МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В.Ломоносова»

«Утверждаю»

Декан факультета ВМК МГУ
имени М.В. Ломоносова

академик _____ Е.И.Моисеев

«___»_____ 2017 г.

РАБОЧАЯ ПРОГРАММА УЧЕБНОЙ ДИСЦИПЛИНЫ

«Аналитическое программное обеспечение SAS»

Уровень высшего образования – подготовка научно-педагогических кадров в аспирантуре

Направление подготовки – 09.06.01 «Информатика и вычислительная техника»

Направленность (профиль) – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (05.13.11)

2017 г.

Рабочая программа дисциплины (модуля)

1. Наименование дисциплины
АНАЛИТИЧЕСКОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ SAS

2. Уровень высшего образования
Подготовка научно-педагогических кадров в аспирантуре.

3. Направление подготовки, направленность (профиль) подготовки
Направление подготовки: 09.06.01 «Информатика и вычислительная техника»;
Направленность (профиль):«Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей» (05.13.11).

4. Место дисциплины в структуре основной образовательной программы
Дисциплина относится к специальным курсам по выбору вариативной части образовательной программы.

5. Перечень планируемых результатов обучения
Дисциплина участвует в формировании следующих компетенций образовательной программы:

Формируемые компетенции *(код компетенции)*	**Планируемые результаты обучения по дисциплине (модулю)**
Владение методологией теоретических и экспериментальных исследований в области профессиональной деятельности (ОПК-1)	ЗНАТЬ: классические математические методы, применяющиеся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий; УМЕТЬ: применять классические методы построения и анализа математических моделей; ВЛАДЕТЬ: базовыми навыками выбора методов и средств

	построения и анализа математических моделей
Владение современными методами построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также методами разработки и реализации алгоритмов их решения на основе фундаментальных знаний в области математики и информатики (ПК-1)	ЗНАТЬ: классические методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также базовые методы разработки и реализации алгоритмов их решения; УМЕТЬ: применять классические методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также базовые методы разработки и реализации алгоритмов их решения; ВЛАДЕТЬ: базовыми навыками выбора методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также методов разработки и реализации алгоритмов их решения.
Владение современными алгоритмами разработки программного обеспечения вычислительных комплексов (ПК-3)	ЗНАТЬ: современные алгоритмы разработки программного обеспечения вычислительных комплексов; УМЕТЬ: применять современные алгоритмы разработки программного обеспечения вычислительных комплексов; ВЛАДЕТЬ: базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов.
Владение современными методами интеллектуального анализа данных (ПК-5)	ЗНАТЬ: современные методы интеллектуального анализа данных; УМЕТЬ: применять современные методы интеллектуального анализа данных; ВЛАДЕТЬ: базовыми навыками выбора методов интеллектуального анализа данных.

Оценочные средства для промежуточной аттестации приведены в Приложении.

6. Объем дисциплины

Объем дисциплины составляет **3**зачетных единицы, всего **108** часов.

72 часа составляет контактная работа с преподавателем (**40** часов занятий лекционного типа, **30** часов занятий семинарского типа (семинары, научно-практические занятия, лабораторные работы и т.п.), **0** часов групповых консультаций, **0** часов индивидуальных консультаций, **0**часов мероприятий текущего контроля успеваемости, **2** часа мероприятий промежуточной аттестации).

36 часов составляетсамостоятельная работа учащихся.

7. Входные требования для освоения дисциплины

Учащиеся должны владеть знаниями по базам данным и языкам программирования, а также по математической статистикеи методам оптимизации в объеме, соответствующем основным образовательным программам бакалавриата и магистратуры по укрупненным группам направлений и специальностей 01.00.00 «Математика и механика», 02.00.00 «Компьютерные и информационные науки».

8. Образовательные технологии

В процессе обучения используется бесплатная облачная аналитическая платформа интеллектуального анализа данных SASEnterpriseMinerOnDemandforAcademics, включающая интерактивные средства разработки аналитических моделей для решения задач поиска ассоциативных правил, тематического моделирования, кластеризации, классификации и прогнозирования, а также библиотеку современных методов прикладной статистики и машинного обучения для решения данных задач. 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.

