

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ
федеральное государственное автономное
образовательное учреждение высшего образования
«Национальный исследовательский
Нижегородский государственный университет им. Н.И. Лобачевского»

УТВЕРЖДЕНО

решением ученого совета ННГУ

протокол от "30"ноября 2022 г. №13

Рабочая программа дисциплины
«Информационные технологии многомерного статистического
анализа»

Уровень высшего образования

Подготовка научных и научно-педагогических кадров

Научные специальности

1.1.2. Дифференциальные уравнения и математическая физика, 1.1.4. Теория вероятностей и математическая статистика, 1.1.5. Математическая логика, алгебра, теория чисел и дискретная математика, 1.1.8. Механика деформируемого твердого тела, 1.2.1. Искусственный интеллект и машинное обучение, 1.2.2. Математическое моделирование, численные методы и комплексы программ, 1.3.11. Физика полупроводников, 1.3.19. Лазерная физика, 1.3.4. Радиофизика, 1.3.7. Акустика, 1.3.8. Физика конденсированного состояния, 1.4.1. Неорганическая химия, 1.4.2. Аналитическая химия, 1.4.3. Органическая химия, 1.4.4. Физическая химия, 1.4.7. Высокомолекулярные соединения, 1.4.8. Химия элементоорганических соединений, 1.5.11. Микробиология, 1.5.15. Экология, 1.5.2. Биофизика, 1.5.21. Физиология и биохимия растений, 1.5.5. Физиология человека и животных, 2.2.2. Электронная компонентная база микро и нанoeлектроники, квантовых устройств, 3.2.7. Аллергология и иммунология, 5.1.1. Теоретико-исторические правовые науки, 5.1.2. Публично-правовые (государственно-правовые) науки, 5.1.3. Частно-правовые (цивилистические) науки, 5.1.4. Уголовно-правовые науки, 5.1.5. Международно-правовые науки, 5.12.1. Междисциплинарные исследования когнитивных процессов, 5.2.3. Региональная и отраслевая экономика, 5.2.4. Финансы, 5.2.6. Менеджмент, 5.3.7. Возрастная психология, 5.4.2. Экономическая социология, 5.4.4. Социальная структура, социальные институты и процессы, 5.4.6. Социология культуры, 5.4.7. Социология управления, 5.5.2. Политические институты, процессы, технологии, 5.5.4. Международные отношения, глобальные и региональные исследования, 5.6.1. Отечественная история, 5.6.2. Всеобщая история, 5.6.7. История международных отношений и внешней политики, 5.7.1. Онтология и теория познания, 5.8.2. Теория и методика обучения и воспитания, 5.8.7. Методология и технология профессионального образования, 5.9.2. Литературы народов мира, 5.9.5. Русский язык. Языки народов России, 5.9.6. Языки народов зарубежных стран (с указанием конкретного языка или группы языков), 5.9.9. Медиакоммуникации и журналистика

Нижний Новгород

2023 год

1. Место и цель дисциплины в структуре ОПОП

Дисциплина «Информационные технологии многомерного статистического анализа» относится к числу факультативных дисциплин образовательного компонента программы аспирантуры и изучается на 3 году обучения в 5 семестре.

Цель дисциплины – изучение современных методов анализа многомерных данных

2. Планируемые результаты обучения по дисциплине

Выпускник, освоивший программу, должен

Знать:

– математическую постановку задач классификации с обучением и без обучения, кластерного анализа, факторного анализа, метода главных компонент, многомерного шкалирования

Уметь:

– решать практические задачи методами многомерного статистического анализа

Владеть:

– современными информационными технологиями выполнения процедур многомерного статистического анализа

3. Структура и содержание дисциплины.

Объем дисциплины (модуля) составляет 2 з.е., всего - 72 часа, из которых 36 часов составляет контактная работа обучающегося с преподавателем (занятия лекционного типа – 18 часов, лабораторные работы – 18 часов), 36 часов составляет самостоятельная работа обучающегося.

