МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Российский государственный гуманитарный университет»
(ФГБОУ ВО «РГГУ»)

ОТДЕЛЕНИЕ ИНТЕЛЛЕКТУАЛЬНЫХ СИСТЕМ В ГУМАНИТАРНОЙ СФЕРЕ

Кафедра математики, логики и интеллектуальных систем в гуманитарной сфере

МАШИННОЕ ОБУЧЕНИЕ

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

46.03.02 Документоведение и архивоведение с дополнительной квалификацией
в области интеллектуальных систем в гуманитарной сфере

Интеллектуальные системы в управлении документами

Уровень высшего образования: бакалавриат

Форма обучения очная

РПД адаптирована для лиц
с ограниченными возможностями
здоровья и инвалидов

Москва 2024

Дискретная математика
Рабочая программа дисциплины
Составители:
к.т.н. Д.В. Стефановский

……

УТВЕРЖДЕНО
Протокол заседания кафедры МЛиИС
№_9____ от_04.04.2024__

ОГЛАВЛЕНИЕ

1. Пояснительная записка

1.1. Цели и задачи дисциплины

Цель дисциплины – усвоение студентами основных идей, моделей и методов машинного обучения, основанных на символьном представлении данных.

Задачи дисциплины: изложение основных алгоритмов машинного обучения, включая регрессионные, байесовские, нейросетевые методы, деревья решений и метод опорных векторов.

1.2. Перечень планируемых результатов обучения по дисциплине, соотнесенных с индикаторами достижения компетенций

Компетенции	**Индикаторы компетенций**	**Результаты обучения по дисциплине**
ОПК-7. Способен к профессиональному росту и самосовершенствованию в области гуманитарных, социальных и лингвистических наук, а также в сфере техники и технологии информатики	ОПК-7.3. Имеет практический опыт работы с поисковыми машинами, справочными и библиотечными системами и системами дистанционного образования	Знать: Алгоритмы машинного обучения: обучение с учителем, обучение без учителя, полууправляемое обучение, обучение с подкреплением Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии
ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельности	ОПК-4.3. Обладает навыками использования информационно-коммуникационных технологий в сфере документационного обеспечения управления и архивного дела	Знать: Определять теоретические верхние оценки переобученности: сложность, разделимость, устойчивость Решать проблемы переобучения и недообучения алгоритма Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии
ПК-4 Способен осуществлять проектирование и внедрение систем	ПК-4.1. обеспечивает доступ пользователей и ведение информационно-	Знать: Формировать предложения по использованию результатов анализа

электронного документооборота в организации	справочной работы в информационной системе	Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии
ПК-5. Способен организовать оперативное и архивное хранение документов с использованием интегрированной среды разработки, включая средства визуального программирования, умеет использовать средства автоматизации этапов анализа и проектирования	ПК-5.2. Имеет практический опыт разработки и тестирования прикладных программ в области оперативного и архивного хранения документов	Знать: Машинное обучение: классификация, кластеризация, обнаружение выбросов, фильтрация Методы и модели классификации: логистическая регрессия, деревья решений, предредукция, постредукция, модели, основанные на правилах, наивный байесовский алгоритм, теорема Байеса, усиление энтропии информации Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии
ПК-6. Способен к участию в разработке архитектур информационных и интеллектуальных систем в управлении документами	ПК-6.2. Знает способы представления архитектуры информационных и интеллектуальных систем и примеры типичных архитектур информационных и интеллектуальных систем в управлении документами и архивном хранении	Знать: Фильтрация шумовых выбросов, виды шумовых выбросов: глобальный, контекстуальный, коллективный Анализ изображений: тепловые карты, анализ сетей, анализ пространственных данных, анализ временных рядов Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии
	ПК-6.3. Умеет применять CASE-технологии для	Знать: Методы идентификации шаблонов

	разработки и наглядного представления архитектуры информационных и интеллектуальных систем в управлении документами и архивном хранении	Методы оценки моделей: оценка качества построенной модели по тестовой выборке и анализ обобщающих способностей алгоритма Распределенный анализ данных Анализ данных в реальном времени Уметь: Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии

1.3. Место дисциплины в структуре основной образовательной программы

Дисциплина «Машинное обучение» относится к обязательной части блока дисциплин учебного плана.

