Федеральное государственное бюджетное образовательное
учреждение высшего образования
Московский государственный университет имени М.В.Ломоносова
Филиал Московского государственного университета имени М.В.Ломоносова
в городе Сарове

УТВЕРЖДАЮ
Директор филиала МГУ в городе Сарове
_____/В.В. Воеводин/

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

Наименование дисциплины:

Интеллектуальный анализ данных

Уровень высшего образования:

магистратура

Направление подготовки / специальность:

02.04.02 "Фундаментальная информатика и информационные технологии" (3++)

Направленность (профиль)/специализация ОПОП:

Суперкомпьютерные технологии и фундаментальная информатика

Форма обучения:

очная

Саров 2022

Рабочая программа дисциплины (модуля) разработана в соответствии с самостоятельно установленным МГУ образовательным стандартом (ОС МГУ) для реализуемых основных профессиональных образовательных программ высшего образования по направлению подготовки 02.04.02 "Фундаментальная информатика и информационные технологии" программы магистратуры - приказ МГУ 30 августа 2019 года № 1054 (в редакции приказа МГУ от 11 сентября 2019 года № 1109)

1. Место дисциплины (модуля) в структуре ОПОП ВО

Дисциплина относится к базовой части образовательной программы и является обязательной для освоения во 2-ом семестре обучения.

2. Входные требования для освоения дисциплины (модуля), предварительные условия (если есть)

Учащиеся должны владеть знаниями по базам данным и языкам программирования, а также по математической статистике и методам оптимизации.

3. Результаты обучения по дисциплине (модулю), соотнесенные с требуемыми компетенциями выпускников.

Формируемые компетенции	**Результаты обучения**
УК-1. Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий, формулировать научно-обоснованные гипотезы, применять методологию научного познания в профессиональной деятельности.	**Знать** историю развития прикладной математики и информатики, особенности возникновения и развития основных методов, понятий, идей, научных теорий в прикладной математики и информатике; методы критического анализа проблемных ситуаций в области прикладной математики и информатики; методологию научного познания. **Уметь** Критически анализировать проблемные ситуации в области прикладной математики и информатики на основе системного подхода **Владеть** Способен формулировать научно-обоснованные гипотезы в профессиональной области.
ОПК-1. Способен находить, формулировать и решать актуальные проблемы в области прикладной математики, фундаментальной информатики и информационно-коммуникационных технологий.	**Знать:** Актуальные проблемы современной прикладной математики и информатики; **Уметь:** анализировать источники информации для поиска новых актуальных проблем и способов их решения; **Владеть:** навыками применения передовых технологий для решения задач прикладной математики и информатики.
ПК-1. Способен в рамках задачи, поставленной специалистом более высокой квалификации, определять теоретическую основу и методологию исследования, разрабатывать план исследования в области информатики и информационно-коммуникационных технологий.	**Знать:** Компьютерные технологии, математический аппарат, вычислительные методы для проведения математического моделирования и обработки данных; типовые методики проведения исследования в области информатики и информационно-коммуникационных технологий; современные методы построения и исследования вычислительных алгоритмов для решения основных классов задач, возникающих в современной науке и технике. **Уметь:** Создавать математические модели реальных явлений и процессов; разрабатывать план исследования математических моделей реальных явлений и процессов; анализировать вычислительные алгоритмы, определять область их применимости; оценивать новизну вычислительных алгоритмов **Владеть:** Способность разрабатывать план исследования в области информатики и информационно-коммуникационных

	технологий; методами построения и исследования вычислительных алгоритмов для решения основных классов задач, возникающих в современной науке и технике.
ПК-2. Способен в рамках задачи, поставленной специалистом более высокой квалификации, проводить научные исследования и (или) осуществлять разработки в области информатики и информационно-коммуникационных технологий с получением научного и (или) научно-практического результата.	**Знать:** Принципы выбора математических моделей реальных явлений и процессов; типовые методы и алгоритмы исследования моделей реальных явлений и процессов. **Уметь:** создавать алгоритмические и математические модели типовых прикладных задач; проводить формализацию задачи, строить описательные и прогнозные модели с помощью современных программных аналитических средств, оценивать и интерпретировать полученные результаты. **Владеть:** опыт проведения научных исследований в области информатики и информационно-коммуникационных технологий с получением научного или научно-практического результата.
МПК-1 Способность понимать и применять в исследовательской и прикладной деятельности современные суперкомпьютерные технологии, математический аппарат, вычислительные методы для проведения крупномасштабного математического моделирования и обработки данных на современных высокопроизводительных вычислительных системах.	**Знать:** компьютерные технологии, математический аппарат, вычислительные методы для проведения крупномасштабного математического моделирования и обработки данных на современных высокопроизводительных вычислительных системах. **Уметь:** применять в исследовательской и прикладной деятельности современные компьютерные технологии, математический аппарат, вычислительные методы для проведения крупномасштабного математического моделирования и обработки данных на современных высокопроизводительных вычислительных системах; **Владеть:** навыками разработки программ для проведения крупномасштабного математического моделирования и обработки данных на современных высокопроизводительных вычислительных системах.

