной интерактивной работы).

Итог занятия №И2: Оценивание компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) по результатам работы на занятиях (активность, инициативность, грамотность, обоснованность защищаемой позиции) и своевременности сдачи отчета по решению реальной практической задачи.

ВАРИАНТЫ ДОМАШНИХ ЗАДАНИЙ К РАЗДЕЛУ 2

Задача Д2.1. Даны матрицы Q - описание 6-ти объектов; R - соответствие номеров объектов и классов.

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{bmatrix} 2 & 1 & 1 & 2 \\ 1 & 3 & 2 & 1 \\ 2 & 1 & 2 & 1 \\ 1 & 3 & 1 & 2 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 3 \end{bmatrix} \end{array}; \quad R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}.$$

Составитель: Колесникова С.И., каф.ЭМИС

Сформировать решающее правило, по которому можно отнести объект S=(2, 1, 2, 3) (не входящий в обучающую выборку) к одному из выделенных классов.

Задача Д2.2. Дано два набора точек:

$$(1,3),(3,3),(4,1),(4,3),(3,4) \in K_1,(8,6),(9,6),(6,9),(6,6),(9,10) \in K_2..$$

Найти разделяющую границу - решающую поверхность.

Задача Д2.3. Перечислите меры расстояния и укажите условия применимости.

Задача Д2.3. Привести пример дерева классификации, с помощью которого решается задача «Выдавать ли кредит клиенту?»

ВАРИАНТЫ КОНТРОЛЬНЫХ ЗАДАНИЙ К РАЗДЕЛУ 2

Вариант I

Имеется база данных о клиентах туристического агентства с информацией о возрасте и доходе за месяц. Есть рекламный материал двух видов: более дорогой и комфортный отдых и более дешевый, молодежный отдых. Соответственно, определены два класса клиентов: класс 1 и класс 2. База данных приведена в таблице 2.2.

Таблица 2.2.

Код клиента	Возраст	Доход	Класс
1	18	25	1
2	22	100	1
3	30	70	1
4	32	120	1
5	24	15	2
6	25	22	1
7	32	50	2
8	19	45	2
9	22	75	1
10	40	90	2

Определить, к какому классу принадлежит новый клиент, и какой из двух видов рекламных материалов ему стоит отсылать.

Вариант II

Имеется база данных о клиентах клиники с информацией о возрасте и диагнозе за анализируемый период. Есть данные-уровни о вредности производства: 1, 2, 3. База данных приведена в таблице 2.3.

Таблица 2.3.

Код клиента	Возраст	Диагноз	Уровень вредности
1	18	грипп	1
2	22	ОРЗ	1
3	30	ОРЗ	1
4	32	Бронхит	3
5	24	Бронхит	3
6	25	ОРЗ	1
7	52	Ревматизм	2
8	49	Ревматизм	2
9	22	Воспаление легких	3
10	40	Ревматизм	2

Определить решающее правило, и ответить на вопрос: «к какому классу принадлежит новый клиент с показателями: (42, Бронхит, 3).

Составитель: Колесникова С.И., каф.ЭМИС

Контрольные вопросы к разделу 2

1. Охарактеризуйте методы построения деревьев решений.
2. Метод наименьших квадратов и условия его оптимальности.
3. Классификационные правила: приведите примеры.
4. Дайте определение опорным векторам.
5. Что такое «Эмпирический риск»?

Раздел 3. Практические работы 11-14
Методы построения математических моделей и прогнозирования временных рядов. Анализ свойств одномерных хаотических моделей.

Цель работы. Знакомство с основами теории временных рядов и систем их прогнозирования. Моделирование рядов динамики.

Примеры типовых аудиторных заданий.

Практическое занятие 11 (2 ч.). Методы нелинейной фильтрации стохастических временных рядов.

Задача 3.1. Методы модовой декомпозиции EMD (Empirical Mode Decomposition) и преобразование Гильберта-Хуанга HHT (Hilbert-Huang Transform). Построить прогноз на основе метода
Решение. Исследовать и апробировать алгоритм из работы:
Давыдов В.А., Давыдов А.В. Очистка геофизических данных от шумов с использованием преобразования Гильберта-Хуанга.// Электронное научное издание "Актуальные инновационные исследования: наука и практика", 2010, № 1. http://www.actualresearch.ru.

Задача 3.2. Апробировать процедуру нелинейной фильтрации на временных рядах, полученных измерением реальных характеристик электромеханического двигателя.

