

МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ
Федеральное государственное автономное образовательное учреждение
высшего образования
«Национальный исследовательский Томский государственный университет»
Механико-математический факультет

«УТВЕРЖДАЮ»

Декан ММФ _____ А.В. Старченко

" ____ " _____ 2014 г.

Рабочая программа дисциплины (модуля)

СОВРЕМЕННЫЙ АНАЛИЗ И ВИЗУАЛИЗАЦИЯ ДАННЫХ

Направление подготовки

01.04.01 Математика

Магистерская программа

Математическое моделирование

Квалификация (степень) выпускника

Магистр

Форма обучения

Очная

Томск - 2014

1. Цели и задачи освоения дисциплины

Целями и задачами освоения дисциплины «Современный анализ и визуализация данных» являются

1.1. Подготовка магистров по направлению «Математика» к использованию современных методов анализа и визуализации данных в профессиональной деятельности,

1.2. Освоение понятий и методов анализа и визуализации данных, закрепления основных теоретико-вероятностных представлений, сформированных ранее в курсах теории вероятностей, математической статистики и многомерного статистического анализа.

1.3. Умение провести статистический анализ с использованием макетов Statistica и MATLAB.

2. Место дисциплины в структуре ООП магистратуры

Дисциплина «Современный анализ и визуализация данных» является обязательным курсом в 11 семестре.

Для изучения курса необходимо освоить курсы «Математический анализ», «Геометрия и линейная алгебра», «Теория вероятностей и математическая статистика», «Компьютерные науки и программирование».

В курсе «Современный анализ и визуализация данных» магистрант получает опыт применения знаний, полученных в математических курсах и курсе компьютерные науки для формирования теоретико-вероятностных моделей данных прикладных задач, проведения анализа данных в пакетах Statistica и MATLAB.

3. Компетенции обучающегося, формируемые в результате освоения дисциплины (модуля) «Современный анализ и визуализация данных»:

Общекультурные:

- способность общаться со специалистами из других областей (ОК-2);
- способность работать в международной среде (ОК-3);
- значительные навыки самостоятельной научно-исследовательской работы и научно-изыскательской работы, а также деятельности в составе группы (ОК-6);
- способность к постоянному совершенствованию и углублению своих знаний, инициативность и стремление к лидерству (ОК-7);

Профессиональные:

- владение методами математического и алгоритмического моделирования при анализе проблем естествознания (ПК-2);
- самостоятельный анализ физических аспектов в классических постановках математических задач (ПК-4);
- умение публично представить собственные новые научные результаты (ПК-5);
- умение ориентироваться в современных алгоритмах компьютерной математики, совершенствовать, углублять и развивать математическую теорию, лежащую в их основе (ПК-7);
- способность к творческому применению, развитию и реализации математически сложных алгоритмов в современных программных комплексах (ПК-9);
- определение общих форм, закономерностей, инструментальных средств для групп дисциплин (ПК-10);
- умение извлекать актуальную научно-техническую информацию из электронных библиотек, реферативных журналов (ПК-16).

В результате освоения дисциплины обучающийся должен овладеть:

- ✓ **знаниями** методов многомерного анализа данных: дискриминантного анализа; регрессионного анализа, кластерного анализа, метода опорных векторов,
- ✓ **умением** имитации наборов данных для тестирования методов и отладки алгоритмов, графического представления данных, визуализации территориально-распределенных данных,
- ✓ **навыками** проведения анализа и визуализации данных в пакетах Statistica и MATLAB.

4. Структура и содержание дисциплины «Современный анализ и визуализация данных»: Общая трудоемкость дисциплины составляет 5 зачетных единиц 180 часов.