4. Формат обучения

Демонстрация примеров использования изучаемых методов и процедур проводится преподавателями на каждой лекции и каждом семинаре. Также данная дисциплина поддерживается практическими заданиями (практическими самостоятельными работами), позволяющими студентам овладеть навыками построения прогнозных и описательных моделей интеллектуального анализа данных, а также навыками анализа результатов и оценки работы реализованных моделей. Обсуждение практических самостоятельных работ, а также их защита, проводятся на семинарах. Дополнительно, на семинарах студенты выполняют небольшие практические задания по тематике последней на момент данного семинара лекции. Темы семинаров соответствуют темам лекций. Семинары направлены на укрепление знаний, полученных на лекциях.

Учебный курс состоит из двух основных блоков. Первый блок посвящен изучению алгоритмов, задач и методов, относящихся к типу обучения без учителя, предназначенных в основном для построения описательных моделей интеллектуального анализа данных, а также выявления скрытых структур и закономерностей в данных. Второй блог посвящен изучению алгоритмов, задач и методов, относящихся к типу обучения с учителем, предназначенных в основном для построения прогнозных моделей интеллектуального анализа данных. В рамках курса читаются лекции, и проводятся практические занятия,

включая выполнение самостоятельных практических заданий по разработке моделей для интеллектуального анализа данных.

5. Объем дисциплины составляет 4 зачетных единицы, всего 144 часа.
В том числе 72 академических часа, отведенных на контактную работу обучающихся с преподавателем, 72 академических часа на самостоятельную работу обучающихся.

6. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий
В курсе рассматриваются современные алгоритмы и методы интеллектуального анализа данных для решения поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования. В первой части курса, посвященной изучению методов обучения без учителя, рассматриваются: задача поиска ассоциативных правил и основные применяемые для этого алгоритмы - apriori и fp-tree; задача выявления скрытых структур в данных на основе тематического моделирования, в частности метод главных компонент, кластеризация переменных, самоорганизующиеся отображения, неотрицательная матричная факторизация; задача кластеризации данных на основе иерархических, метрических и вероятностных методов. Также обсуждаются методы предобработки данных для эффективного решения данных задач. Вторая часть курса посвящена изучению методов прогнозирования, используемых в системах интеллектуального анализа данных, связанные с этим проблемы, алгоритмы и терминология. Рассматриваются следующие вопросы: понятие проклятия размерности и проблема переобучения; вопросы и критерии для оценки и выбора моделей с использованием валидации и кросс-валидации; алгоритмы и методы необходимой предобработки данных для решения задачи прогнозирования. Далее рассматриваются наиболее популярные и современные алгоритмы и модели машинного обучения и прикладной статистики для решения задач прогнозирования в системах интеллектуального анализа данных, в частности: линейные регрессионные модели; пошаговые методы отбора переменных, регуляризация, преобразование пространства признаков для решения задач прогнозирования; нелинейные регрессионные модели, сплайны, локальная взвешенная регрессия; нейронные сети, их типовые архитектуры RBF и MLP, алгоритмы ранней остановки обучения, методы оптимизации для обучения нейронных сетей; метод опорных векторов для бинарной классификации, виды ядерных функций, алгоритмы оптимизации для обучения модели на основе опорных векторов; деревья решений, алгоритмы и критерии поиска разбиения при их построении, вопросы управление процессом роста и обрубания ветвей деревьев для борьбы с переобучением; ансамбли моделей на основе бустинга и бэгинга, случайный лес и градиентный бустинг. Демонстрация примеров использования изучаемых методов и процедур проводится преподавателями на каждой лекции и каждом семинаре. Также данная дисциплина поддерживается практическими заданиями (практическими самостоятельными работами), позволяющими студентам овладеть навыками построения прогнозных и описательных моделей интеллектуального анализа данных, а также навыками анализа результатов и оценки работы реализованных моделей. Обсуждение практических самостоятельных работ, а также их защита, проводятся на семинарах. Дополнительно, на семинарах студенты выполняют небольшие практические задания по тематике последней на момент данного семинара лекции. Темы семинаров соответствуют темам лекций. Семинары направлены на укрепление знаний, полученных на лекциях.

Наименование и краткое содержание	**Всего (час**	В том числе	
		Контактная работа (работа во взаимодействии с	**Самостоятельная работа обучающегося, часы**

разделов и тем дисциплины (модуля), **форма промежуточной аттестации по дисциплине (модулю)**	**ы)**	**преподавателем), часы** из них					из них			
		Занятия лекционного типа	Занятия семинарского типа	Групповые консультации	Индивидуальные консультации	Учебные занятия, направленные на проведение текущего контроля успеваемости коллоквиумы, практические контрольные занятия и др)*	**Всего**	Выполнение домашних заданий	Подготовка рефератов и т.п..	**Всего**
Тема 1. Введение в методы интеллектуального анализа данных	8	3	2	–	–	–	5	3	0	3
Тема 2. Выявление структур в данных. Поиск ассоциативных правил. Алгоритмы apriori и fp-tree.	8	3	2	–	–	–	5	3	0	3
Тема 3. Выявление структур в данных. Тематическое моделирование. Метод главных компонент, кластеризация переменных, самоорганизующиеся отображения.	8	3	2	–	–	–	5	3	0	3
Тема 4. Выявление структур в данных.	8	3	2	–	–	–	5	3	0	3