Решение. Приведем алгоритм нелинейной фильтрации (Г.М.Кошкин).
Алгоритм: *Нелинейная фильтрация сигнала с аддитивным шумом*
Вход: наблюдения дискретного сигнала $Y(t)$ глубины *p*: $Y_{i+1}, Y_{i+2}, \ldots, Y_{i+p+1}$, $i = 1, 2, \ldots,; t_i = t(T - N - p + i), i = \overline{1, N + p}$ - моменты времени, в которые производились измерения
Выход: оценки дискретного сигнала $\hat{Y}(t)$
Тело алгоритма:
1: Преобразуем сигнал к виду: $Y(t) = Y(t-1) + \Delta Y(t)$, где
$\Delta Y(t) = (Y(t) - Y(t-1)) + \xi(t)$ $\{\xi(t)$ - произвольный ограниченный шум, $\Delta Y(t)$ - зашумленное изменение сигнала на интервале $[t-1, t]\}$;
2: Осуществляем выбор параметров размытости выборки $h^o_{[1]}, h^o_{[2]}, \ldots, h^o_{[N]}$ с помощью следующей рекуррентной процедуры на основе аппарата из книги [18] в смысле критерия полного скользящего контроля.
3: Оцениваем функцию $F(Y)$ (условное математическое ожидание выхода стохастического объекта относительно входов $Y(t)$, или функция регрессии) в момент

времени $t(T)$ по следующим данным: $\{t_1,\dots,t_{N+p}\}$, $\{\Delta Y^1,\dots,\Delta Y^{N+p}\}$, где $t_i = t(T-N-p+i)$, $i=\overline{1,N+p}$ - моменты времени, в которые производились измерения, $\Delta Y^i = \Delta Y\left(t(T-N-p+i)\right), i=\overline{1,N+p}$ - выборочные значения изменения сигнала, T – величина интервала наблюдения. Оценка $\hat{F} = \hat{F}\left(\Delta Y_N,\dots,\Delta Y_{N+p-1}, t(T)\right)$ функции $F(\cdot)$ в момент времени $t(T)$ имеет вид:

$$\hat{F} = \frac{\sum_{i=1}^{N} \frac{\Delta Y_{i+p}}{h^t_{[i]}\prod_{z=1}^{p} h^{\Delta}_{[i]z}} K\left(\frac{\Delta Y_N - \Delta Y_i}{h^{\Delta}_{[i]1}}, \frac{\Delta Y_{N+1} - \Delta Y_{i+1}}{h^{\Delta}_{[i]2}}, \dots, \frac{\Delta Y_{N+p-1} - \Delta Y_{i+p-1}}{h^{\Delta}_{[i]p}}, \frac{t(T) - t_{i+p}}{h^t_{[i]}}\right)}{\sum_{i=1}^{N} \frac{1}{h^t_{[i]}\prod_{z=1}^{p} h^{\Delta}_{[i]z}} K\left(\frac{\Delta Y_N - \Delta Y_i}{h^{\Delta}_{[i]1}}, \frac{\Delta Y_{N+1} - \Delta Y_{i+1}}{h^{\Delta}_{[i]2}}, \dots, \frac{\Delta Y_{N+p-1} - \Delta Y_{i+p-1}}{h^{\Delta}_{[i]p}}, \frac{t(T) - t_{i+p}}{h^t_{[i]}}\right)}.$$

$h_{[i]k} > 0$ - последовательность чисел (параметров), сходящаяся к нулю для каждого $k = \overline{1,l}$.

Дополнительная литература.

Васильев В.А. Непараметрическое оценивание функционалов от распределений стационарных последовательностей / В.А. Васильев, А.В. Добровидов, Г.М. Кошкин. – М.: Наука, 2004. – 508 с.

Практическое занятие 12 (2 ч.). Анализ свойств временных рядов, порождаемых хаотическими моделями. Роль хаотических моделей в описании реальных ситуаций.

Задача 3.4. Анализировать свойства одномерной хаотической модели Хатчинсона.

Решение. Моделируем поведение модели Хатчинсона: $\frac{dN(t)}{dt} = r\left(1 - \frac{N(t-h)}{K}\right)N(t)$, где N – число членов популяции (животтных), K – средний размер популяции, r – относительный коэффициент роста (Мальтуса), h – время запаздывания, обусловленное возрастной структурой популяции. На рис.3.1 приведена зависимость функции N от параметра r, при изменении параметра r в интервале [0.554, 0.555] функция $N\to\infty$ (поведение интерпретируется как катастрофа).