Таблица 2

Структура дисциплины

Наименование раздела дисциплины	Всего, часов	В том числе					
		Контактная работа, часов					Самостоятельная работа обучающегося, часов
		Занятия лекционного типа	Занятия семинарского типа	Занятия лабораторного типа	Консультации	Всего	
Распознавание образов и типологизация объектов.	32	8	0	8	0	16	16
Снижение размерности признакового пространства	24	6	0	6	0	12	12
Многомерное шкалирование	16	4	0	4	0	8	8
Промежуточная аттестация: – зачет							
Итого	72	18	0	18	0	36	36

Таблица 3

Содержание дисциплины

№ п/п	Наименование раздела дисциплины	Содержание раздела	Форма проведения занятия	Форма текущего контроля*
1.	Распознавание образов и типологизация объектов.	Классификация с обучением: дискримантный анализ.	лекции, лабораторные	собеседование, отчеты по лабораторным

		Классификация без обучения: кластерный анализ	работы	работам
2.	Снижение размерности признакового пространства	Факторный анализ. Метод главных компонент.	лекции, лабораторные работы	собеседование, отчеты по лабораторным работам
3.	Многомерное шкалирование	Задача многомерного шкалирования. Меры близости на основе условных вероятностей. Метрические меры различия профилей. Модель Торнгенсона. Неметрические методы. Снижение размерности признакового пространства.	лекции, лабораторные работы	собеседование, отчеты по лабораторным работам

4. Формы организации и контроля самостоятельной работы обучающихся

В процессе изучения дисциплины применяются лекционные и семинарские занятия. Применяются следующие виды лекций: 1) лекция-информация (ориентированная на изложение и объяснение научной информации, подлежащей осмыслению и запоминанию); 2) лекция-беседа (предполагающая непосредственное общение со слушателями посредством вопросов информационного, проблемного характера и вопросов для напоминания пройденного материала). Самостоятельная работа реализуется в форме изучения лекций и выполнения домашних заданий. Самостоятельная работа контролируется преподавателем как в ходе аудиторных занятий, так и во время внеаудиторной работы, в том числе посредством консультаций по электронной почте, видеосвязи, в социальных сетях.

5. Фонд оценочных средств для аттестации по дисциплине

5.1. Критерии и процедуры оценивания результатов обучения по дисциплине.

При выполнении всех работ учитываются следующие **основные критерии**:

- уровень теоретических знаний (подразумевается не только формальное воспроизведение информации, но и понимание предмета, которое подтверждается правильными ответами на дополнительные, уточняющие вопросы, заданные членами комиссии);
- умение использовать теоретические знания при анализе конкретных проблем, ситуаций;
- качество изложения материала, то есть обоснованность, четкость, логичность ответа, а также его полнота (то есть содержательность, не исключающая сжатости);
- способность устанавливать внутри- и межпредметные связи,
- оригинальность мышления, знакомство с дополнительной литературой и другие факторы.

Описание шкалы оценивания на промежуточной аттестации в форме зачета

Оценка	Уровень подготовленности, характеризующий оценкой
<i>Зачтено</i>	владение программным материалом, понимание сущности рассматриваемых процессов и явлений, умение самостоятельно обозначить проблемные ситуации в организации научных исследований, способность критически анализировать и сравнивать существующие подходы и методы к оценке результативности научной деятельности, свободное владение источниками, умение четко и ясно излагать результаты собственной работы, следовать нормам, принятым в научных дискуссиях.
<i>Не зачтено</i>	непонимание смысла ключевых проблем, недостаточное владение науковедческой терминологией, неумение самостоятельно обозначить проблемные ситуации, неспособность анализировать и сравнивать существующие концепции, подходы и методы, неумение ясно излагать результаты собственной работы, следовать нормам, принятым в научных дискуссиях.