Кластеризация: иерархическая, метрическая, вероятностная. Предобработка данных для кластеризации.										
Тема 5. Задача прогнозирования. Проклятие размерности, переобучение, оценка и выбор моделей, валидация и кросс-валидация.	8	3	2	–	–	–	5	3	0	3
Тема 6. Задача прогнозирования. Предобработка данных для задачи прогнозирования. Метод k ближайших соседей.	8	3	2	–	–	–	5	3	0	3
Тема 7. Задача прогнозирования. Регрессионные модели. Пошаговые методы отбора переменных, регуляризация, преобразован	8	3	2	–	–	–	5	3	0	3

ие пространства признаков.										
Тема 8. Задача прогнозирования. Нелинейные регрессионные модели, сплайны, локальная взвешенная регрессия.	8	3	2	–	–	–	5	3	0	3
Тема 9. Задача прогнозирования. Нейронные сети: типовые архитектуры RBF и MLP, ранняя остановка обучения, алгоритмы оптимизации для обучения нейронных сетей.	8	3	2	–	–	–	5	3	0	3
Тема 10. Задача прогнозирования. Метод опорных векторов для бинарной классификации. Виды ядерных функций. Алгоритмы оптимизации.	10	4	3	–	–	–	7	3	0	3
Тема 11. Задача прогнозирования. Деревья решений. Алгоритмы и	12	6	3	–	–	–	9	3	0	3

критерии поиска разбиения. Управление процессом роста и обрубания ветвей деревьев.										
Тема 12. Задача прогнозирования. Ансамбли моделей. Бустинг и бэгинг ансамбли. Случайный лес.	12	6	3	–	–	–	9	3	0	3
Промежуточная аттестация – ЭКЗАМЕН	38	2					36			
Итого	144	72					72			

7. Фонд оценочных средств (ФОС) для оценивания результатов обучения по дисциплине (модулю)

7.1. Типовые контрольные задания или иные материалы для проведения текущего контроля успеваемости.

Типовые задания для ПСР и методические рекомендации к их выполнению

Домашнее задание №1 (ПСР №1)

Целью Домашнего задания №1 является освоение алгоритмов и методов прогнозирования для решения учебной задачи анализа данных в условиях, близких к реальным условиям, возникающим при решении прикладных задач анализа данных.

Формулировка задания:

Дано:

Тренировочный набор adult_train.

Файл содержит более 30 тысяч записей о людях с 14 атрибутами:

- **age:** continuous.
- **workclass**: Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
- **fnlwgt**: continuous.
- **education**: Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th,

12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.

- **education-num**: continuous.
- **marital-status**: Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
- **occupation**: Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
- **relationship**: Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
- **race**: White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
- **sex**: Female, Male.
- **capital-gain**: continuous.
- **capital-loss**: continuous.
- **hours-per-week**: continuous.
- **native-country**: United-States, Cambodia, England, Puerto-Rico, Canada, Germany, Outlying-US(Guam-USVI-etc), India, Japan, Greece, South, China, Cuba, Iran, Honduras, Philippines, Italy, Poland, Jamaica, Vietnam, Mexico, Portugal, Ireland, France, Dominican-Republic, Laos, Ecuador, Taiwan, Haiti, Columbia, Hungary, Guatemala, Nicaragua, Scotland, Thailand, Yugoslavia, El-Salvador, Trinadad&Tobago, Peru, Hong, Holand-Netherlands.

Задача заключается в том, чтобы построить модель, которая по атрибутам человека будет предсказывать богатый он или нет, т.е. атрибут **target_rich.**

Требуется:

1) Провести предобработку данных:
    a. балансировку (если нужно)
    b. фильтрацию (если нужно)
    c. разбиение на тренировочный и валидационный наборы (если нужно)
    d. подстановку пропущенных значений (если нужно)
    e. удаления корреляций (если нужно)
    f. группировку категориальных (если нужно) и дискретизацию непрерывных (если нужно)
2) Построить модель прогнозирования:
    a. Если ваша фамилия начинается на гласную, то с помощью деревьев решений и их ансамблей (если нужно)
    b. Если ваша фамилия начинается на согласную и имя на согласную, то с помощью нейросетей и их ансамблей (если нужно)
    c. Если ваша фамилия начинается на согласную, а имя на гласную, то с помощью регрессий и их ансамблей (если нужно)
3) Качество модели:
    a. Индекс Джини должен быть выше 0.8 на ТЕСТОВОМ наборе данных, который не выдается студентам (подбирайте правильно валидационный и получайте наивысший ROC индекс или индекс Джини)
    b. Top 5 лучших моделей на тестовом наборе получают автомат за экзамен по разделу Enterprise Miner.
    c. Выберите порог отсечения по максимуму значения статистики Колмогорова-Смирнова.