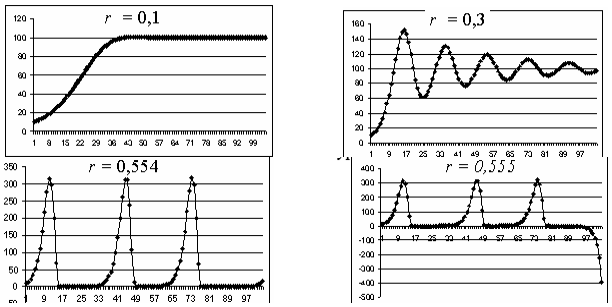

Рис. 3.1.

Задача 3.5. Анализировать свойства одномерной хаотической модели Фейгенбаума.

Решение. Математическая форма отображения:

$$x_{n+1} = \alpha x_n(1 - x_n),$$

где x_n принимает значения от 0 до 1 и отражает численность популяции в n-ом году, x_0 - начальная численность (в год номер 0); α - положительный параметр, характеризующий

скорость размножения (роста) популяции. Иногда данная формулировка называется отображением Ферхюльста (или Ферхюльста-Пирла), а логистическим отображением называется другая, но эквивалентная по свойствам формула: $x_{n+1} = 1 - \alpha x_n^2$. Это нелинейное отображение описывает два эффекта: размножение популяции, со скоростью, пропорциональной ее численности в момент, когда численность мала; конкуренцию (смертность при высокой плотности) за жизненные ресурсы, при которой скорость размножения падает из-за ограничения на «максимальную емкость» среды, в которой обитает популяция.

Задача 3.6..Применить синергетическое управление к модели Фейгенбаума с целью управления хаосом (устремления к стабильному положению - аттрактору).
Решение. Рассмотрим регулятор для уравнения Фейгенбаума:

$$\begin{cases} x_{k+1} = \alpha x_k(1 - x_k) + u_k, \\ u_k = \alpha x_k^2 - (\alpha + L)x_k + (1 + L)x_c, \\ u_0 = 0, \\ y_k - x_k + \gamma_k, \end{cases}$$

где x_k - сигнал, α – параметр, отвечающий за рост, u_k – управление,L-параметр, отвечающий за время достижения аттрактора, γ_k – шум.
Для численного исследования модели Фейгенбаума произведён комплекс вычислений для разных групп изменений α (единичный скачок, двойной скачок и линейный рост α) с присутствием и отсутствием зашумления основного сигнала.
Ниже для примера рассмотрен случай: α_1=2, α_2=2.5, x_c =0.5, x_0=0.8, L =0.626, σ=0.001, и показана возможность стабилизации системы.

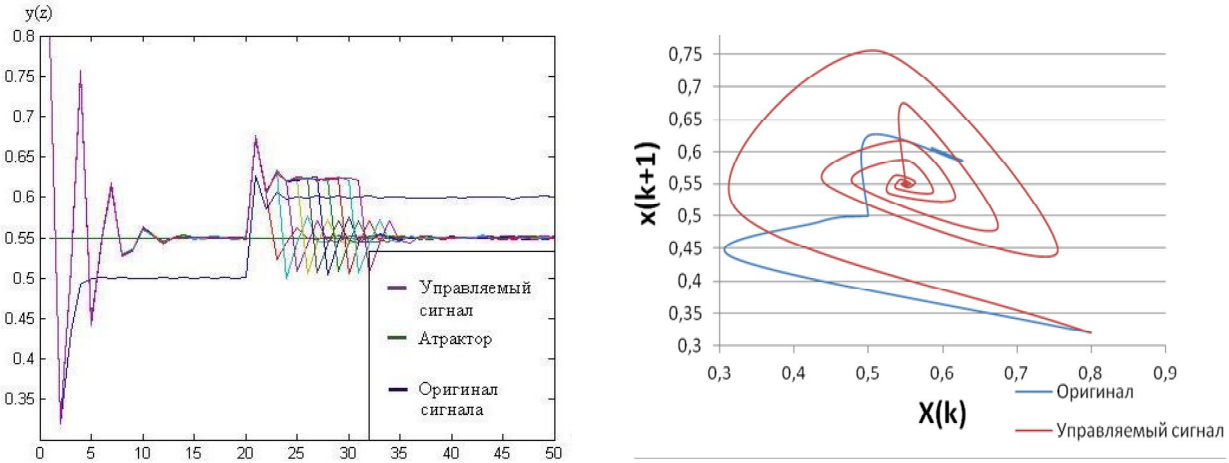

Рис. 3.2.