Для освоения дисциплины необходимы знания, умения и владения, сформированные в ходе изучения следующих дисциплин и прохождения практик: Информатика, Дискретная математика , Теория вероятностей.

В результате освоения дисциплины формируются знания, умения и владения, необходимые для изучения следующих дисциплин и прохождения практик: Технологии искусственного интеллекта в управлении документами.

2. Структура дисциплины

Общая трудоемкость дисциплины составляет 5 з.е., 180 академических часа.

Структура дисциплины для очной формы обучения

Объем дисциплины в форме контактной работы обучающихся с педагогическими работниками и (или) лицами, привлекаемыми к реализации образовательной программы на иных условиях, при проведении учебных занятий:

Семестр	Тип учебных занятий	Количество часов
5	Лекции	40
5	Семинары	60
Всего:		100

Объем дисциплины в форме самостоятельной работы обучающихся составляет 80 академических часов.

3. Содержание дисциплины

№ п/п	Наименование раздела дисциплины	Содержание
1	Введение в машинное обучение. Цели и основная проблематика машинного обучения	Существующие, наборы данных, визуализация модели классификации. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, ранжирование. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль. Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия.
2	Методы оценки точности полученных решений, включая ROC анализ.	Линейный регрессионный анализ, чувствительность, специфичность и точность. Корреляционный анализ. Анализ выживаемости и многомерная статистика. Таблицы дожития (mortality table) и метод Каплана-Мейера (Kaplan-Meier method). Лог-ранк тест. Модель Кокса.
3	Современные регрессионные методы, включая эластичные сети, регрессионные деревья и леса. Стандартный метод наименьших квадратов. Методы распознавания	Логистическая регрессия. Автокорреляционная функция. Алгоритм Левенберга-Марквардта. Алгоритмы выбора линейных регрессионных моделей. Вспомогательные функции. Анализ регрессионных остатков. Аппроксимация Лапласа. Регрессионные деревья и леса. Методы распознавания.
4	Байесовские методы и другие статистические модели, включая логистическую регрессию и др.	Понятие о случайном процессе. Байесовский подход к статистическому оцениванию. Априорные распределения, сопряженные с наблюдаемой генеральной совокупностью. Байесовский прогноз зависимой переменной, основанный на нормальной линейной модели множественной регрессии. Проверка статистических гипотез при байесовском подходе.
5	Нейросетевые методы. Современные подходы и идеи.	Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевых функций. Теоремы Колмогорова, Стоуна, Горбаня (без доказательства). Алгоритм обратного распространения ошибок. Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети. Метод послойной настройки сети. Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage). Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM. Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена.

6	Метод опорных векторов	Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция потерь. Задача квадратичного программирования и двойственная задача. Понятие опорных векторов. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы конструктивного построения ядер. Примеры ядер. SVM-регрессия. Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM. Метод релевантных векторов RVM.
7	Решающие деревья и леса	Понятие логической закономерности. Параметрические семейства закономерностей: конъюнкции пороговых правил, синдромные правила, шары, гиперплоскости. Переборные алгоритмы синтеза конъюнкций: стохастический локальный поиск, стабилизация, редукция. Двухкритериальный отбор информативных закономерностей, парето-оптимальный фронт в (p,n)-пространстве. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Вывод критериев ветвления. Мера нечистоты (impurity) распределения. Энтропийный критерий, критерий Джини. Редукция решающих деревьев: предредукция и постредукция. Алгоритм C4.5. Деревья регрессии. Алгоритм CART. Небрежные решающие деревья (oblivious decision tree). Решающий лес. Случайный лес (Random Forest).
8	Комбинаторно-логические методы, АВО. Представление о графических моделях (Байесовские сети)	Аппарат графических моделей (байесовские и марковские сети). Аппарат байесовского вывода. Некоторые методы дискретной оптимизации. Методы структурного обучения. Факторизация байесовских сетей. Потенциалы и энергия клик, связь с байесовскими сетями.