Домашнее задание №2 (ПСР №2)

Целью Домашнего задания №2 является освоение алгоритмов и методов прогнозирования для решения учебной задачи анализа данных в условиях, близких к реальным условиям, возникающим при решении прикладных задач анализа данных.

Формулировка задания:

Дан набор данных, в котором описана история предложений клиентам банка застраховать свои вклады. Целевая бинарная переменная INS, содержит признак, согласился ли клиент приобрести такую услугу или нет. Каждый клиент имеет свой уникальный ID. Остальные переменные – входные. Информацию о них можно посмотреть, включив опции «label» при добавлении переменных.

Определите метаданные для источника указано в таблице ниже:

	Variable Name	Role	Measurement Level	Label
1	ACCTAGE	INPUT	INTERVAL	Age of Oldest Account
2	AGE	INPUT	INTERVAL	Age
3	ATM	INPUT	BINARY	ATM
4	ATMAMT	INPUT	INTERVAL	ATM Withdrawal Amount
5	BRANCH	INPUT	NOMINAL	Branch of Bank
6	CASHBK	INPUT	INTERVAL	Number Cash Back
7	CC	INPUT	BINARY	Credit Card
8	CCBAL	INPUT	INTERVAL	Credit Card Balance
9	CCPURC	INPUT	INTERVAL	Credit Card Purchases
10	CD	INPUT	BINARY	Certificate of Deposit
11	CDBAL	INPUT	INTERVAL	CD Balance
12	CHECKS	INPUT	INTERVAL	Number of Checks
13	CRSCORE	INPUT	INTERVAL	Credit Score
14	DDA	INPUT	BINARY	Checking Account
15	DDABAL	INPUT	INTERVAL	Checking Balance
16	DEP	INPUT	INTERVAL	Checking Deposits
17	DEPAMT	INPUT	INTERVAL	Amount Deposited
18	DIRDEP	INPUT	BINARY	Direct Deposit
19	HMOWN	INPUT	BINARY	Owns Home
20	HMVAL	INPUT	INTERVAL	Home Value
21	id	ID	NOMINAL	
22	ILS	INPUT	BINARY	Installment Loan
23	ILSBAL	INPUT	INTERVAL	Loan Balance
24	INAREA	INPUT	BINARY	Local Address
25	INCOME	INPUT	INTERVAL	Income
26	INS	TARGET	BINARY	Insurance Product
27	INV	INPUT	BINARY	Investment
28	INVBAL	INPUT	INTERVAL	Investment Balance

29	IRA	INPUT	BINARY	Retirement Account
30	IRABAL	INPUT	INTERVAL	IRA Balance
31	LOC	INPUT	BINARY	Line of Credit
32	LOCBAL	INPUT	INTERVAL	Line of Credit Balance
33	LORES	INPUT	INTERVAL	Length of Residence
34	MM	INPUT	BINARY	Money Market
35	MMBAL	INPUT	INTERVAL	Money Market Balance
36	MMCRED	INPUT	INTERVAL	Money Market Credits
37	MOVED	INPUT	BINARY	Recent Address Change
38	MTG	INPUT	BINARY	Mortgage
39	MTGBAL	INPUT	INTERVAL	Mortgage Balance
40	NSF	INPUT	BINARY	Number Insufficient Fund
41	NSFAMT	INPUT	INTERVAL	Amount NSF
42	PHONE	INPUT	NOMINAL	Number Telephone Banking
43	POS	INPUT	INTERVAL	Number Point of Sale
44	POSAMT	INPUT	INTERVAL	Amount Point of Sale
45	RES	INPUT	NOMINAL	Area Classification
46	SAV	INPUT	BINARY	Saving Account
47	SAVBAL	INPUT	INTERVAL	Saving Balance
48	SDB	INPUT	BINARY	Safety Deposit Box
49	TELLER	INPUT	INTERVAL	Teller Visits
50	_dataobs_	REJECTED	INTERVAL	Observation Number

Необходимо построить модель прогнозирования для бинарного отклика, которая будет наилучшим образом его предсказывать. Можно использовать любые методы, рассмотренные на лекциях.

Оцениваться качество модели будет:

1) По критерию ROC Index (площадь под ROC кривой).

2) На тестовом наборе, где реальный отклик не будет известен аналитику (вам), но будет известен проверяющему (мне).

3) Будет проверяться не только результат для тестового набора, но и что предоставленная модель действительно генерирует представленный тестовый набор.

4) Для зачета по заданию необходимо получить ROC на тестовом наборе больший или равный 0.777

(заметьте, что оценки на тестовом и валидационном наборах могут сильно отличаться).

Данные ПСР соответствуют изучаемым темам следующим образом:

№	Изучаемая тема	Соответствующая ПСР
1	Темы 1-12	ПСР №1
2	Темы 1-12	ПСР №2

7.2. Типовые контрольные задания или иные материалы для проведения промежуточной аттестации.