Интерактивное занятие №14 (№И3) по теме: «Методы анализа нестационарных стохастических временных рядов»

Цель занятия: знакомство с методами анализа и прогнозирования нестационарных временных рядов.

Форма текущего контроля освоения компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) (табл.3.1); *отчет* по решению практических задач (по выбору):

Задача И3.1. Типичной в сфере маркетинга является задача прогнозирования рынков. В результате решения данной задачи оцениваются перспективы развития конъюнктуры определенного рынка, изменения рыночных условий на будущие периоды, определяются тенденции рынка (структурные изменения, потребности покупателей, изменения цен).

Составитель: Колесникова С.И., каф.ЭМИС

Обычно в этой области решаются следующие практические задачи:
• прогноз продаж товаров (например, с целью определения нормы товарного запаса);
• прогнозирование продаж товаров, оказывающих влияние друг на друга;
• прогноза продаж в зависимости от внешних факторов.
Для заданной *базы данных* (student\Колесникова\ Анализ данных) торговых временных рядов провести исследование по всем вышеперечисленным задачам на базе методов «Гусеница» и модовой декомпозиции. Сравнить результаты.

***Подготовка** занятия №И3.* Выбор ведущего студента, ответственного за выбор и подачу необходимой информации, и обсуждение с ним алгоритма занятия.

Вступление. Сообщение темы (далее, на примере задачи 1) и обоснование ее актуальности через задачи анализа систем массового обслуживания.
Основная часть:

I. Сообщение в виде доклада-презентации ответственным (студентом) за проведение занятия И3, в котором излагается суть обсуждаемого явления:
   1) Методы главных компонент и «Гусеница»;
   2) Метод модовой декомпозиции;

II. Выяснение позиций участников с зафиксированными точками зрения на решение основной задачи И3.1.

Итог II-го этапа: формирование целевых групп по общности позиций каждой из групп.

III. Организация коммуникации между группами: 1) выяснение позиции-варианта решения выявленных групп и защита занятой позиции; 2) формирование нового набора вариантов решений на основании общего обсуждения; 3) выбор одного решения голосованием;

IV. Повторная защита позиций-вариантов групп после проведения расчетов с целью оценки отклонения от «истинного» решения.

Выводы: реализован самостоятельный поиск учащимися путей и вариантов решения поставленных учебных задач (выбор одного из предложенных вариантов или нахождение собственного варианта и обоснование решения на базе коллективной интерактивной работы).
Итог занятия №И3: Оценивание компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) по результатам работы на занятиях (активность, инициативность, грамотность, обоснованность защищаемой позиции) и своевременности сдачи отчета по решению практических задач согласно табл. 3.1.

Таблица3.1

№	№ задачи	Вид интерактивной работы (совмещение нескольких видов)	Трудоемкость (час.)	Отрабатываемые компетенции/ ожидаемый уровень освоения	Оценка личностных качеств	Контроль выполнения работы (участие в полемике, индивидуальные групповые задания (ИГЗ) на базе выбранного программного продукта и т.д)
1	И3.1	Работа в команде. Решение ситуационных задач.	4	ОК-1, ОК-3, ОК-12/ З-Эл, У-Эл, В-Эл ПК-26/ З-Пр, У-Пр,	Качество работы; своевременность сдачи отчета по решению	ИГЗ. Критерии оценивания поведения на занятии: активность, инициативность, грамотность,

				В-Пр	ИГЗ	обоснованность защищаемой позиции.
Всего			4			

ВАРИАНТЫ ДОМАШНИХ ЗАДАНИЙ К РАЗДЕЛУ 3

Задача Д3.1. Известен вид функции регрессии: $\widetilde{F} = \widetilde{f}\left(x; a_0, a_1, a_2\right) = \widetilde{a}_0 + \widetilde{a}_1 x^2 + \widetilde{a}_2 x^4$. Измерены значения некоторого параметра *Y* в шести точках:

x	0	0,00359	0,00718	0,01077	0,01436	0,01795
F	369	359	328	273	188	63

Необходимо найти коэффициенты a_i функции регрессии.

1. Найти левые и правые пределы доверительных интервалов для дисперсии с.в., заданной в таблице (преподавателем).
2. Построить гистограмму с.в., заданной в таблице (преподавателем).