№ п/п	Раздел Дисциплины	Семестр	Неделя семестра	Виды учебной работы, включая самостоятельную работу студентов и трудоемкость (в часах)			Формы текущего контроля успеваемости (по неделям семестра) Форма промежуточной аттестации (по семестрам)
				Лекции	Лабораторные работы	СРС	
1	Данные	3	1	2	2	20	Тест 1
2	Структура пакетов Statistica и MATLAB.	3	2	4	2	10	Индивидуальное задание 1
3	Описание Данных. Описательные Статистики.	3	3		2	20	Индивидуальное задание 1
4	Зависимость данных.	3	3-6	4	6	20	Индивидуальное задание 2
5	Множественная линейная регрессия	3	7-9	4	4	6	Лабораторная работа 1 Коллоквиум
6	Методы решения задачи классификации	3	10-12	4	4	24	Тест 2 Лабораторная работа 2

7	Проблема отбора наиболее информативных признаков.	3	12	2	2	10	Индивидуальная работа 3
8	Методы кластерного анализа.	3	13	2	2	20	Коллоквиум Лабораторная работа 3
9	Метод опорных векторов (SVM).	3	14-15	6	4	12	Коллоквиум
10	Кластеризация в Statistica и MATLAB.	3	16	4	4	10	Лабораторная работа 4
	ИТОГО:			32	32	152	Зачет

Темы и краткое содержание дисциплины

- 1) **Данные.** Качество данных. Структура и объем данных. Этапы и разновидности анализа данных.
- 2) **Структура пакетов Statistica и MATLAB.** Импорт и экспорт данных. Общие вопросы выполнения анализа и визуализации данных в пакетах Statistica и MATLAB.
- 3) **Описание Данных. Описательные Статистики.** Моделирование выборки одномерной и многомерной случайной величины на ЭВМ. Модели данных и задача прогнозирования. Алгоритм проверки гипотез. Критерии согласия Пирсона, Колмогорова-Смирнова, Вилкоксона-Манна-Уитни в пакетах Statistica и MATLAB.
- 4) **Зависимость данных.** Алгоритм однофакторного дисперсионного анализа. Численные методы выявления корреляции количественных признаков. Методы оценивания корреляции порядковых и номинальных признаков. Построение таблицы сопряженных признаков в пакетах Statistica и MATLAB.
- 5) **Методы решения задачи классификации.** Частные постановки задач классификации и прогноза из различных областей знаний: биологии и медицины; экологии; экономики и страхового дела. Факторный анализ. Дискриминантный анализ в пакетах Statistica и MATLAB.
- 6) **Множественная линейная регрессия** в пакетах Statistica и MATLAB. Probit и Logit регрессия в пакете Statistica.
- 7) **Проблема отбора наиболее информативных признаков.** Задача снижения размерности вектора признаков. Реализации безмодельной процедуры отбора тарификационных переменных и проведения численного эксперимента в пакетах Statistica и MATLAB.
- 8) **Методы кластерного анализа.** Функции расстояния и близости. Расстояния между кластерами. Параллельная реализация процедуры предварительного обнаружения кластеров. Эвристические алгоритмы «Форель» и «Краб» и их реализация в пакетах Statistica.
- 9) **Метод опорных векторов (SVM).** Метод опорных векторов в задачах классификации. Понятие оптимальной разделяющей гиперплоскости. Линейно

разделимая выборка. Линейно неразделимая выборка. Ядра и спрямляющие пространства.

10) **Расстояния между кластерами** в пакетах Statistica и MATLAB. Иерархическая агломеративная кластеризация и метод К-средних в пакетах Statistica и MATLAB.

5. Образовательные технологии

Методы \ ФОО	Лекции	Лаб. раб.	СРС
IT-методы	*	*	*
Работа в команде		*	*
Методы проблемного обучения.	*	*	*
Опережающая самостоятельная работа		*	*
Проектный метод		*	*
Поисковый метод		*	*
Исследовательский метод	*	*	*

6. Учебно-методическое обеспечение самостоятельной работы студентов. Оценочные средства для текущего контроля успеваемости, промежуточной аттестации по итогам освоения дисциплины.

6.1. Самостоятельную работу студентов (СРС) можно разделить на текущую и творческую.

Текущая СРС – Проработка лекций, изучение рекомендованной литературы.

Творческая проблемно-ориентированная самостоятельная работа

(ТСР) – Анализ источников по темам индивидуальных заданий, поиск существующих аналогов. Создание программ средств, реализующих разрабатываемые алгоритмы.