9. Содержание дисциплины

В курсе рассматриваются современные алгоритмы иметоды 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 частностиметод 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 аспирантам 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 аспиранты 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 разделов и тем дисциплины (модуля), **форма промежуточной аттестации по дисциплине (модулю)**	**Всего (часы)**	В том числе								
		Контактная работа (работа во взаимодействии с преподавателем), часы из них						**Самостоятельная работа обучающегося, часы** 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 рефератови т.п..	**Всего**
Тема 1. Введение в методы интеллектуального анализа данных	4	2	2	–	–	–	4	0	0	0
Тема 2. Выявление структур в данных. Поиск ассоциативных правил. Алгоритмы apriori и fp-tree.	8	3	3	–	–	–	6	2	0	2
Тема 3. Выявление структур в данных. Тематическое моделирование. Метод главных компонент, кластеризация переменных, самоорганизующиеся отображения.	8	3	3	–	–	–	6	2	0	2

Тема 4. Выявление структур в данных. Кластеризация: иерархическая, метрическая, вероятностная. Предобработка данных для кластеризации.	8	3	3	–	–	–	6	2	0	2
Тема 5. Задача прогнозирования. Проклятие размерности, переобучение, оценка и выбор моделей, валидация и кросс-валидация.	6	3	3	–	–	–	6	0	0	0
Тема 6. Задача прогнозирования. Предобработка данных для задачи прогнозирования. Метод kближайших соседей.	8	3	3	–	–	–	6	2	0	2
Тема 7. Задача прогнозирования. Регрессионные модели.Пошаговые методы отбора переменных, регуляризация, преобразование пространства признаков.	8	3	3	–	–	–	6	2	0	2
Тема 8. Задача прогнозирования. Нелинейные регрессионные	6	3	3	–	–	–	6	0	0	0

модели, сплайны, локальная взвешенная регрессия.										
Тема 9. Задача прогнозирования. Нейронные сети: типовые архитектуры RBF и MLP, ранняя остановка обучения, алгоритмы оптимизации для обучения нейронных сетей.	8	3	3	–	–	–	6	2	0	2
Тема 10. Задача прогнозирования. Метод опорных векторов для бинарной классификации. Виды ядерных функций. Алгоритмы оптимизации.	8	3	3	–	–	–	6	2	0	2
Тема 11. Задача прогнозирования. Деревья решений. Алгоритмы и критерии поиска разбиения. Управление процессом роста и обрубания ветвей деревьев.	8	3	3	–	–	–	6	2	0	2
Тема 12. Задача прогнозирования. Ансамбли моделей. Бустинг и бэгинг ансамбли. Случайный лес.	8	3	3	–	–	–	6	2	0	2
Промежуточная	20	2						18		

аттестация– ЭКЗАМЕН			
Итого	108	72	36

10. Учебно-методические материалы для самостоятельной работы учащихся

Самостоятельная работа учащихся проводится в виде выполнения практических самостоятельных работ (ПСР).

Текущий контроль осуществляется путем проверки ПСР преподавателями, а также индивидуального обсуждения (защиты) с преподавателями выполненной ПСР на семинарах. За каждую ПСР аспиранту проставляется определенное количество баллов. Итоговая сумма баллов влияет на итоговую оценку учащегосяпо данной дисциплине.

Также семинарахаспиранты выполняют небольшие практические задания по тематике последней на момент данного семинара лекции. Качество выполнения заданий также влияет на итоговую оценку.

Методика выставления оценки по данной дисциплине (с учетом ПСР, а также практических заданий на семинарах) приведена в разделе«Методические материалы для проведения процедур оценивания результатов обучения» Приложения.

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.sas7bat лежит в директории "/courses/u_cmc.msu.ru1/i_889205/c_6043".

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) Создать библиотеку, привязав ее к "/courses/u_cmc.msu.ru1/i_889205/c_6043"
2) Подключить источник данных и правильно задать типы и роли переменных (target_rich – отклик, ID - нет)
3) 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)

4) 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)
5) Качество модели:
   a. Индекс Джини должен быть выше 0.8 на ТЕСТОВОМ наборе данных, который не выдается аспирантам (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 EnterpriseMiner.
   c. Выберите порог отсечения по максимуму значения статистики Колмогорова-Смирнова

Результат:

Запишите диаграмму в формате xml (нажав на диаграмму в дереве проекта и выбрав Save as) и отправьте на проверку преподавателям.

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, которые реализованы в SAS Enterprise Miner.