Список вопросов для устной части экзамена

1. Понятие процесса интеллектуального анализа данных, основные типы решаемых задач, исходных данных и приложений.
2. Поиск ассоциативных правил. Алгоритмы apriori и fp-tree.
3. Тематическое моделирование. Метод главных компонент, кластеризация переменных, самоорганизующиеся отображения.
4. Кластеризация: иерархическая, метрическая, вероятностная. Предобработка данных для кластеризации.
5. Задача прогнозирования. Проклятие размерности, переобучение, оценка и выбор моделей, валидация и кросс-валидация.
6. Предобработка данных для задачи прогнозирования. Метод k ближайших соседей.
7. Регрессионные модели. Пошаговые методы отбора переменных, регуляризация, преобразование пространства признаков.
8. Нелинейные регрессионные модели, сплайны, локальная взвешенная регрессия.
9. Нейронные сети: типовые архитектуры RBF и MLP, ранняя остановка обучения, алгоритмы оптимизации для обучения нейронных сетей.
10. Метод опорных векторов для бинарной классификации. Виды ядерных функций. Алгоритмы оптимизации.
11. Деревья решений. Алгоритмы и критерии поиска разбиения. Управление процессом роста и обрубания ветвей деревьев.
12. Ансамбли моделей. Бустинг и бэгинг ансамбли. Случайный лес. Процедуры и инструменты для поиска выбросов.

Примеры вопросов для письменной части экзамена

Письменная часть экзамена охватывает материал всего курса и состоит из заданий следующего типа:

1. Тестовые вопросы с выбором одного варианта ответа из списка предложенных;
2. Тестовые вопросы на выбор нескольких верных утверждений из списка предложенных;
3. Расчетные задачи без выбора вариантов ответа;
4. Задачи на написание программы.

Регрессионные модели:

Ниже даны утверждения относительно некоторых процедур построения регрессионных моделей. Какие из перечисленных ниже утверждений истины относительно алгоритма LARS, использующего метод LASSO, а какие относительно регрессии частичных квадратов PLS. Некоторые утверждения справедливы для обоих алгоритмов, некоторые ни для одного.

a. Может применяться для выбора важных переменных
b. Может применяться для прогнозирования бинарного отклика
c. Может применяться для прогнозирования категориального отклика
d. Использует линейное преобразование пространства признаков
e. Использует нелинейное преобразование пространства признаков
f. Использует регуляризацию L2 в пространстве коэффициентов регрессионной модели
g. Использует регуляризацию L1 в пространстве коэффициентов регрессионной модели
h. Использует пошаговые методы выбора важных переменных
i. Может обрабатывать пропущенные значения
j. Корректно оценивает важность категориальных предикторов
k. Может строить нелинейные модели
l. Может строить обобщенные линейные модели
m. Максимально сложная модель, построенная таким методом, всегда совпадает с регрессией, построенной методом наименьших квадратов на всем наборе входных переменных.

LARS+LASSO:

_____

PLS: _____

Поиск ассоциативных правил:

Дан набор из 50 транзакций (чеки из магазина): {bourbon},{baguette},{artichok avocado},{bourbon},{avocado apples},{apples bourbon},{baguette bourbon},{baguette bourbon},{bordeaux bourbon},{avocado apples baguette},{apples artichok avocado baguette},{apples},{avocado apples baguette},{apples},{artichok bourbon},{artichok avocado baguette},{apples},{baguette bourbon},{bourbon},{baguette apples},{bourbon artichok avocado baguette},{artichok bourbon},{bourbon},{baguette apples},{baguette apples bourbon},{avocado bourbon},{artichok bourbon},{bourbon},{bourbon},{bourbon},{avocado apples baguette},{bordeaux bourbon},{bordeaux artichok avocado baguette},{baguette},{artichok avocado},{apples bourbon},{artichok avocado baguette},{artichok},{artichok bourbon},{baguette apples},{bourbon},{bourbon},{bourbon},{avocado apples baguette},{artichok baguette},{bourbon},{artichok avocado}
Найдите методом FP-tree (или априори) частые наборы и достоверные правила при заданных ограничениях: Min support = 10%. Min confidence = 60%. Распишите процедуру поиска частых эпизодов и правил по шагам с учетом выбранного алгоритма. У какого правила самый высокий lift? Дайте словесную интерпретацию этому правилу и всем его числовым характеристикам.