4. Образовательные технологии

№ п/п	Наименование раздела	Виды учебных занятий	Образовательные технологии
1	**2**	**3**	4
1	Введение в машинное обучение. Цели и основная проблематика	Лекции 1-2. Семинар 1-3	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.

	машинного обучения		
2	Методы оценки точности полученных решений, включая ROC анализ.	Лекции 3-5. Семинар 4-8	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
3	Современные регрессионные методы, включая эластичные сети, регрессионные деревья и леса. Стандартный метод наименьших квадратов. Методы распознавания	Лекции 6-7. Семинар 9-12	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
4	Байесовские методы и другие статистические модели, включая логистическую регрессию и др.	Лекции 8-10. Семинар 13-16	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
5	Нейросетевые методы. Современные подходы и идеи.	Лекция 11-13 Семинар 17-20	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
6	Метод опорных векторов	Лекция 14-15 Семинар 21-23	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
7	Решающие деревья и леса	Лекция 16-17 Семинар 24-26	Проблемная лекция. Обсуждение теоретического материала. Практикум по решению задач.
8	Комбинаторно-логические	Лекция 18-20	Проблемная лекция.

	методы, АВО. Представление о графических моделях (Байесовские сети)	Семинар 27-30	Обсуждение теоретического материала. Практикум по решению задач.

В период временного приостановления посещения обучающимися помещений и территории РГГУ для организации учебного процесса с применением электронного обучения и дистанционных образовательных технологий могут быть использованы следующие образовательные технологии:
– видео-лекции**;**
– онлайн-лекции в режиме реального времени;
– электронные учебники, учебные пособия, научные издания в электронном виде и доступ к иным электронным образовательным ресурсам;
– системы для электронного тестирования;
– консультации с использованием телекоммуникационных средств.

5. Оценка планируемых результатов обучения

5.1. Система оценивания

Форма контроля	***Макс. количество баллов***	
	За одну работу	***Всего***
Текущий контроль:		
● Домашнее задание	3 балла	30 баллов
● Контрольная работа	30 баллов	30 баллов
Промежуточная аттестация (экзамен)		40 баллов
Итого за семестр (дисциплину)		100 баллов

Полученный совокупный результат конвертируется в традиционную шкалу оценок и в шкалу оценок Европейской системы переноса и накопления кредитов (European Credit Transfer System; далее – ECTS) в соответствии с таблицей:

100-балльная шкала	Традиционная шкала		Шкала ECTS
95 – 100	отлично	зачтено	A
83 – 94			B
68 – 82	хорошо		C
56 – 67	удовлетворительно		D
50 – 55			E
20 – 49	неудовлетворительно	не зачтено	FX
0 – 19			F

5.2. Критерии выставления оценки по дисциплине

Баллы/ Шкала ECTS	**Оценка по дисциплине**	**Критерии оценки результатов обучения по дисциплине**
100-83/ A,B	отлично/ зачтено	Выставляется обучающемуся, если он глубоко и прочно усвоил теоретический и практический материал, может продемонстрировать это на занятиях и в ходе промежуточной аттестации. Обучающийся исчерпывающе и логически стройно излагает учебный материал, умеет увязывать теорию с практикой, справляется с решением задач профессиональной направленности высокого уровня сложности, правильно обосновывает принятые решения. Свободно ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «высокий».
82-68/ C	хорошо/ зачтено	Выставляется обучающемуся, если он знает теоретический и практический материал, грамотно и по существу излагает его на занятиях и в ходе промежуточной аттестации, не допуская существенных неточностей. Обучающийся правильно применяет теоретические положения при решении практических задач профессиональной направленности разного уровня сложности, владеет необходимыми для этого навыками и приёмами. Достаточно хорошо ориентируется в учебной и профессиональной литературе. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «хороший».
67-50/ D,E	удовлетво-рительно/ зачтено	Выставляется обучающемуся, если он знает на базовом уровне теоретический и практический материал, допускает отдельные ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает определённые затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, владеет необходимыми для этого базовыми навыками и приёмами. Демонстрирует достаточный уровень знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции, закреплённые за дисциплиной, сформированы на уровне – «достаточный».
49-0/ F,FX	неудовлет-ворительно/ не зачтено	Выставляется обучающемуся, если он не знает на базовом уровне теоретический и практический материал, допускает грубые ошибки при его изложении на занятиях и в ходе промежуточной аттестации. Обучающийся испытывает серьёзные затруднения в применении теоретических положений при решении практических задач профессиональной направленности стандартного уровня сложности, не владеет необходимыми для этого навыками и приёмами. Демонстрирует фрагментарные знания учебной литературы по дисциплине. Оценка по дисциплине выставляются обучающемуся с учётом результатов текущей и промежуточной аттестации. Компетенции на уровне «достаточный», закреплённые за дисциплиной, не сформированы.