5.2. Примеры типовых контрольных заданий или иных материалов, используемых для оценивания результатов обучения по дисциплине

Вопросы для собеседования

1. Многомерный статистический анализ: его особенности и технологии. Основные разделы многомерного статистического анализа.
2. Кластерный анализ. Области применения кластерного анализа. Расстояние между объектами. Меры сходства. Расстояние между кластерами.
3. Кластерный анализ. Метод k – средних. Метод поиска сгущений. Функционалы качества разбиения. Число кластеров.
4. Кластерный анализ. Иерархический кластерный анализ. Дендограммы.
5. Примеры применения кластерного анализа. Использование кластерного анализа для характеристики деятельности коммерческих банков на основе данных журнала "Профиль".
6. Дискриминантный анализ. Проблемы классификации. Принципы правильной классификации. Функции потерь и вероятности неправильной классификации. Решение задачи байесовских процедур классификации.
7. Дискриминантный анализ. Параметрический дискриминантный анализ в случае нормальных классов Две генеральные совокупности, имеющие известные многомерные нормальные распределения и равные матрицы ковариаций.
8. Дискриминантный анализ. Параметрический дискриминантный анализ в случае нормальных классов Две генеральные совокупности, имеющие известные многомерные нормальные распределения и неравные матрицы ковариаций.
9. Дискриминантный анализ. Параметрический дискриминантный анализ в случае нормальных классов Многомерные нормальные распределения, параметры которых оцениваются по выборке.
10. Дискриминантный анализ. Классификация наблюдений в случае нескольких генеральных совокупностей. Определение числа и вида дискриминирующих функций. Классификация объектов с помощью функции расстояния
11. Дискриминантный анализ. Классификация без обучения (параметрический случай): расщепление смесей вероятностных распределений. Определение вероятности ошибки дискриминации. ЕМ-алгоритм.
12. Применение дискриминантного анализа: диагностика причин кризисного состояния и банкротства предприятий.

13. Факторный анализ. Основные понятия факторного анализа.
14. Метод главных компонент. Статистическая оценка надежности решений методами главных компонент.
15. Задача о числе факторов. Критерий Кайзера. Критерий «каменистой осыпи».
16. Факторный анализ как метод классификации. Факторные нагрузки. Вращение факторной структуры. Методы вращения.
17. Многомерное шкалирование. Задача многомерного шкалирования. Основные подходы к многомерному шкалированию. Меры близости на основе условных вероятностей.
18. Многомерное шкалирование. Метрические меры различия профилей. Модель Торнгенсона. Примеры.
19. Неметрическое многомерное шкалирование.
20. Использование многомерного шкалирования в маркетинговых исследованиях

Задания лабораторных работ:

Задача 1. Деятельность пяти сельскохозяйственных предприятий характеризуется показателями объема реализованной продукции $x^{(1)}$ – растениеводства и $x^{(2)}$ – животноводства с одного гектара пашни (тыс.руб./га). Значения показателей представлены в таблице:

Номер хозяйства (i)	1	2	3	4	5
$x^{(1)}$	24,9	15,1	11,7	16,7	27,3
$x^{(2)}$	9,8	11,1	8,8	8,9	9,4

Требуется с помощью иерархического агломеративного алгоритма провести классификацию сельскохозяйственных предприятий и построить дендограмму при использовании обычной евклидовой метрики методом «дальнего соседа». Сделать выводы.

Задача 2. Результаты работы фермеров района оценивались по двум показателям: объем реализованной продукции растениеводства и объем реализованной продукции животноводства с гектара посевной площади (млн.руб./га), и были выделены хозяйства с высоким (А) и низким (В) уровнями использования земли. Используя данные таблицы, с помощью дискриминантного анализа провести классификацию трех последних хозяйств (Z), считая, что $\Sigma_1 \neq \Sigma_2$.

№ п/п	Группы районов	Растениеводство	Животноводство
1	Группа А X_1	25	21
2		31	37
3		27	28
4		33	36

5	Группа В X_2	52	49
6		50	61
7		47	48
8		53	46
9		55	50
10	Подлежат дискриминации (Z)	46	44
11		54	47
12		35	39

Задача 3. Потребительское поведение пяти семейств характеризуется удельными (на душу) расходами за летние месяцы на: культуру, спорт, отдых ($x^{(1)}$ – в тыс. руб.) и питание ($x^{(2)}$ – в тыс. руб.). Значения показателей представлены в следующей таблице:

Номер семьи (i)	1	2	3	4	5
$x^{(1)}$	2	4	8	12	13
$x^{(2)}$	10	7	6	11	9

Требуется с помощью агломеративного иерархического алгоритма провести классификацию семей и построить дендограмму при использовании взвешенной евклидовой метрики (с весами $w_1 = 0.05$, $w_2 = 0.95$) методом ближайшего соседа.