_____ _____ФИО _____ группа

Кластерный анализ:

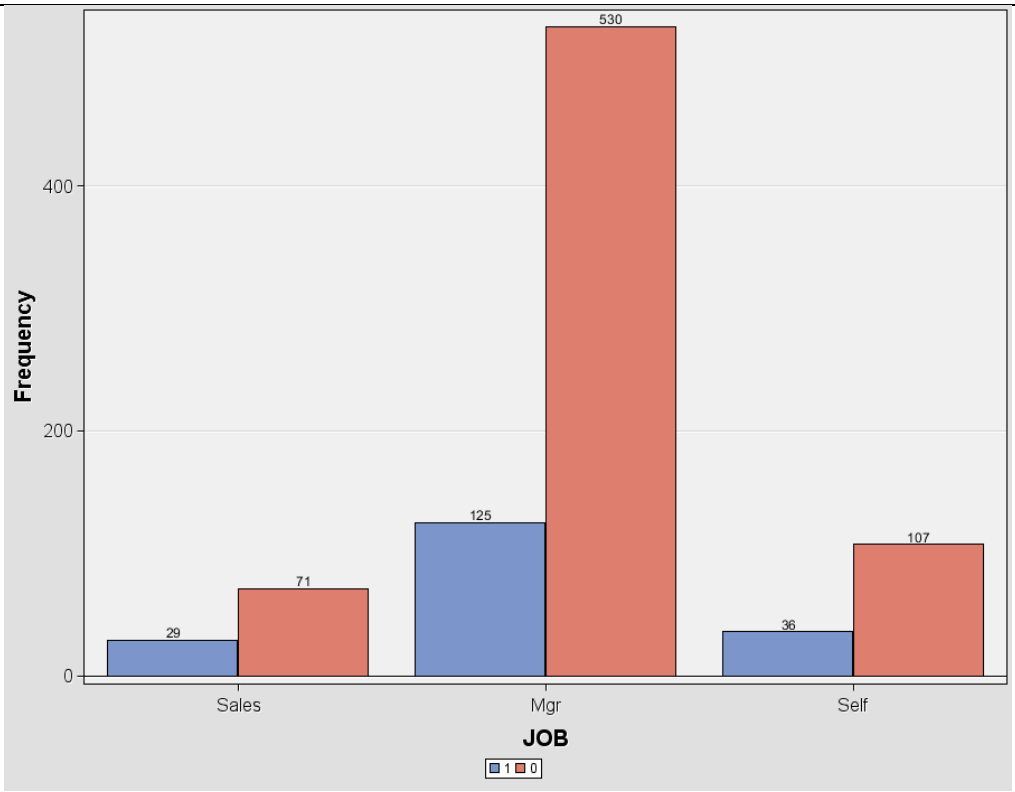
Дано множество точек:

A	B	C	D	E	F	G	H	I	J
(2,10)	(2,5)	(8,4)	(5,8)	(7,5)	(6,4)	(1,2)	(4,9)	(7,6)	(1,0)

Проведите иерархическую кластеризацию по методу Single-link, выпишите последовательно полученные кластеры после каждого шага или нарисуйте дендрограмму. Везде указывать межкластерное расстояние для каждого шага. Для удобства рекомендуется сначала выписать матрицу попарных квадратов расстояний.

Деревья решений:

Дана информация о распределении значений переменной JOB (с тремя значениями) в срезе отклика BAD:



Среди всех возможных бинарных вариантов разбиения по переменной JOB найдите лучший по критерию Xi2. Выпишите таблицы сопряженности для каждого варианта (т.к. число степеней свободы одинаковое, то p-value можно не рассчитывать). Сравните по Джини (или энтропии) победивший по Xi2 вариант с вариантом разбиения, когда для каждого значения атрибута JOB будет выделена своя ветвь.

_____ ____ФИО _____ группа

Предобработка данных:

Ниже даны утверждения относительно некоторых процедур предобработки данных. Укажите, какие из них справедливы для процедуры семплирования, а какие для процедуры фильтрации выбросов? Некоторые пункты справедливы для обоих процедур, некоторые ни для одной.

a. Может добавлять новые переменные в обрабатываемый набор данных
b. Может удалять переменные в обрабатываемом наборе данных
c. Может добавлять новые записи в обрабатываемый набор данных
d. Может удалять записи в обрабатываемом наборе данных
e. Может использоваться для балансировки классов
f. Может находить и удалять артефакты и выбросы
g. Может находить и исправлять артефакты и выбросы
h. Может использовать информацию о распределении категориальной переменной при формировании выходных наборов данных
i. Может использовать информацию о распределении числовой переменной при формировании выходных наборов данных
j. Производит кластеризацию для последующей обработки с сохранением пропорций размеров кластеров
k. Определяет важность входных переменных для последующей обработки с фильтрацией незначимых предикторов
l. Позволяет объединять несколько наборов данных в один
m. Позволяет разбивать один набор данных на несколько

Sample: _____

Filter: _____

Методические материалы для проведения процедур оценивания результатов обучения

В течение семестра студенты выполняют небольшие практические задания на семинарах (по тематике последней на момент данного семинара лекции), а также две ПСР дома (которые обсуждаются с преподавателями на семинарах и «защищаются»).
За работу на семинарах студенты могут получить 0–40 баллов.
За каждую ПСР студенты могут получить 0–30 баллов (таким образом, всего за ПСР можно получить 0–60 баллов).
Таким образом, за семестр студенты могут набрать 0–100 баллов.
По результатам работы в семестре, всем студентам ставится предварительная оценка по следующей схеме:

Количество баллов, набранных в семестре	**Предварительная оценка**
Не менее 80 баллов	«ОТЛ»
Не менее 60 баллов и не более 79 баллов	«ХОР»
Не менее 40 баллов и не более 59 баллов	«УДОВЛ»
Не более 39 баллов	«НЕУД»

Далее, на экзамене студенты пишут письменную работу, за которую также получают оценку (вся работа оценивается в 100 баллов, оценка за письменную работу ставится аналогично оценке за работу в семестре).
Итоговая оценка за дисциплину вычисляется как среднее арифметическое между оценкой за работу в семестре и оценкой за письменную на экзамене работу. В случае возникновения спорной ситуации, преподаватели устно задают студенту любые три вопроса из списка вопросов для устной части экзамена. По результатам ответа студента на вопросы, ставится итоговая оценка.

РЕЗУЛЬТАТ ОБУЧЕНИЯ по дисциплине (модулю)	**КРИТЕРИИ и ПОКАЗАТЕЛИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТА ОБУЧЕНИЯ по дисциплине (модулю)**					**ОЦЕНОЧНЫЕ СРЕДСТВА**
	1	2	3	4	5	
Знать: основные типовые и прикладные задачи, которые решаются с помощью методов интеллектуального анализа данных, условия применимости и ограничения этих методов, подходы для интерпретации полученных результатов, типовые методы предобработки и данных.	Отсутствие знаний	Фрагментарные представления	В целом сформированные, но неполные знания	Сформированные, но содержащие отдельные пробелы знания	Сформированные систематические знания	Устная часть экзамена; Письменная часть экзамена; Обсуждение ПСР с преподавателями.
Уметь: создавать алгоритмические и математические модели типовых прикладных задач, решаемых с помощью методов	Отсутствие умений	Фрагментарные умения	В целом успешное, но не систе	Успешное, но содержащее отдельные пробелы	Сформированное умение	Письменная часть экзамена; ПСР.

интеллектуального анализа данных, проводить формализацию задачи, строить описательные и прогнозные модели с помощью современных программных аналитических средств, оценивать и интерпретировать полученные результаты, реализовывать алгоритмы предобработки и постобработки данных.			матическое умение	умение		
Владеть: Навыками построения описательных и прогнозных аналитических моделей с использованием современных инструментов интеллектуального анализа данных.	Отсутствие навыков	Фрагментарное владение навыками	В целом успешное, но не полное владение навыками	Успешное, но содержащее отдельные пробелы владение навыками	Сформированное владение навыками	Устная часть экзамена; Письменная часть экзамена; ПСР.
Знать: основные методы машинного обучения и прикладной статистики для решения задач поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования;	Отсутствие знаний	Фрагментарные представления	В целом сформированные, но неполные знания	Сформированные, но содержащие отдельные пробелы знания	Сформированные систематические знания	Устная часть экзамена; Письменная часть экзамена; Обсуждение ПСР с преподавателями.
Уметь: формализовать задачу интеллектуального анализа данных в части методов поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования; настраивать, оценивать, сравнивать и выбирать модели интеллектуального анализа данных.	Отсутствие умений	Фрагментарные умения	В целом успешное, но не систематическое умение	Успешное, но содержащее отдельные пробелы умение	Сформированное умение	Письменная часть экзамена; ПСР.
Владеть: теоретическими знаниями в области прикладной статистики и машинного обучения; практическими навыками работы с современными	Отсутствие навыков	Фрагментарное владение навыками	В целом успешное, но не полное	Успешное, но содержащее отдельные пробелы владение	Сформированное владение навыками	Устная часть экзамена; Письменная часть экзамена; ПСР.

программными системами интеллектуального анализа.			владение навыками	навыками		
Знать: Понятие процесса интеллектуального анализа данных, последовательности типовых шагов этого процесса и задач, решаемых на каждом шаге. основные алгоритмы и методы обработки, и анализа данных, используемые в системах интеллектуального анализа данных.	Отсутствие знаний	Фрагментарные представления	В целом сформированные, но неполные знания	Сформированные, но содержащие отдельные пробелы знания	Сформированные систематические знания	Устная часть экзамена; Письменная часть экзамена; Обсуждение ПСР с преподавателями.
Уметь: использовать современные программные аналитические средства для разработки, оценки и применения прогнозных и описательных моделей интеллектуального анализа данных.	Отсутствие умений	Фрагментарные умения	В целом успешное, но не систематическое умение	Успешное, но содержащее отдельные пробелы умение	Сформированное умение	Письменная часть экзамена; ПСР.
Владеть: навыками решения практических задач, связанных с анализом данных с использованием современных аналитических программных средств.	Отсутствие навыков	Фрагментарное владение навыками	В целом успешное, но не полное владение навыками	Успешное, но содержащее отдельные пробелы владение навыками	Сформированное владение навыками	Устная часть экзамена; Письменная часть экзамена; ПСР.
Знать: особенности методов обработки и анализа данных; факторы, влияющие на эффективность работы методов анализа данных.	Отсутствие знаний	Фрагментарные представления	В целом сформированные, но неполные знания	Сформированные, но содержащие отдельные пробелы знания	Сформированные систематические знания	Устная часть экзамена; Письменная часть экзамена; Обсуждение ПСР с преподавателями.
Уметь: разрабатывать и оценивать модели интеллектуального анализа данных для решения прикладных задач.	Отсутствие умений	Фрагментарные умения	В целом успешное, но не систематическое умени	Успешное, но содержащее отдельные пробелы умение	Сформированное умение	Письменная часть экзамена; ПСР.