5.3. Оценочные средства (материалы) для текущего контроля успеваемости, промежуточной аттестации обучающихся по дисциплине

Примеры домашних заданий

При выполнении домашнего задания студентам предлагается разработать программу, решающую одну из задач использование одного из методов, предложенных на лекциях.
Студенты самостоятельно выбирает источники данных для разрабатываемой программы. Если преподаватель считает выбранные источники достаточными для успешного выполнения задания, то студент может приступать к выполнению задания. В противном случае студент должна устранить все замечания преподавателя и повторно представить тему на согласование.
После согласования темы, студентом проводится анализ с разработкой необходимых программ и алгоритмов.
Например, домашняя работа по теме «***Решающие деревья и леса***» представляется в письменной форме и содержит следующие разделы:
1. Для набора объектов производится расчет основных статистических показателей динамики.
2. Подбирается и, если необходимо, модицируется алгоритм(ы) рассматриваемый(ые) в рамках темы.
3. Представляются в письменном виде результаты машинного обучения, модифицированные алгоритмы и программы, также указываются параметры качества параметры ошибок.

Пример контрольной работы 1

Задача 1.

Вариант 1. Критерий AIC для модели регрессии построенной для некоторого набора объектов показывает насколько хорошо модель предсказывает целевой показатель. Для данного набора Выберите целевую переменную и подберите модель на основе критерия AIC (50 баллов из 100)

Вариант 2.
Выделить на числовой оси области значений показателя X с отнесениям классам К1 и К2
Показатель X в классах К1 и К2 распределён нормально с пара метрами:
К1 : математическое ожидание 0, стандартное отклонение 4;
К2 : математическое ожидание 1, стандартное отклонение 1.
Выделить на числовой оси области значений показателя X с отнесениям классам К1 и К2 байесовским классификатором. Априорные вероятности классов К1 -0.6 и К2-0.4.

Вариант 3.

Класс 1

	X1	X2	X3	X4
Об. 1	1	1	1	0
Об. 2	0	0	1	1
Об. 3	1	0	0	1
Об.	1	0	1	1

Класс 1

	X1	X2	X3	X4
Об. 1	0	1	0	0
Об. 2	0	1	1	0
Об. 3	1	0	0	0
Об.	0	1	1	0

	4							4					

Указать один тупиковый тест и один из представительных наборов

Задача 2. (50 баллов из 100) Ответьте на вопросы:

1. Что такое решающее дерево? Как по построенному дереву найти прогноз для объекта? 2. Зачем в вершинах нужны предикаты? Какие типы предикатов вы знаете? Приведите примеры.

3. Почему для любой выборки можно построить решающее дерево, имеющее нулевую ошибку на ней?

4. Почему не рекомендуется строить небинарные деревья (т.е. имеющие больше двух потомков у каждой вершины)?

5. Как задается критерий ошибки классификации? Критерий Джини? Энтропийный критерий? Какой у них смысл?