Задача 4. По данным опроса практиков-экономистов построена матрица ковариационной зависимости характерных признаков: x_1 – объем выпускаемой продукции, x_2 – себестоимость.

Признак	x_1	x_2
x_1	25	3
x_2	3	4

Проведите анализ данных матрицы методом главных компонент (найдите собственные числа и собственные векторы), определите относительную долю суммарной дисперсии, обусловленной одной и двумя главными компонентами.

Задача 5. По семи предприятиям имеются следующие данные о результатах работы за отчетный период:

Таблица

Номер предприятия	Выпуск продукции на одного работающего, млн. руб.	Прибыль от реализации продукции млн. руб.
1	51	28
2	63	39
3	48	29
4	39	37
5	30	18
6	58	36
7	61	55

По иерархическому агломеративному алгоритму провести классификацию $n = 7$ предприятий. В качестве расстояния между объектами принять евклидово расстояние. Расстояние между кластерами измерять по принципу «средней связи».

Задача 6. По данным опроса практиков-экономистов построена матрица корреляционной зависимости R характерных признаков: X_1 – уровень оплаты труда, X_2 – возраст, X_3 – трудовой стаж:

Признак	X_1	X_2	X_3
X_1	1	-0.388	0.665
X_2	-0.388	1	0.740
X_3	0.665	0.740	1

Проведите анализ данных матрицы парных корреляций R методом главных компонент, определите уровень информативности каждой из главных компонент и ее признаковый состав. Покажите распределение элементарных признаков в пространстве двух первых главных компонент F_1 и F_2 .

Задача 7. По данным, представленным в таблице, провести классификацию $n = 4$ предприятий по двум показателям.

Номер предприятия	1	2	3	4
$x_i^{(1)}$	3	5	5	8
$x_i^{(2)}$	6	2	7	3

Классификацию провести по иерархическому агломеративному принципу с использованием обычного евклидова расстояния, а расстояние между кластерами определять по принципу «ближайшего соседа» и центра тяжести.

Задача 8. Деятельность $n = 5$ строительных организаций характеризуется численностью рабочих ($x^{(1)}$) и фондом зарплаты ($x^{(2)}$). Значения показателей, полученных по данным годовых отчетов, представлены в следующей таблице:

Номер предприятия	1	2	3	4	5
$x^{(1)}$ (тыс. чел.)	3	6	8	2	7
$x^{(2)}$ (млн.руб.)	4	5	9	3	6

Вычислить главные компоненты и их относительный вклад (%) в суммарную вариацию. Определить нагрузки главных компонент. Ранжировать предприятия по первой главной компоненте. Представить графически результаты

Задача 9. Используя обычное евклидово расстояние между объектами и расстояние между кластерами по принципу «средней связи» групп, провести классификацию 5-ти объектов по представленной таблице.

Объекты	x_1	x_2
1	1	2
2	3	2
3	6	6
4	10	7
5	8	8

Задача 10. Эффективность использования земельных угодий двенадцатью сельскохозяйственными районами области оценивалась по объемам реализованной продукции растениеводства ($x^{(1)}$ тыс.руб./га) и животноводства ($x^{(2)}$ тыс.руб./га). Значения показателей приводятся в следующей таблице:

	Группы предприятий	Рентабельность ($x_i^{(1)}$)	Производительность труда ($x_i^{(2)}$)
1 2 3 4	Высокий уровень (X_1)	23.4 19.1 17.5 17.2	9.1 6.6 5.2 10.0
1 2 3 4 5	Низкий уровень (X_2)	5.4 6.6 8.0 9.7 9.1	4.3 5.5 5.7 5.5 6.6
1 2 3	Подлежат классификации (Z)	9.9 14.2 12.9	7.4 9.4 6.7

Предварительно известно, что в первых четырех районах земля используется неэффективно, а в следующих пяти районах – эффективно, причем случайный разброс показателей описывается внутри совокупности районов двумерным нормальным законом с неизвестными средними и а) неизвестными, но равными матрицами ковариаций.