Задача Д3.2. Имеется база данных «Прибыль компании» за период с 2007 по 2012 года, она представлена в табличном виде в таблице 3.2.

Таблица 3.2. Прибыль компании А

год	прибыль
2007	1100
2008	1101
2009	1104
2010	1105
2011	1106
2012	1107

Иллюстрировать динамику роста прибыли компании за указанный период. Как влияет масштаб и формат оси y (значения параметров) на вид динамику роста прибыли?

ВАРИАНТЫ КОНТРОЛЬНЫХ ЗАДАНИЙ К РАЗДЕЛУ 3

Указание. Работы выполняются в любом пакете: MatLab, Excel, MatCad, Statistica и др.

Вариант I

Методом скользящего среднего произвести сглаживание временного ряда, приведенного в таблице. Использовать интервал усреднения, в который входили бы 2 последовательных года наблюдения. Результаты представить в таблице. Сравнить с результатами линейной регрессии.

Вариант II

Построить прогноз для стохастического временного ряда двумя кривыми регрессии: квадратичной и логарифмической. Сравнить на базе двух критериев: детерминации и СКО. Чем можно объяснить разницу результатов?

Вариант III

Построить прогноз для стохастического временного ряда на базе генетического алгоритма.

Контрольные вопросы к разделу 3

Составитель: Колесникова С.И., каф.ЭМИС

1. Привести два принципиальных отличия временного ряда от простой последовательности наблюдений.
2. Назовите вопросы, на которые необходимо ответить перед началом прогнозирования.
3. Точность прогноза характеризуется ошибкой прогноза. Назовите наиболее распространенные виды ошибок.

Раздел 4. Практические работы 15-20
Основные модели управления данными, многомерный анализ данных

Цель работы – знакомство с основными моделями управления данными, с элементами многомерного анализа данных.

Форма текущего контроля освоения компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) (см. табл. 4.1): отчет по решению следующих практических текстовых задач:

Примеры типовых аудиторных задач

Практическое занятие 15 (2 ч.). Комплексный подход к внедрению Data Mining, OLAP и хранилищ данных в СППР.

Задача 4.1. Осуществить классификацию систем поддержки принятия решений (СППР), учитывая одну из существующих по типу ориентации: на данные, на модели. на знания, на документы, на коммуникации, интер- и интра-СППР, на специфику функций, на базу Web.

Задача 4.2. Процесс-исследование Data Mining: основные этапы. Охарактеризовать этапы.
Решение. Традиционный процесс Data Mining включает этапы:
• анализ предметной области;
• постановка задачи;
• подготовка данных;
• построение моделей;
• проверка и оценка моделей;
• выбор модели;
• применение модели;
• коррекция и обновление модели.

Задача 4.3. Исследовать чувствительность Data Mining к выбросам. Методы очистки данных.
Решение. Нанести указанные данные на диаграмму. Указать возможные выбросы. Сформировать таксоны-эталоны по одному из методов (например, FRiS-столпы (Н.Г.Загоруйко)) и удалить эталоны с малым весом.

Практическое занятие 16 (2 ч.). Разные подходы к определению знаний и данных, информации.

Задача 4.4. Задать понятия: знания, данные, информация своими свойствами. Привести разные определения указанным понятиям.

Составитель: Колесникова С.И., каф.ЭМИС

Решение. Свойства информации: запоминаемость, передаваемость, воспроизводимость, преобразуемость, стираемость, объективность и субъективность, достоверность, полнота, адекватность, доступность.
По Д.А. Поспелову выделяют 6 основных факторов, характеризующих знания: 1) внутренняя интерпретируемость, 2) структурированность, 3) связность, 4) шкалирование, 5) семантическая метрика, 6) наличие активности.
Данные – это факты, характеризующие объекты, процессы и явления, а также их свойства.

Задача 4.5. Модели представления данных и знаний (по Дюку В.Д., Янковской А.Е. и др.).

Практическое занятие 17 (2 ч.). OLAP: оперативная аналитическая обработка данных (On-Line Analytical Processing).

Задача 4.6. Продемонстрируйте технологию OLAP-анализа для примера (рис.4.1[2]).