6.2. Содержание самостоятельной работы студентов по дисциплине

Самостоятельная работа организуется в двух формах:

- аудиторной (на лабораторных работах при решении поставленных и индивидуальных задач);
- внеаудиторной (проработка лекций – 18 часов; изучение рекомендованной литературы по темам, выносимым на самостоятельное изучение – 30 часов; подготовка к тесту – 4 часа; подготовка к контрольной работе – 4 часа; подготовка к выполнению лабораторной работы – 6 часов; выполнение индивидуальных работ – 24 часа; оформление отчетов по индивидуальным и лабораторным работам -6 часа; подготовка к коллоквиуму и зачету - 12).

6.3. Контроль самостоятельной работы

Контроль результатов самостоятельной работы осуществляется при проведении письменных тестов, самостоятельных, контрольных работ, устных коллоквиумах по проверке уровня усвоения студентом лекционного материала и проверкой уровня освоения студентом теоретических знаний и практических навыков при выполнении и защите им лабораторных работ и индивидуальных работ.

6.4. Учебно-методическое обеспечение самостоятельной работы студентов

Студентам для самостоятельной работы предлагается учебно-методическое обеспечение дисциплины в электронном виде, презентации лекций, учебники.

6.5. Текущий и итоговый контроль оценки качества

Текущий контроль оценки качества усвоения дисциплины заключается в проведении теста, индивидуальных, четырех лабораторной и двух коллоквиумов.

Для теста и коллоквиума подготовлены списки вопросов. Тестирование проводится в системе Moodle ТГУ. Для успешной сдачи коллоквиумов студент устно в режиме реального времени должен ответить на 5 вопросов из указанного списка. Во время выполнения лабораторных работ преподаватель на основе серии контрольных вопросов проверяет теоретические знания студента по темам лабораторной работы. При защите индивидуальной работы студент излагает теоретическое обоснование постановки задачи и методов ее решения. Для зачета сформированы билеты. В каждом билете содержится 2 вопроса и одно практическое задание.

7. Учебно-методическое и информационное обеспечение дисциплины (модуля)

а) основная литература:

1. Айвазян С.А., Мхитарян В.С. Теория вероятностей и прикладная статистика. Том 1. М.: ЮНИТИ-ДАНА, 2001. – 656 с.
2. Гайдышев И. Анализ и обработка данных. Специальный справочник. С.Пб.: ПИТЕР, 2001. – 751с.
3. Боровиков В. П. Боровиков И. П. STATISTICA. – Статистический анализ и обработка данных в среде Windows. – М.: Информационно-издательский дом «Филинь», 1997. – 608с.
4. Дубров А. М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы. М.: Финансы и статистика, 2003. – 352 с.
5. Глинский В.В., Ионин В.Г. Статистический анализ: Учебное пособие. М.: ИНФРА-М; Новосибирск: Сибирское соглашение, 2002. – 241 с.
6. Мартынов Н.Н., Иванов А.П. Вычисления, визуализация, программирование в среде MatLab 5.x. Издательство: КУДИЦ-ОБРАЗ, 2002. – 336 с.
7. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Изд. ИМ СО РАН, 1999. – 270 с.

б) дополнительная литература:

1. Кокс Д., Снелл Э. Прикладная статистика. Принципы и примеры. М.: Мир, 1984 – 200с.
2. Андерсон Т. Введение в многомерный анализ.- М.: Физматгиз, 1963. – 500 с.
3. Афифи А.А., Эйзен С.П. Статистический анализ: Подход с использованием ЭВМ.– Пер. с англ.– М.: Мир, 1982.– 488 с.
4. Голенко Д.И. Моделирование и статистический анализ псевдослучайных чисел на ЭВМ.– М: Наука, 1965.– 228 с.
5. Левитан Ю.Л., Соболев И.М. О датчике псевдослучайных чисел для персональных компьютеров // Математическое моделирование.– 1990.– Т.2, №8.– с. 119-126.