			е			
Владеть: Аналитическими программными средствами интеллектуального анализа данных.	Отсутствие навыков	Фрагментарное владение навыками	В целом успешное, но не полное владение навыками	Успешное, но содержащее отдельные пробелы владение навыками	Сформированное владение навыками	Устная часть экзамена; Письменная часть экзамена; ПСР.

8. Ресурсное обеспечение:

Основная литература

1. Айвазян С.А., Бухтштабер В.М., Енюков И.С., Мешалкин Л.Д. /Прикладная статистика: Классификации и снижение размерности. Справочное издание. - М.: Финансы и статистика, 1989. - 607с.
2. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. – Springer, 2001.

Активные и интерактивные формы проведения занятия

Каждое занятие (лекция и семинар) сопровождается демонстрацией преподавателями изучаемых на данном занятии технологий. В рамках данных демонстраций студенты проделывают необходимые действия по настройке программного обеспечения, написанию и запуску программ на своих компьютерах, задают вопросы. Дополнительно, на семинаре студенты выполняют небольшие практические задания (как индивидуально, так и в группах) по тематике последней на момент данного семинара лекции. Также на каждом занятии проводится обсуждение домашних заданий, а также все студенты имеют возможность задать преподавателям свои вопросы по изучаемой теме.

Лицензионное программное обеспечение, в том числе отечественного производства

При реализации дисциплины может быть использовано следующее программное обеспечение:

1. Программный продукт Red Hat Enterprise Linux Server for HPC Compute Node for Power, LE, Self-support 4 шт. №5540331
2. Программный продукт Red Hat Enterprise Linux Server for HPC Head Node for Power, LE, Standard 1 шт. №5540332
3. Операционная система SUSE Linux Enterprise Server 11 SP4 for x86_64 16 шт.
4. Операционная система Red Hat Enterprise Linux Server 5.0 for x86_64 14шт.
5. Операционная система SUSE Linux Enterprise Server 10 SP3 for ppc64 7 шт.
6. Операционная система Ubuntu 18.04.
7. Программное обеспечение для виртуализации Oracle VM VirtualBox
8. Операционная система ALTLinuxMATEStarterkit 9 лицензияGPL
9. Программный продукт JetBrains IntelliJ IDEA Community Edition Free Educational Licenses
10. Программный продукт JetBrainsPyCharm Community Edition Free Educational Licenses

11. Программный продукт JetBrainsCLion Community Edition Free Educational Licenses
12. Программный продукт UPPAAL (http://www.uppaal.org/) академическая лицензия
13. Программный продукт Java 8 (64-bit)Oracle Corporation
14. Программный продукт Java SE Development Kit 8(64-bit) Oracle Corporation
15. Программный продукт NetBeans IDE 8.2 NetBeans.org
16. Программный продукт Dev-C++ Bloodshed Software
17. Программный продуктCodeBlocksThe Code::Blocks Team
18. Программный продукт Free Pascal 3.0.0Free Pascal Team
19. Программный продукт Python 3.5.1 (64-bit)Python Software Foundation
20. Программный продукт R for Windows 3.2.2 R Core Team
21. Программный продуктHaskell Platform 7.10.3 Haskell.org
22. Операционная система Microsoft Windows 7 корпоративная академическая лицензия
23. Операционная система Microsoft Windows 10 Education академическая лицензия
24. Программный продукт Microsoft ProjectProfessional 2013 академическая лицензия
25. Программный продукт Microsoft VisioProfessional 2013 академическая лицензия
26. Программный продуктMicrosoft VisualStudioProfessional 2013 - RUS [Русский (Россия)] академическая лицензия

Профессиональные базы данных и информационные справочные системы

1. http://www.edu.ru – портал Министерства образования и науки РФ
2. http://www.ict.edu.ru – система федеральных образовательных порталов «ИКТ в образовании»
3. http://www.openet.ru - Российский портал открытого образования
4. http://www.mon.gov.ru - Министерство образования и науки Российской Федерации
5. http://www.fasi.gov.ru - Федеральное агентство по науке и инновациям

Материально-техническая база

Для преподавания дисциплины требуется класс, оборудованный маркерной или меловой доской и проектором (и компьютером с разъемом VGA / HDMI для подключения к проектору);

9. Язык преподавания.
Русский.

11. Автор (авторы) программы.
1. Доцент кафедры Интеллектуальных Информационных Технологий, Петровский Михаил Игоревич (michael@cs.msu.su)