Решение. Согласно технологии OLAP (по Кодду), одновременный анализ по нескольким измерениям определяется как многомерный анализ. Каждое измерение включает направления консолидации данных, состоящие из серии последовательных уровней обобщения, где каждый вышестоящий уровень соответствует большей степени агрегации данных по соответствующему измерению. Здесь измерение Исполнитель может определяться направлением консолидации, состоящим из уровней обобщения "предприятие - подразделение - отдел - служащий". Измерение Время может даже включать два направления консолидации - "год - квартал - месяц - день" и "неделя - день", поскольку счет времени по месяцам и по неделям несовместим. В этом случае становится возможным произвольный выбор желаемого уровня детализации информации по каждому из измерений. Операция спуска соответствует движению от высших ступеней консолидации к низшим; напротив, операция подъема означает движение от низших уровней к высшим.

Практическое занятие 18 (2 ч.). Системы анализа распределённых данных.

Задача 4.7. Проанализируйте способы[3] анализа распределённых данных. Решите задачу выбора недвижимости (офисов, складов, квартир). Измерения - обычные для этого рынка. Город, Район, Количество комнат, Расстояние до метро, Этаж, Тип дома, Дата и т.д. Фактов три - средняя цена, максимальная цена, минимальная цена. Манипулируя измерениями, покупатель может определиться со своими возможностями, а продавец проанализировать зависимости цен, динамику цен и назначить правильную цену.
Соответствующая база данных находится в папке: student\Колесникова\ Анализ данных.

[2] http://citforum.ru/seminars/cis99/sch_03.shtml
[3] http://www.rarus.nn.ru/products/olap/30.htm

Интерактивное занятие №19 (№И4) по теме: «Основные модели управления данными и распределённый анализ данных»

Цель занятия*:* активное воспроизведение ранее полученных знаний в «незнакомых» условиях (применение знакомой модели для решения незнакомых задач); ознакомиться с максимально широким кругом понятий раздела 4 «Основные модели управления данными, многомерный анализ данных». Раскрыть взаимосвязь понятий, их внутреннюю логику. Научиться правильно формулировать цель в каждой задаче.

Форма текущего контроля освоения компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) (табл.3); отчет по решению практических задач (по выбору):

Задача И4.1. Приведите примеры задача Data Mining в промышленном производстве:
• комплексный системный анализ производственных ситуаций;
• краткосрочный и долгосрочный прогноз развития производственных ситуаций;
• выработка вариантов оптимизационных решений;
• прогнозирование качества изделия в зависимости от некоторых параметров технологического процесса;
• обнаружение скрытых тенденций и закономерностей развития производственных процессов;
• прогнозирование закономерностей развития производственных процессов;
• обнаружение скрытых факторов влияния;
• обнаружение и идентификация ранее неизвестных взаимосвязей между производственными параметрами и факторами влияния;
• анализ среды взаимодействия производственных процессов и прогнозирование изменения ее характеристик;
• выработка оптимизационных рекомендаций по управлению производственными процессами;
• визуализация результатов анализа, подготовку предварительных отчетов и проектов допустимых решений с оценками достоверности и эффективности возможных реализаций.

Задача И4.2. Системы анализа распределённых данных. Разработка Web-сайта на основе базы данных.
Темы для предварительного изучения. Настройка Internet Information Server, знакомство с HTML, ASP, настройка DSN по эл. источнику: Сороковиков В. Н. Разработка Web-сайта на основе базы данных. – Томск. – 2002. (student\Колесникова\ Анализ данных)
Опубликовать спроектированную БД согласно указанному алгоритму.

Подготовка *занятия №И4.* Выбор ведущих студентов, ответственного за выбор и подачу необходимой информации и разработка с ними алгоритма занятия.

Таблица 4.1

№	№ задачи	Вид интерактивной работы (совмещение нескольких видов)	Трудо-емкость (час.)	Отрабатываемые компетенции/ ожидаемый уровень освоения	Оценка личностных качеств	Контроль выполнения работы (участие в полемике, индивидуальные групповые задания (ИГЗ) на базе выбранного программного продукта и т.д)
1.	И4.1	Работа в команде. Решение	2	ОК-1, ОК-3, ОК-12/ З-Эл, У-Эл,	Качество работы; своевременнос	ИГЗ. Критерии оценивания поведения на занятии: активность,

		ситуационных задач. Поисковый метод.		В-Эл ПК-26/ З-Пр, У-Пр, В-Пр	ть сдачи отчета по решению ИГЗ	инициативность, грамотность, обоснованность защищаемой позиции.
2	И4.2	Работа в команде. Решение ситуационных задач.	2	ОК-1, ОК-3, ОК-12/ З-Эл, У-Эл, В-Эл ПК-26/ З-Пр, У-Пр, В-Пр	Качество работы; своевременность сдачи отчета по решению ИГЗ	ИГЗ. Критерии оценивания поведения на занятии: активность, инициативность, грамотность, обоснованность защищаемой позиции.
Всего			4			

Вступление. **С**ообщение темы (далее, на примере задачи И4.1) и обоснование ее актуальности через логику задач математической статистики.