6. Осоков Г.А., Тихоненко Е.А. Новый генератор случайных чисел на базе двумерного клеточного автомата // Математическое моделирование.– 1996.– Т.8, № 12.– с. 77-84.
7. Смирнов Н. В. Приближение законов распределения случайных величин по эмпирическим данным // Усп. матем. наук.– 1944, 10.–с. 179-206.
8. Соболев И.М. Численные методы Монте-Карло.– М: Наука, 1973.– 312с.
9. Фёдорова О.П., Бакакина Ж.В., Каминская Е.В. исследование датчиков случайных чисел// Моделирование неравновесных систем. – 2000:Материалы III Всероссийского семинара, 20-22 октября 2000., Красноярск/ Красноярск: ИПЦ КПТУ.2000, 268-269 стр.
10. Гольдин В.Д., Федорова О.П. Алгоритм выявления аномалий// Вычислительная гидродинамика. – Томск: ТГУ, 1999. с.21-26.
11. Федорова О.П. Выявление аномалий координатно-привязанных объектов // Математическое моделирование и теория вероятностей. -Томск: Пеленг, 1998. с.270-275.
12. Бородачева А.В. Аппроксимация сплайнами функций с заданными свойствами. Дипломная работа. Рукопись.- Томск: ММФ ТГУ, 2003.- 66с.
13. Сокал Р. Р. Кластер-анализ и классификация: предпосылки и основные направления.// Классификация и кластер. – М.: Мир, 1980. – 230с.
14. Апраушева Н. Н. Предварительное обнаружение идеальных кластеров и оценивание их числа. – М.:ВЦ АН СССР, 1987. – 22с.
15. Апраушева Н. Н. Три алгоритма естественной кластеризации объектов. – М.:ВЦ АН СССР, 1986. – 21с.
16. Апраушева Н. Н. Новый подход к обнаружению идеальных кластеров. – М.:ВЦ РАН, 1993. – 61с.
17. Загоруйко Н. Г., Елкина В. Н., Лбов Г. С. Алгоритмы обнаружения эмпирических закономерностей. – Новосибирск.: Наука, 1985. – 106с.
18. Борисова И., Загоруйко Н., Кутненко. Критерии информативности и пригодности подмножества признаков. Proceedings of the 8-th International Conf. KDS – 2007. Sofia Изд-во ITNEA, 2007. Vol. 2, p.567-571.
19. Перегудов Ф. И., Тарасенко Ф. П. Введение в системный анализ: учебное пособие для вузов. – М.: Высш. Шк., 1989. – 367с.

в) примеры теста 1 и вопросы для формирования теста 2:

Тест 1, пример варианта

- 1)** Вычислить максимальное, минимальное значения и размах выборки:
3.480; 5.293; 1.066; -0.040; -0.470; 0.031; 4.158; -2.844; -2.244; 0.311; 0.897;
5.545; 3.153; 4.514; -2.236; -2.080; -0.869; 4.399; 4.581;
- 2)** Построить вариационный статистический ряд дискретной случайной величины, вычислить выборочные значения медианы, моды и построить гистограмму
21; 20; 19; 19; 21; 20; 20; 19; 21; 21; 19; 20; 19; 19; 21; 21; 19; 21; 21; 19; 19; 20;
21; 21;
- 3)** Вычислить оценки среднего и дисперсии выборки:
9.0; -3.0; 12.0;

Тест 2, вопросы для формирования теста

- 1) Цели факторного анализа.
- 2) Задачи факторного анализа.
- 3) Предпосылки факторного анализа.
- 4) Фундаментальная теорема.
- 5) Как выполнить стандартизацию данных.
- 6) Что такое факторная нагрузка.
- 7) Что такое факторное отображение.
- 8) На какие типы подразделяются факторы.