Вопросы для промежуточной аттестации

1. Расскажите об основных понятиях: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль.
2. Расскажите о следующих понятиях: линейная модель регрессии и классификации, метод наименьших квадратов, полиномиальная регрессия.
3. Для чего используются такие понятия как: чувствительность, специфичность и точность?
4. Как и для чего осуществляется корреляционный анализ?
5. В каких случая используется анализ выживаемости и многомерная статистика и для чего?
6. Что такое таблицы дожития (mortality table), также расскажите о методе Каплана-Мейера (Kaplan-Meier method)?
7. В чем назначения лог-ранк тест?
8. Что такое логистическая регрессия?
9. Для чего нужна автокорреляционная функция?
10. Расскажите о алгоритме Левенберга-Марквардта?
11. Расскажите о алгоритмах выбора линейных регрессионных моделей?
12. Как проводится анализ регрессионных остатков?
13. Что такое аппроксимация Лапласа?
14. Для чего используются регрессионные деревья и леса?
15. Расскажите о методах распознавания?

Пример задания на экзамене

Задача 1. Критерий BIC показывает насколько хорошо модель предсказывает целевой показатель. Для данного набора Выберите целевую переменную и подберите модель на основе критерия BIC

Задача 2. Ответьте на вопросы:

1. Что такое центры кластеров? Как используются на практике значения центров?
2. Зачем в модели нужны веса? Какие типы весов вы знаете? Приведите примеры.
3. Какие алгоритмы кластеризации вы знаете приведите принципы работы двух трех алгоритмов?

6. Учебно-методическое и информационное обеспечение дисциплины

6.1. Список источников и литературы

а) *Основная литература*

1. Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. — 3-е изд., перераб. и доп. — СПб.: БХВ-Петербург, 2009. — 512 с.: ил. + CD-ROM — (Учебная литература для вузов) ISBN 978-5-9775-0368-6.
2. Тель Ж. Введение в распределенные алгоритмы. Пер. с англ. – М.: МЦНМО, 2009. – 616 с.
3. Kshemkalyani A. D., Singhal M. Distributed Computing: Principles, Algorithms, and Systems. Cambridge University Press, 2008.
4. Таненбаум Э. и др. Распределенные системы. Принципы и парадигмы. – СПб.: Питер, 2003.
5. Bollobas B. Modern Graph Theory. – Corrected ed. – Springer, 2013. – 394 p.
6. Handbook of Graph Theory. Edited by Gross J.L., Yellen J., Zhang P. – 2th ed. – CRC Press, 2014. – 1632 p.
7. Handbook of Graph Drawing and Visualization. Edited by Tamassia R. – CRC Press, 2013. – 862 p.
8. Барсегян А. А. и др. Анализ данных и процессов: учеб. пособие. 3-е изд. – 2009.
9. Бизнес-аналитика. От данных к знаниям (+ CD-ROM). Авторы Николай Паклин, Вячеслав Орешков
10. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика: учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
11. Натан Марц, Джеймс Уоррен. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени.
12. Юре Лесковец, Ананд Раджараман, Джеффри Д. Ульман. Анализ больших наборов данных
13. «Управление мастер-данными». Алекс Берсон, Лоуренс Дубов
14. Gifkins, Mike; Hitchcock, David (1988). The EDI handbook. London: Blenheim Online.
15. Эдвард Тафти. Визуальное представление больших объемов информации.
16. «Искусство визуализации в бизнесе. Как представить сложную информацию простыми образами» Нейтан Яу, «Манн, Иванов и Фербер», 2013 г."
17. Корпоративные хранилища данных. Планирование, разработка и реализация. Эрик Спирли.
18. Интеграция хранилищ данных с открытыми и большими данными для решения задач финансовой организации: проблемы и подходы к решению

б) *Дополнительная литература*

1. Карау Х. и др. Изучаем Spark: молниеносный анализ данных //ДМК Пресс, М. – 2015.
2. Майер-Шенбергер, В. М14 Большие данные. Революция, которая изменит то, как мы живем, работаем и мыслим / Виктор Майер-Шенбергер, Кеннет Кукьер ; пер. с англ. Инны Гайдюк. — М. : Манн, Иванов и Фербер, 2014. — 240 с.
3. Робинсон Ян, Вебер Джим, Эифрем Эмиль Графовые базы данных: новые возможности для работы со связанными данными / пер. с англ. Р. Н. Рагимова; науч. ред. А. Н. Кисилев. – 2-е изд. – М.: ДМК Пресс, 2016. – 256 с.: ил.