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 зачет по заданию необходимо получить ROC на тестовом наборе больший или равный 0.777 (заметьте, что оценки на тестовом и валидационном наборах могут сильно отличаться).

Входные данные:

1) Закачены на сервер и состоят из тренировочного набора train и тестового набора test.

2) Чтобы получить к ним доступ нужно в вашем проекте добавить библиотеку (с любым именем), привязав ее к директории. Для этого в вашем проекте выбираете File->New->Library и в запушенном визарде создаете библиотеку (например, с именем Task4) и путем /courses/d2db66e5ba27fe300

3) После этого можно добавлять в проект наборы файлов также как из обычных библиотек.

В качестве базового примера предоставлена диаграмма sample, которая реализует простую пошаговую регрессию и дерево решений.

Среди этих моделей выбирается лучшая (в узле Model Comparison), она применяется к тестовому набору данных (в узле Score).

Для сдачи этого задания необходимо прислать:

1) Проэкспортированную диаграмму в формате xml (как обычно)

2) Скопированный код из результатов узла Score из раздела “SAS Code”

3) ROC Index должен быть больше 0.777 (у модели в примере он меньше 0.75).

Данные ПСР соответствуют изучаемым темам следующим образом:

№	Изучаемая тема	Соответствующая ПСР
1	Темы1-12	ПСР №1
2	Темы 1-12	ПСР №2

11. 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.
3. Kattamuri S. Sarma Predictive Modeling with SAS® Enterprise Miner™: Practical Solutions for Business Applications, Third Edition - SAS Institute, 2017

Ресурсы информационно-телекоммуникационной сети «Интернет»

1. SASEnterpriseMinerOnDemandforAcademics, руководство по установке и использованию. https://www.sas.com/ru_ru/software/on-demand-for-academics.html

Информационные технологии, используемые в процессе обучения

1. БесплатнаяоблачнаяаналитическаяплатформаSASEnterprise Miner OnDemand for Academics

Активные и интерактивные формы проведения занятия

Каждое занятие (лекция и семинар) сопровождается демонстрацией преподавателями изучаемых на данном занятии технологий. В рамках данных демонстраций аспиранты проделывают необходимые действия по настройке программного обеспечения, написанию и запуску программ на своих компьютерах, задают вопросы. Дополнительно, на семинаре аспиранты 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 аспиранты имеют возможность задать преподавателям свои вопросы по изучаемой теме.

Материально-техническая база

1. Для преподавания дисциплины требуется класс, оборудованный маркерной или меловой доской и проектором (и компьютером с разъемом VGA / HDMI для подключения к проектору);
2. Для демонстрации современных технологий аналитики больших данных требуется компьютерный класс с доступом в Интернет со следующим установленным графическим клиентом SASEnterpriseMinerOnDemandforAcademics.

12. Язык преподавания
Русский.

13. Разработчики программы, Преподаватели
Доцент кафедры Интеллектуальных Информационных Технологий, Петровский Михаил Игоревич (michael@cs.msu.su)

Приложение

Оценочные средства для промежуточной аттестации по дисциплине «АНАЛИТИЧЕСКОЕ ПРОГРАММНОЕ ОБЕСПЕЧЕНИЕ SAS»

Средства для оценивания планируемых результатов обучения, критерии и показатели оценивания приведены ниже.

РЕЗУЛЬТАТ ОБУЧЕНИЯ по дисциплине (модулю)	**КРИТЕРИИ и ПОКАЗАТЕЛИ ОЦЕНИВАНИЯ РЕЗУЛЬТАТА ОБУЧЕНИЯ по дисциплине (модулю)**					**ОЦЕНОЧНЫЕ СРЕДСТВА**
	1	2	3	4	5	
ЗНАТЬ: современные математические методы, применяющиеся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий **Код З1 (ОПК-1)**	Отсутствие знаний	Фрагментарные представления о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	В целом сформированные, но неполные знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Сформированные, но содержащие отдельные пробелы знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Сформированные систематические знания о современных математических методах, применяющихся для решения задач в области естественных наук, экономики, социологии и информационно-коммуникационных технологий	Устный экзамен
УМЕТЬ: применять современные методы постановки и анализа задач в области математики и информатики	Отсутствие умений	Фрагментарные умения применять современные методы	В целом успешное, но не систематическое умение применять современные	Успешное, но содержащее отдельные пробелы умение применять	Сформированное умение применять современные методы постановки и	отчет