- 9) Сформулируйте основную задачу дискриминантного анализа.
- 10) Что называют дискриминантной функцией?
- 11) Сформулируйте основную задачу дискриминантного анализа при наличии обучающих выборок.
- 12) Какие методы дискриминации называются непараметрическими?
- 13) Какие методы дискриминации называются параметрическими?
- 14) Опишите линейный дискриминантный анализ в случае известных параметров нормальных совокупностей.
- 15) Напишите алгоритм линейного дискриминантного анализа нормальных совокупностей, если даны обучающие выборки для каждого класса.
- 16) Сформулируйте проблему снижения размерности в анализе данных.
- 17) Какие методы анализа данных позволяют решить проблему снижения размерности?
- 18) Сформулируйте основную идею уменьшения числа информативных бинарных признаков, лежащую в основе безмодельной процедуры отбора.
- 19) Опишите подходы к построению параллельной реализации безмодельной процедуры отбора.
- 20) Дайте определение расстояния (метрики)?
- 21) Запишите формулу вычисления евклидова расстояния между двумя k – мерными векторами X и Y .
- 22) Запишите формулу вычисления L_1 расстояния между двумя k – мерными векторами X и Y .
- 23) Запишите формулу вычисления расстояния Хемминга между двумя k – мерными двоичными векторами X и Y .
- 24) Запишите и объясните формулу вычисления расстояния Махаланобиса между двумя k – мерными векторами X и Y .
- 25) Как переводится английское слово Cluster?
- 26) В чем состоит задача кластерного анализа?
- 27) Какие иерархические методы кластерного анализа вы знаете?
- 28) В чем суть агломеративных иерархических методов кластерного анализа?
- 29) В чем суть дивизимных иерархических методов кластерного анализа?
- 30) Опишите основные этапы кластеризации с помощью итеративного метода k -средних.
- 31) Перечислите виды расстояний между двумя кластерами.
- 32) Запишите формулу расстояния по принципу «ближнего соседа» между двумя кластерами.
- 33) Запишите формулу расстояния по принципу «дальнего соседа» между двумя кластерами.
- 34) Запишите формулу расстояния по «центрам тяжести» между двумя кластерами.
- 35) Запишите формулу расстояния по принципу «средней связи» между двумя кластерами.
- 36) В чем состоит идея Уорда (Ward) объединения объектов в кластеры?
- 37) Что такое дендрограмма

г) примеры индивидуальных заданий:

1. Представление данных (построение гистограмм) и исследование взаимозависимости морфометрических признаков особей муравьев из района падения Тунгусского метеорита и визуализация результатов исследования.
2. Представление данных и анализ взаимозависимости признаков, описывающих иммунитет человека (по данным иммунологической лаборатории Томского онкоцентра) и визуализация результатов исследования.

3. Представление данных и анализ взаимозависимости показателей экономической деятельности предприятий и визуализация результатов исследования.
4. Представление данных и анализ взаимозависимости показателей внутренней и внешней миграции в Томской области в период 2000 – 20013 гг и визуализация результатов исследования.
5. Представление данных и анализ взаимозависимости социально-экономических показателей народонаселения Томской области в период 2000 – 20013 гг и визуализация результатов исследования.

д) пример лабораторной работы:

Провести кластерный анализ предприятий машиностроения по значениям показателей производственно-хозяйственной деятельности.

Задание 1. Провести эксперимент, в котором сравниваются результаты анализа для случая различных заданий начальных центров кластеров:

- ✓ **Choose observations to maximize initial between-cluster distances**
- ✓ **Choose the first N(Number of clusters) observations.**

Задание 2. Перенести таблицы для каждого кластера в рабочую тетрадь.

Задание 3. Для отчета представить графики средних значений по кластерам (рис.7) в формате, изображенном на рис.7.(шрифт Times New Roman 10pt, размеры окна графика: ширина 8 см, высота 7 см). Процесс форматирования описать в отчете.

Задание 4. Сравнить результаты анализа для случая различных заданий начальных центров кластеров и представить в отчете.

Задание 5. Выполнить кластерный анализ для своего варианта при $k=2$; $k=3$; $k=4$. Провести анализ получившихся результатов.

е) задачи для самостоятельной работы:

1. Составить и отладить программу в MATLAB вычисления аналитического решения задачи Коши системы неоднородных линейных ОДУ вида $\frac{dX}{dt} = AX + B$, где X - вектор столбец неизвестных, A – квадратная матрица, B – вектор столбец свободных членов;
2. Составить и отладить программу в MATLAB вычисления численного решения задачи Коши системы неоднородных линейных ОДУ вида $\frac{dX}{dt} = AX + B$, где X - вектор столбец неизвестных, A – квадратная матрица, B – вектор столбец свободных членов, неявным методом Эйлера.

ж) пример задачи для контрольной работы:

Познакомиться с описанием предметной области (см. статья Каменецкий И.С. Математическая статистика в археологии // Природа, 1979, № 9. – С. 51-59). Составить математическую модель процесса накопления обломков амфор в слоях.