4. Уэс Маккинли Python и анализ данных / Пер. с англ. Слинкин А. А. – М.: ДМК Пресс, 2015. – 482 с.: ил."
5. Руководство пользователя SAP BusinessObjects Web Intelligence
6. Райан Митчелл: Скрапинг веб-сайтов с помощью Python. Сбор данных из современного интернета
7. Spark для профессионалов: современные паттерны обработки больших данных. Риза С., Лезерсон У., Оуэн Ш., Уиллс Д.
8. Онтологическая модель представления и организации знаний Учебное пособие для вузов Цуканова Н.И.
9. Криптография и защита сетей: принципы и практика 2-е издание.Вильям Столлингс.
10. Современная криптография: теория и практика. Венбо Мао
11. R в действии. Анализ и визуализация данных на языке R.
12. Постреляционные хранилища данных. Учебное пособие. Юрий Парфенов
13. Х. Карау, Э. Конвински, П. Венделл, М. Захария "Изучаем Spark. Молниеносный анализ данных" ДМК Пресс, 2015 год, 304 стр.
14. Чак Лэм. Hadoop в действии. – М.: ДМК Пресс, 2012. – 424с.: ил.
15. Горбань А.Н., Россиев Д.А. Нейронные сети на персональном компьютере.— Новосибирск: Наука, 1996. 275 с.
16. Уэс Маккинли Python и анализ данных/ Пер. с англ. Слинкин А. А. - М.: ДМК Пресс, 2015. - 482 с.: ил.

6.2. Перечень ресурсов информационно-телекоммуникационной сети «Интернет»

Национальная электронная библиотека (НЭБ) www.rusneb.ru
ELibrary.ru Научная электронная библиотека www.elibrary.ru
Электронная библиотека Grebennikon.ru www.grebennikon.ru
Cambridge University Press
ProQuest Dissertation & Theses Global
SAGE Journals
Taylor and Francis
JSTOR
http://www.wolframalpha.com

6.3. Профессиональные базы данных и информационно-справочные системы

Доступ к профессиональным базам данных: https://liber.rsuh.ru/ru/bases

Информационные справочные системы:
1. Консультант Плюс
2. Гарант

7. Материально-техническое обеспечение дисциплины

Для обеспечения дисциплины используется материально-техническая база образовательного учреждения: учебные аудитории, оснащённые доской, а также компьютером и проектором для демонстрации учебных материалов, компьютеры для студентов.

Состав программного обеспечения:

1. Windows
2. Microsoft Office

8. Обеспечение образовательного процесса для лиц с ограниченными возможностями здоровья и инвалидов

В ходе реализации дисциплины используются следующие дополнительные методы обучения, текущего контроля успеваемости и промежуточной аттестации обучающихся в зависимости от их индивидуальных особенностей:

- для слепых и слабовидящих: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением или могут быть заменены устным ответом; обеспечивается индивидуальное равномерное освещение не менее 300 люкс; для выполнения задания при необходимости предоставляется увеличивающее устройство; возможно также использование собственных увеличивающих устройств; письменные задания оформляются увеличенным шрифтом; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.
- для глухих и слабослышащих: лекции оформляются в виде электронного документа, либо предоставляется звукоусиливающая аппаратура индивидуального пользования; письменные задания выполняются на компьютере в письменной форме; экзамен и зачёт проводятся в письменной форме на компьютере; возможно проведение в форме тестирования.
- для лиц с нарушениями опорно-двигательного аппарата: лекции оформляются в виде электронного документа, доступного с помощью компьютера со специализированным программным обеспечением; письменные задания выполняются на компьютере со специализированным программным обеспечением; экзамен и зачёт проводятся в устной форме или выполняются в письменной форме на компьютере.

При необходимости предусматривается увеличение времени для подготовки ответа.

Процедура проведения промежуточной аттестации для обучающихся устанавливается с учётом их индивидуальных психофизических особенностей. Промежуточная аттестация может проводиться в несколько этапов.