Код У1 (ОПК-1)		постановки и анализа задач в области математики и информатики	методы постановки и анализа задач в области математики и информатики	современные методы постановки и анализа задач в области математики и информатики	анализа задач в области математики и информатики	
ВЛАДЕТЬ: навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики **Код В1 (ОПК-1)**	Отсутствие навыков	Фрагментарное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	В целом успешное, но не полное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	Успешное, но содержащее отдельные пробелывладение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	Сформированное владение навыками оптимального выбора современных методов и средств постановки и анализа задач в области математики и информатики	отчет
ЗНАТЬ: современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения **Код З1 (ПК-1)**	Отсутствие знаний	Фрагментарные представления о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных	В целом сформированные, но неполные знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также	Сформированные, но содержащие отдельные пробелы знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также	Сформированные систематические знания о современных методах построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных	Устный экзамен

		методах разработки и реализации алгоритмов их решения	современных методах разработки и реализации алгоритмов их решения	современных методах разработки и реализации алгоритмов их решения	методах разработки и реализации алгоритмов их решения	
УМЕТЬ: применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения **Код У1 (ПК-1)**	Отсутствие умений	Фрагментарные умения применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	В целом успешное, но не систематическое умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	Успешное, но содержащее отдельные пробелы умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	Сформированное умение применять современные методы построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современныеметоды разработки и реализации алгоритмов их решения	отчет
ВЛАДЕТЬ: навыками оптимального выбора современных методов построения и анализа математических моделей, возникающих при решении	Отсутствие навыков	Фрагментарное владение навыками оптимального выбора современных	В целом успешное, но не полное владение навыками оптимального выбора	Успешное, но содержащее отдельные пробелывладение навыками оптимального	Сформированное владение навыками оптимального выбора современных	отчет

естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения **Код В1 (ПК-1)**		методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	выбора современных методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	методов построения и анализа математических моделей, возникающих при решении естественнонаучных задач, а также современных методов разработки и реализации алгоритмов их решения	
ЗНАТЬ: современные алгоритмы разработки программного обеспечения вычислительных комплексов; **Код З1 (ПК-3)**	Отсутствие знаний	Фрагментарные представления о современных алгоритмах разработки программного обеспечения вычислительных комплексов	В целом сформированные, но неполные знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Сформированные, но содержащие отдельные пробелы знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Сформированные систематические знания о современных алгоритмах компьютерной математики, о математической теории, лежащей в их основе	Устный экзамен
УМЕТЬ: применять современные алгоритмы разработки программного обеспечения	Отсутствие умений	Фрагментарные умения применять современные алгоритмы	В целом успешное, но не систематическое умение применять	Успешное, но содержащее отдельные пробелы	Сформированное умение применять современные алгоритмы	отчет

вычислительных комплексов **Код У1 (ПК-3)**		разработки программного обеспечения вычислительных комплексов	современные алгоритмы разработки программного обеспечения вычислительных комплексов	умениеприменять современные алгоритмы разработки программного обеспечения вычислительных комплексов	разработки программного обеспечения вычислительных комплексов	
ВЛАДЕТЬ: базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов **Код В1 (ПК-3)**	Отсутствие навыков	Фрагментарное владение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	В целом успешное, но не полное владение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	Успешное, но содержащее отдельные пробелывладение базовыми навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	Сформированное владениебазовым и навыками выбора современных алгоритмов разработки программного обеспечения вычислительных комплексов	отчет
ЗНАТЬ: современные методы интеллектуального анализа данных; **Код З1 (ПК-5)**	Отсутствие знаний	Фрагментарные представления о современных методах интеллектуального анализа данных	В целом сформированные, но неполные знания о современных методах интеллектуального анализа данных	Сформированные, но содержащие отдельные пробелы знания осовременных методах интеллектуального анализа данных	Сформированные систематические знания осовременных методах интеллектуального анализа данных	Устный экзамен
УМЕТЬ: применять современные методы интеллектуального анализа данных	Отсутствие умений	Фрагментарные уменияприменять современные методы	В целом успешное, но не систематическое умениеприменять	Успешное, но содержащее отдельные пробелы	Сформированное умение применять современные методы	отчет