Провести классификацию трех последних предприятий

Задача 11. Пусть $X = (X_1, X_2, X_3)$. Значения случайных величин X_1, X_2, X_3 из нормальных совокупностей с неизвестными средними и неизвестными, но равными матрицами ковариаций для обучающих выборок приведены в таблице.

Таблица .

	1-й класс			2-й класс		
	X_1 ,	X_2 ,	X_3	X_1 ,	X_2 ,	X_3
1	9.9	0.34	1.68	5.5	0.05	1.02
2	9.1	0.09	1.89	5.6	0.48	0.88
3	9.4	0.21	2.3	4.3	0.41	0.62
4	9.4	0.28	2.03	7.4	0.62	1.09
5				6.6	0.5	1.32

Классифицировать шесть объектов со значениями признаков X_1, X_2, X_3 , указанными в следующей таблице

x_1 ,	x_2 ,	x_3	x_1 ,	x_2 ,	x_3
9.4	0.15	1.91	5.2	0.74	1.82
5.5	1.20	0.68	10.0	0.32	2.62
5.7	0.66	1.43	6.7	0.39	1.24

ч

Задача 12. В следующей таблице представлены общие затраты на рубль товарной продукции ($x^{(1)}$) и фондоотдача ($x^{(2)}$) по $n = 10$ предприятиям приборостроения

Таблица

Номер предприятия	$x^{(1)}$	$x^{(2)}$
1	0.92	0.51
2	0.93	0.59
3	0.83	1.09
4	0.81	1.21
5	0.95	0.63

6	0.88	0.68
7	0.89	0.57
8	0.80	1.52
9	0.72	1.04
10	0.82	0.99

Построить оценку матрицы коэффициентов корреляции. Найти собственные числа и собственные векторы. Рассчитать факторные нагрузки. Представить графически данные на плоскости в пространстве двух главных компонент и ранжировать предприятия по первой главной компоненте.

Вопросы к зачету:

1. Дискриминантный анализ. Проблемы классификации. Принципы правильной классификации. Функции потерь и вероятности неправильной классификации. Решение задачи байесовских процедур классификации.
2. Статистическая оценка надежности решения методом главных компонент.
3. Факторный анализ. Факторные нагрузки. Простая факторная структура. Вращение факторной структуры. Методы вращения.
4. Факторный анализ – метод главных компонент. Собственные значения, задача о числе факторов. Критерий Кайзера. Критерий каменистой осыпи.
5. Дискриминантный анализ. Определение ошибки классификации.
6. Дискриминантный анализ. Классификация без обучения: расщепление смесей распределения. ЕМ-алгоритм.
7. Дискриминантный анализ. Классификация наблюдений в случае нескольких генеральных совокупностей. Классификация с помощью функции расстояния.
8. Дискриминантный анализ в случае нормальных распределений. Две генеральные совокупности с неравными ковариациями.
9. Дискриминантный анализ в случае нормальных распределений. Две генеральные совокупности с равными ковариациями.
10. Дискриминантный анализ. Принципы правильной классификации. Функции потерь и вероятности неправильной классификации. Решение задачи байесовских процедур классификации.
11. Иерархический кластерный анализ. Дендограммы. Примеры.
12. Кластерный анализ. Функционалы качества разбиений. Число кластеров.
13. Кластерный анализ. Метод средних. Метод поиска сгущений. Метод Ворда.
14. Кластерный анализ. Меры сходства между объектами. Расстояние между кластерами.
15. Применение кластерного и дискриминантного анализа.

Задания (оценочные средства), выносимые на зачет

Список вопросов для зачета приведен в пункте 3.1.

Задача 1. По данным, представленным в таблице, провести классификацию $n = 4$ семей по двум показателям.