Основная часть:

I. **С**ообщение в виде доклада-презентации ответственными (студентами) за проведение занятия И4, в котором излагается суть обсуждаемого явления:

1) Обзор систем анализа распределённых данных и принципы их построения;
2) Существующие стандарты Data mining и OLAP;
3) Понятие «мобильных агентов» и системы мобильных агентов, используемых для анализа данных;
4) Метод инвариантных многообразий для управления сложными системами.

II. Выяснение позиций участников с зафиксированными точками зрения на решение задач И4.1-И4.2.

Итог II-го этапа: формирование целевых групп по общности позиций каждой из групп.

III. Организация коммуникации между группами: 1) выяснение позиции-варианта решения выявленных групп и защита занятой позиции; 2) формирование нового набора вариантов решений на основании общего обсуждения; 3) выбор одного решения голосованием;

IV. Повторная защита позиций-вариантов групп после проведения расчетов с целью оценки отклонения от «истинного» решения.

Выводы: реализован самостоятельный поиск учащимися путей и вариантов решения поставленной учебной задачи (выбор одного из предложенных вариантов или нахождение собственного варианта и обоснование решения на базе коллективной интерактивной работы).

Итог занятия №И4: Оценивание компетенций (ОК-1, ОК-3, ОК-12, уровни З-Эл, У-Эл, В-Эл; ПК-26 уровни З-Пр, У-Пр, В-Пр) по результатам работы на занятиях (активность, инициативность, грамотность, обоснованность защищаемой позиции) и своевременности сдачи отчета по решению практических задач согласно табл. 4.1.

ВАРИАНТЫ ДОМАШНИХ ЗАДАНИЙ К РАЗДЕЛУ 4

Подготовьте сообщения на один из заданных примеров по теме: «Некоторые бизнес-приложения Data Mining»

1. Розничная торговля
2. Банковское дело
3. Телекоммуникации
4. Страхование

Составитель: Колесникова С.И., каф.ЭМИС

5. Автопроизводители: развитие автомобильной промышленности.
6. Производители разной продукции: политика гарантий.
7. Авиакомпании: поощрение часто летающих клиентов.

ВАРИАНТЫ КОНТРОЛЬНЫХ ЗАДАНИЙ К РАЗДЕЛУ 4

Даны таблицы данных. Следует определить (найти) закономерности. Построить решающее правило. Обосновать выбор и вид решающего правила.

Вариант I

Данные по «угоняемости» автомобилей[4].

Таблица 4.2

Цвет	Тип	Производство	Повреждения (1/0=Да/Нет)	Признак угона (1, 0.-)
Красный	Спортивный	США	0	1
Желтый	Спортивный	Япония	0	1
Желтый	Джип	Япония	0	1
Красный	Спортивный	Япония	1	1
Желтый	Спортивный	США	1	0
Желтый	Джип	США	0	0
Красный	Джип	Япония	1	0
Желтый	Спортивный	США	1	0
Красный	Спортивный	Германия	0	1
Черный	Джип	Япония	0	-

Вариант II

Дана таблица данных. Следует определить (найти) закономерности. Построить решающее правило. Обосновать выбор и вид решающего правила.

Таблица 4.3

Классы	Объекты	Значения признаков			
		X_1	X_2	X_3	X_4
Ω_1	ω_1 ω_2 ω_3	1 1 0	1 0 0	1 1 1	1 1 0
Ω_2	ω_4 ω_5 ω_6	1 0 0	0 0 0	1 0 1	0 0 1
	ω'	1	1	0	0

Вариант III

Дана таблица данных. Следует определить (найти) закономерности. Построить решающее правило. Обосновать выбор и вид решающего правила.