При проведении процедуры оценивания результатов обучения предусматривается использование технических средств, необходимых в связи с индивидуальными особенностями обучающихся. Эти средства могут быть предоставлены университетом, или могут использоваться собственные технические средства.

Проведение процедуры оценивания результатов обучения допускается с использованием дистанционных образовательных технологий.

Обеспечивается доступ к информационным и библиографическим ресурсам в сети Интернет для каждого обучающегося в формах, адаптированных к ограничениям их здоровья и восприятия информации:

- для слепых и слабовидящих: в печатной форме увеличенным шрифтом, в форме электронного документа, в форме аудиофайла.
- для глухих и слабослышащих: в печатной форме, в форме электронного документа.
- для обучающихся с нарушениями опорно-двигательного аппарата: в печатной форме, в форме электронного документа, в форме аудиофайла.

Учебные аудитории для всех видов контактной и самостоятельной работы, научная библиотека и иные помещения для обучения оснащены специальным оборудованием и учебными местами с техническими средствами обучения:

- для слепых и слабовидящих: устройством для сканирования и чтения с камерой SARA CE; дисплеем Брайля PAC Mate 20; принтером Брайля EmBraille ViewPlus;
- для глухих и слабослышащих: автоматизированным рабочим местом для людей с нарушением слуха и слабослышащих; акустический усилитель и колонки;
- для обучающихся с нарушениями опорно-двигательного аппарата: передвижными, регулируемыми эргономическими партами СИ-1; компьютерной техникой со специальным программным обеспечением.

9. Методические материалы

9.1. Планы семинарских занятий

Тема 1. Введение в машинное обучение. Цели и основная проблематика машинного обучения.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Существующие наборы данных, визуализация модели классификации. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, прогнозирование, ранжирование.

Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль.

Линейные модели регрессии и классификации. Метод наименьших квадратов. Полиномиальная регрессия.

Тема 2. Методы оценки точности полученных решений, включая ROC анализ.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Линейный регрессионный анализ, чувствительность, специфичность и точность. Корреляционный анализ. Анализ выживаемости и многомерная статистика. Таблицы дожития (mortality table) и метод Каплана-Мейера (Kaplan-Meier method). Лог-ранк тест. Модель Кокса.

Тема 3. Современные регрессионные методы, включая эластичные сети, регрессионные деревья и леса. Стандартный метод наименьших квадратов. Методы распознавания.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Логистическая регрессия. Автокорреляционная функция. Алгоритм Левенберга-Марквардта. Алгоритмы выбора линейных регрессионных моделей. Вспомогательные функции. Анализ регрессионных остатков. Аппроксимация Лапласа.

Регрессионные деревья и леса. Методы распознавания.

Тема 4. Байесовские методы и другие статистические модели, включая логистическую регрессию и др.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Понятие о случайном процессе. Байесовский подход к статистическому оцениванию. Априорные распределения, сопряженные с наблюдаемой генеральной совокупностью.

Байесовский прогноз зависимой переменной, основанный на нормальной линейной модели множественной регрессии. Проверка статистических гипотез при байесовском подходе.

Тема 5. Нейросетевые методы. Современные подходы и идеи.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Биологический нейрон, модель МакКаллока-Питтса как линейный классификатор. Функции активации. Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевых функций. Теоремы Колмогорова, Стоуна, Горбаня (без доказательства). Алгоритм обратного распространения ошибок. Эвристики: формирование начального приближения, ускорение сходимости, диагональный метод Левенберга-Марквардта. Проблема «паралича» сети. Метод послойной настройки сети. Подбор структуры сети: методы постепенного усложнения сети, оптимальное прореживание нейронных сетей (optimal brain damage). Нейронная сеть Кохонена. Конкурентное обучение, стратегии WTA и WTM.

Самоорганизующаяся карта Кохонена. Применение для визуального анализа данных. Искусство интерпретации карт Кохонена.

Тема 6. Метод опорных векторов.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin).

Случаи линейной разделимости и отсутствия линейной разделимости. Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция потерь. Задача квадратичного программирования и двойственная задача. Понятие опорных векторов. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера.

Способы конструктивного построения ядер. Примеры ядер.

SVM-регрессия.