Номер семьи	1	2	3	4
$x_i^{(1)}$	8	4	10	2

$x_i^{(2)}$	6	2	5	4
-------------	---	---	---	---

Классификацию провести по иерархическому агломеративному принципу с использованием обычного евклидова расстояния, а расстояние между кластерами определять по принципу «дальнего соседа» и средней связи.

Задача 2. Используя взвешенное евклидово расстояние между объектами весами $\omega_1 = 0.3$ и $\omega_2 = 0.7$ и расстояние между кластерами по принципу «центра тяжести» групп, провести классификацию 5-ти объектов по представленной таблице.

Объекты	x_1	x_2
1	1	1
2	1	2
3	6	3
4	8	2
5	8	0

Задача 3. Деятельность каждого производственного объединения отрасли оценивалась по следующим трем показателям:

X_1 — среднегодовой стоимости основных производственных фондов (ОПФ); X_2 — среднесписочной численности промышленно-производственного персонала (ППП); X_3 — балансовой прибыли (БП).

В отрасли выделены две группы: передовая, состоящая из четырех объединений, и остальная, включающая пять объединений. Данные представлены в таблице.

Таблица.

Показатели Группа объединений	Стоимость ОПФ	Численность ППП	Балансовая Прибыль
Передовая	224	17	23
	152	15	21
	147	14	29
	152	11	10
Остальная	47	4	11
	29	6	6
	52	4	12
	37	6	12
	64	4	13

В предположении, что (X_1, X_2, X_3) имеют нормальное распределение с одинаковыми матрицами ковариаций проверить, можно ли отнести новое объединение к передовой группе предприятий отрасли.

6. Учебно-методическое и информационное обеспечение дисциплины.

а) Основная литература

1. **Айвазян С.А., Мхитарян В.С.** Прикладная статистика. Основы эконометрики: учеб. для вузов: в 2 т.. – М.: ЮНИТИ, 2001. т.1 – 656 с (3 экз), т.2 – 432 с. (18 экз).
2. **Дубров А.М., Мхитарян В.С., Л.И. Трошин.** Многомерные статистические методы. - М., Финансы и статистика, 2000. – 352 с. (4 экз.)

б) Дополнительная литература

3. **Магнус Я.Р., Катышев П.К., Пересецкий А.А.** Эконометрика: начальный курс. - М.: Дело, 2005. – 504 с. (2 экз.).
4. **Сошникова Л.А., Тамашевич В.Н., Уеббе Г., Шефер М.** Многомерный статистический анализ в экономике. - 1999, М., Юнити, – 598 с. (3 экз).

в) Программное обеспечение и Интернет-ресурсы

архивы математических журналов на сайте mathnet.ru

7. Материально-техническое обеспечение дисциплины

- помещения для проведения занятий: лекционного типа, семинарского типа, групповых и индивидуальных консультаций, текущего контроля и промежуточной аттестации, а также помещения для хранения и профилактического обслуживания оборудования и помещения для самостоятельной работы обучающихся, оснащенные компьютерной техникой с возможностью подключения к сети "Интернет" и обеспечением доступа в электронную информационно-образовательную среду ННГУ;
- материально-техническое обеспечение, необходимое для реализации дисциплины, включая лабораторное оборудование;
- лицензионное программное обеспечение: *Windows, Microsoft Office, R-Studio*;
- обучающиеся из числа лиц с ограниченными возможностями здоровья обеспечиваются электронными и (или) печатными образовательными ресурсами в формах, адаптированных к ограничениям их здоровья.

ресурсам.

Рабочая программа учебной дисциплины составлена в соответствии с учебным планом, Положением о подготовке научных и научно-педагогических кадров в аспирантуре (адъюнктуре) (Постановление Правительства РФ от 30.11.2021 № 2122), Федеральными государственными требованиями к структуре программ подготовки научных и научно-педагогических кадров в аспирантуре (адъюнктуре) (Приказ Минобрнауки РФ от 20.10.2021 № 951).

Авторы:

Авторы: Тихов Михаил Семенович, профессор кафедры теории вероятностей и анализа данных Института ИТММ

Рецензент(ы) _____

Заведующий кафедрой _____

Программа одобрена на заседании Методической комиссии Института /факультета от _____ 2022 года, протокол № ____.