Таблица 4.4

Классы	Объекты	Значения признаков

[4] http://www.interface.ru/home.asp?artId=24809

Составитель: Колесникова С.И., каф.ЭМИС

		X_1	X_2	X_3	X_4	X_5	X_6
Ω_1	ω_{11} ω_{12} ω_{13} ω_{14} ω_{15}	5 4 9 7 2	11 10 5 13 14	9 2 4 3 8	3 7 6 4 5	3 12 11 6 9	1 1 1 2 1
Ω_2	ω_{12} ω_{22} ω_{23} ω_{24} ω_{25}	5 4 6 7 3	9 6 11 10 10	2 7 9 4 5	8 3 11 2 9	14 13 5 12 7	1 1 1 1 1
	ω'	3	13	7	4	8	2

Вариант IV

Даны матрицы Q - описание 6-ти объектов; R - соответствие номеров объектов и классов. Следует определить (найти) закономерности. Построить решающее правило. Обосновать выбор и вид решающего правила.

$$Q = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & z_3 & z_4 \end{array} \\ \begin{bmatrix} 0 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 2 & 1 & 0 \\ 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 2 \end{bmatrix} \end{array}; \; R = \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 2 \\ 2 \\ 2 \end{bmatrix}.$$

Контрольные вопросы к разделу 4

1. Перечислить основные шаги при осуществлении OLAP-анализа.
2. Сформулировать понятие «мобильный агент».
3. Изложить суть метода инвариантных многообразий.
4. Что понимается под стандартами Data mining.
5. Назовите стандартные типы закономерностей, которые позволяют выявлять методы Data Mining.

Составитель: Колесникова С.И., каф.ЭМИС

При составлении методических указаний использовался материал нижеуказанной литературы, а также материал интернет-ресурсов.

ИСПОЛЬЗОВАННАЯ ЛИТЕРАТУРА

1. Васильев В.А. Непараметрическое оценивание функционалов от распределений стационарных последовательностей / В.А. Васильев, А.В. Добровидов, Г.М. Кошкин. – М.: Наука, 2004. – 508 с
2. Воронцов К.В. Лекции по методам оценивания и выбора моделей. 2007. Режим доступа: www.ccas.ru/voron/download/Modeling.pdf.
3. Воронцов К.В. Обзор современных исследований по проблеме качества обучения алгоритмов. Таврический вестник информатики и математики. – 2004. – № 1. – С. 5 – 24. http://www.ccas.ru/frc/papers/voron04twim.pdf.
4. Горелик А. Л., Скрипкин В. А. Методы распознавания: Учебное пособие для вузов. - 4-е изд., испр. - М.: Высшая школа, 2004. – 260 с.
5. Дюк В. Обработка данных на ПК в примерах / Вячеслав Дюк. - СПб.:Питер, 1997. - 240с.
6. З. Брандт. Анализ данных. Статистические и вычислительные методы для научных работников и инженеров: Пер. с англ.: Учебное пособие/ З. Брандт; пер.: О.И.Волкова; ред. пер.: Е. В. Чепурин. - М. :Мир, 2003 ; М.: АСТ, 2003. – 686 с.
7. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. - Новосибирск: ИМ СО РАН, 1999, стр. 270 http://www.sernam.ru/book_zg.php
8. Кокс, Д. Р. Анализ данных типа времени жизни : Пер. с англ. / Д. Р. Кокс, Д. Оукс; Пер. О. В. Селезнева, Ред. Ю. К. Беляев, Предисл. Ю. К. Беляев. - М. : Финансы и статистика, 1988. – 189с.
9. Лапко А.В. Непараметрические системы обработки информации : Учебное пособие для вузов / А. В. Лапко, С. В. Ченцов; Российская Академия наук. Сибирское отделение, Институт вычислительного моделирования. - М. : Наука, 2000. - 349 с.
10. Прикладная статистика. Классификация и снижение размерности : справочное издание / С. А.Айвазян [и др.] ; ред. С. А. Айвазян. - М. : Финансы и статистика, 1989. - 608 с.
11. Тюрин, Ю.Н.. Анализ данных на компьютере : учебное пособие для вузов / Ю. Н. Тюрин, А. А. Макаров. - 4-е изд., перераб. - М. : Форум, 2008. – 366.
12. http://itteach.ru/predstavlenie-znaniy/metodi-i-sredstva-intellektualnogo-analiza-dannichhttp://www.bsu.ru/content/hec/biometria/modules/stdatmin.html
13. http://www.arshinov74.ru/files/files/10.pdf
14. http://www.intuit.ru/department/database/dataanalysis/
15. http://www.olap.ru/basic/dm2.asp
16. http://www.sernam.ru/book_zg.php
17. http://www.statlab.kubsu.ru/node/4

Составитель: Колесникова С.И., каф.ЭМИС