Регуляризации для отбора признаков: LASSO SVM, Elastic Net SVM, SFM, RFM.

Метод релевантных векторов RVM.

Тема 7. Решающие деревья и леса.

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Понятие логической закономерности.

Параметрические семейства закономерностей: конъюнкции пороговых правил, синдромные правила, шары, гиперплоскости.

Переборные алгоритмы синтеза конъюнкций: стохастический локальный поиск, стабилизация, редукция. Двухкритериальный отбор информативных закономерностей, парето-оптимальный фронт в (p,n)-пространстве. Решающее дерево. Жадная нисходящая стратегия «разделяй и властвуй». Алгоритм ID3. Недостатки жадной стратегии и способы их устранения. Проблема переобучения. Вывод критериев ветвления. Мера нечистоты (impurity) распределения. Энтропийный критерий, критерий Джини. Редукция решающих деревьев: предредукция и постредукция. Алгоритм C4.5. Деревья регрессии. Алгоритм CART. Небрежные решающие деревья (oblivious decision tree). Решающий лес. Случайный лес (Random Forest).

Тема 8. Комбинаторно-логические методы, АВО. Представление о графических моделях (Байесовские сети)

Студент изучает предложенные материалы темы, решает задачи по теме и готовится к опросу по следующей тематике:

Аппарат графических моделей (байесовские и марковские сети). Аппарат байесовского вывода. Некоторые методы дискретной оптимизации. Методы структурного обучения. Факторизация байесовских сетей. Потенциалы и энергия клик, связь с байесовскими сетями.

АННОТАЦИЯ РАБОЧЕЙ ПРОГРАММЫ ДИСЦИПЛИНЫ

Цель дисциплины – усвоение студентами основных идей, моделей и методов машинного обучения, основанных на символьном представлении данных.

Задачи дисциплины: изложение основных алгоритмов машинного обучения, включая регрессионные, байесовские, нейросетевые методы, деревья решений и метод опорных векторов.

Дисциплина направлена на формирование следующих компетенций:
ОПК-7. Способен к профессиональному росту и самосовершенствованию в области гуманитарных, социальных и лингвистических наук, а также в сфере техники и технологии информатики
ОПК-4. Способен понимать принципы работы современных информационных технологий и использовать их для решения задач профессиональной деятельностиПК-4 Способен осуществлять проектирование и внедрение систем электронного документооборота в организации
ПК-5. Способен организовать оперативное и архивное хранение документов с использованием интегрированной среды разработки, включая средства визуального программирования, умеет использовать средства автоматизации этапов анализа и проектирования
ПК-6. Способен к участию в разработке архитектур информационных и интеллектуальных систем в управлении документами

В результате освоения дисциплины обучающийся должен:

Знать:
Алгоритмы машинного обучения: обучение с учителем, обучение без учителя, полууправляемое обучение, обучение с подкреплением
Определять теоретические верхние оценки переобученности: сложность, разделимость, устойчивость
Решать проблемы переобучения и недообучения алгоритма
Формировать предложения по использованию результатов анализа
Машинное обучение: классификация, кластеризация, обнаружение выбросов, фильтрация
Методы и модели классификации: логистическая регрессия, деревья решений, предредукция, постредукция, модели, основанные на правилах, наивный байесовский алгоритм, теорема Байеса, усиление энтропии информации
Фильтрация шумовых выбросов, виды шумовых выбросов: глобальный, контекстуальный, коллективный
Анализ изображений: тепловые карты, анализ сетей, анализ пространственных данных, анализ временных рядов
Методы идентификации шаблонов
Методы оценки моделей: оценка качества построенной модели по тестовой выборке и анализ обобщающих способностей алгоритма
Распределенный анализ данных
Анализ данных в реальном времени

Уметь:

Осуществлять поиск информации о новых и перспективных методах анализа больших данных, сравнительный анализ методов

Владеть: простейшими навыками встраивания алгоритмов машинного обучения в новые информационные технологии

По дисциплине предусмотрена промежуточная аттестация в форме экзамена.

Общая трудоемкость освоения дисциплины составляет 5 зачетных единиц, 180 часа.