Провести расчеты.

з) список вопросов для зачета:

1. В чем состоят цели многомерного анализа данных?
2. Дайте формальное описание данных, используемое в многомерном статистическом анализе и опишите представление данных в электронном виде.
3. Функция распределения и плотность вероятностей непрерывной многомерной случайной величины.

4. Подсистема многомерной случайной величины. Маргинальные распределения.
5. Условные распределения. Независимость случайных величин и подсистем случайных величин.
6. Моменты l – го порядка: относительно компонент некоторого вектора, центральные, начальные. Математическое ожидание, дисперсии.
7. Ковариационная и корреляционная матрица и их свойства.
8. Многомерное нормальное распределение. Зависимости между нормально распределенными случайными величинами.
9. Генеральная совокупность, выборка. Репрезентативность выборки.
10. Определите основные шкалы измерений в многомерных выборках.
11. Как произвести преобразование данных из одной шкалы в другую?
12. Какие и как рассчитываются основные характеристики многомерной выборки в предварительном анализе данных?
13. Определите средства визуализации данных в предварительном анализе данных?
14. Статистические гипотезы. Проверка статистической гипотезы. Ошибки 1-го и 2-го рода. Критерии проверки статистической гипотезы.
15. Критерий согласия Хи-квадрат.
16. Два взгляда на выборку. Статистики. Определения несмещенности, состоятельности и эффективности.
17. Оценки математического ожидания, дисперсии, стандартного (среднего квадратического отклонения).
18. Оценки элементов ковариационной матрицы и коэффициентов парных корреляций. Поэлементная и матричная запись.
19. Модель множественной регрессии.
20. Постановка задачи кластерного анализа.
21. Охарактеризовать типы кластерного анализа.
22. Расстояния между объектами.
23. Расстояния между кластерами.
24. Метод k-means.
25. Метод Forel.
26. Постановка задачи факторного анализа.
27. Основная теорема факторного анализа.
28. Факторные нагрузки и факторное отображение.
29. Типы факторов.
30. Проблемы факторного анализа и их решение.
31. Общая характеристика методов дискриминантного анализа.

8. Материально-техническое обеспечение дисциплины (модуля)

Для проведения лабораторных работ и самостоятельной работы используются аудитории 314, 316, оснащенные (каждая) компьютерами (13 шт.), LCD мониторами BENQ 21.5”, имеющими процессоры Intel core i5-2400, с тактовой частотой 3.40 ГГц, оперативной памятью: 4 Гб, жестким диском (винчестер) 500 Гб, видеокартой Nvidia GTS 450.

Свободным и лицензионным программным обеспечением, которое включает в себя

- операционные системы: Microsoft Windows XP, Microsoft Windows 7, GNU/Linux SLES 10, GNU/Linux CentOS 6;

- офисные и издательские пакеты Microsoft Office 2003, Microsoft Office 2010, MikTeX 2.9;
- средства разработки приложений и СУБД Microsoft Visual Studio 2010, Delphi 2006 (для работы с базами данных - Borland Database Engine, Database Desktop), Lazarus, Borland Pascal, PascalABC.NET, Intel Fortran Compiler 12, CUDA Toolkit 4;
- математические пакеты PTC Mathcad 15, Mathematica 8, Maple 15, Matlab R2011b; Statistica

В образовательном процессе также используются пакеты математической и графической обработки данных Golden Software Grapher, Golden Software Surfer; пакеты для решения задач вычислительной гидродинамики Ansys CFD 14, Fluent Flowlab; софт для удаленного доступа Winscp, Putty, FreeNX

Программа составлена в соответствии с требованиями ФГОС ВПО с учетом рекомендаций и ПрООП ВПО по

Направлению подготовки: 01.04.01 «Математика»

По программе магистратуры: «Математическое моделирование»

Автор

К.ф.-м.н., доцент Федорова О.П. _____

К.ф.-м.н., доцент Пчелинцев Е.А. _____

Рецензент

Д.ф.-м.н., профессор _____

Программа одобрена на заседании *методической комиссии* ММФ

от 19 декабря 2014 г., протокол № 13.

Председатель методической комиссии ММФ _____ Федорова О.П.