Код У1 (ПК-5)		интеллектуального анализа данных	современные методы интеллектуального анализа данных	умениеприменять современные методы интеллектуального анализа данных	интеллектуального анализа данных	
ВЛАДЕТЬ: базовыми навыками выбора методов интеллектуального анализа данных **Код В1 (ПК-5)**	Отсутствие навыков	Фрагментарное владение базовыми навыками выбора методов интеллектуального анализа данных	В целом успешное, но не полное владение базовыми навыками выбора методов интеллектуального анализа данных	Успешное, но содержащее отдельные пробелывладение базовыми навыками выбора методов интеллектуального анализа данных	Сформированное владениебазовыми навыками выбора методов интеллектуального анализа данных	отчет

Фонды оценочных средств, необходимые для оценки результатов обучения

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 kближайших 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: Minsupport = 10%. Minconfidence = 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:

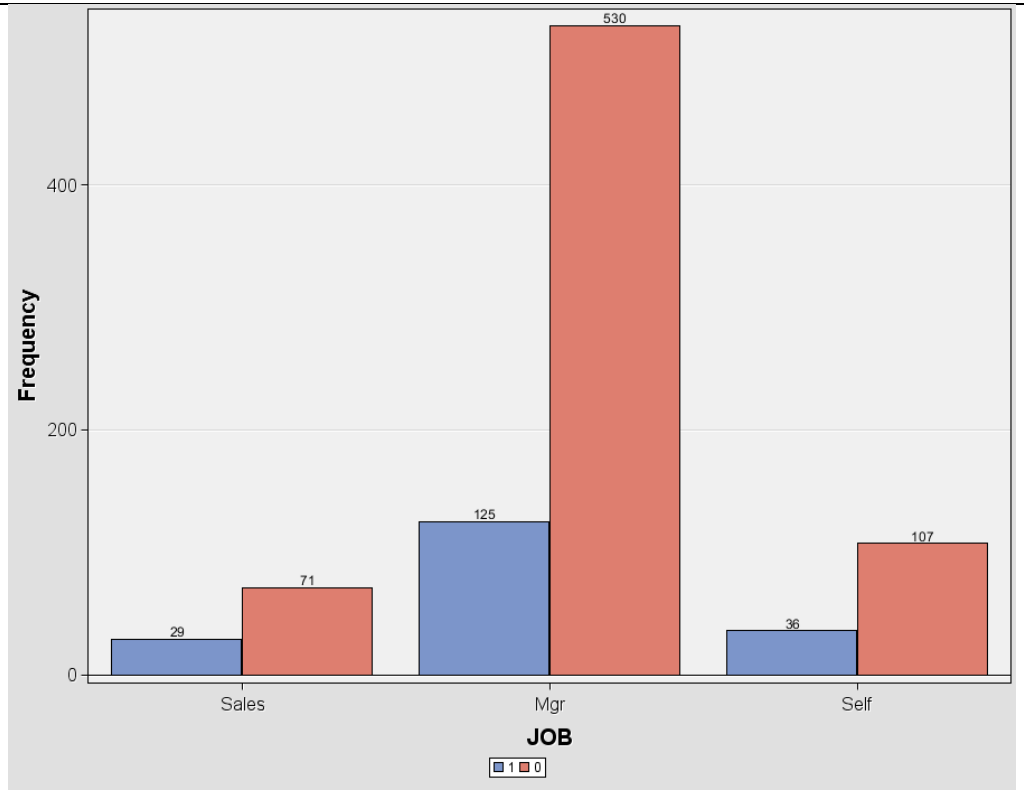
Дано множество точек:

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 (в частности для узла обработки данных Sample в SASEnterpriseMiner), а какие для процедуры фильтрации выбросов (в частности для узла обработки данных Filter в SASEnterpriseMiner)? 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: _____

Примеры ПСРприведены выше.

Методические материалы для проведения процедур оценивания результатов обучения

В течение семестра аспиранты 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 аспиранты могут получить 0–40 баллов.
За каждую ПСР аспиранты могут получить 0–30 баллов (таким образом, всего за ПСР можно получить 0–60 баллов).
Таким образом, за семестр аспиранты могут набрать 0–100 баллов.
По результатам работы в семестре, всем аспирантам 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 аспиранты 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 аспиранту любые три вопроса из списка вопросов для устной части экзамена. По результатам ответа аспиранта на вопросы, ставится итоговая оценка.