МИНОБРНАУКИ РОССИИ
Ярославский государственный университет им. П.Г. Демидова

Кафедра вычислительных и программных систем

УТВЕРЖДАЮ
Декан факультета ИВТ
_____Д.Ю. Чалый
«_23_»__мая_____2023 г.

Рабочая программа дисциплины
«Машинное обучение»

Направление подготовки
01.03.02 Прикладная математика и информатика

Направленность (профиль)
«Искусственный интеллект»

Квалификация выпускника
Бакалавр

Форма обучения
очная

Программа рассмотрена на
заседании кафедры
от 21 апреля 2023 г.,
протокол № 8

Программа одобрена НМК
факультета ИВТ
протокол № 6 от
28 апреля 2023 г.

Ярославль

1. Цели освоения дисциплины

Целями дисциплины «Машинное обучение» являются изучение современных методов машинного обучения, изучение основных прикладных математических моделей и алгоритмов, формирование представления об устройстве современных программно-аппаратных комплексах обработки данных, формирование практических навыков разработки, реализации и алгоритмов машинного обучения.

2. Место дисциплины в структуре образовательной программы бакалавриата *(магистратуры, специалитета)*

Дисциплина «Машинное обучение» входит в модуль «Искусственный интеллект» и изучается в 4 семестре на основе знаний, полученных при изучении дисциплин модулей «Математика», «Дискретная математика», а также дисциплин: «Элементы действительного анализа», «Алгоритмы и алгоритмические языки», «Структуры и алгоритмы обработки данных», «Введение в искусственный интеллект». Результаты изучения дисциплины востребованы при освоении последующих дисциплин модуля «Искусственный интеллект», в ходе производственной практики и при подготовке выпускной квалификационной работы.

3. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы бакалавриата *(магистратуры, специалитета)*

Процесс изучения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО и приобретения следующих знаний, умений, навыков и (или) опыта деятельности:

Формируемая компетенция (код и формулировка)	**Индикатор достижения компетенции (код и формулировка)**	**Перечень планируемых результатов обучения**
Общепрофессиональные компетенции		

ПК-3. Способен разрабатывать и тестировать программные компоненты решения задач в системах искусственного интеллекта.	ИПК3.1 Настраивает программное обеспечение и участвует в разработке программных компонентов систем искусственного интеллекта. ИПК3.2 Разрабатывает приложения систем искусственного интеллекта. ИПК3.3 Проводит тестирование систем искусственного интеллекта.	Знать: современные методымашинного обучения такие, как классификация без(с) учителя, регрессия, градиентный спуск, SVM, деревья, бустинг, нейронные сети. Уметь: применять полученные знания в процессе анализа, разработки и реализации прикладного программного обеспечения. Владеть: практическими навыками разработки, реализации и применения алгоритмов классификации, а также оценки их качества.
ПК-4. Способен разрабатывать и применять методы машинного обучения для решения задач.	ИПК4.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения. ИПК4.2 Определяет метрики оценки результатов моделирования и критерии качества построенных моделей. ИПК4.3 Принимает участие в оценке, выборе и при необходимости разработке методов машинного обучения.	Знать: современные методымашинного обучения такие, как классификация без(с) учителя, регрессия, градиентный спуск, SVM, деревья, бустинг, нейронные сети. Уметь: применять полученные знания в процессе анализа, разработки и реализации прикладного программного обеспечения. Владеть: практическими навыками разработки, реализации и применения алгоритмов классификации, а также оценки их качества.

ПК-5. Способен использовать инструментальные средства для решения задач машинного обучения.	ИПК5.1 Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи. ИПК5.2 Разрабатывает модели машинного обучения для решения задач. ИПК5.3 Создает, поддерживает и использует системы искусственного интеллекта, включающие разработанные модели и методы, с применением выбранных инструментов машинного обучения.	Знать: современные методымашинного обучения такие, как классификация без(с) учителя, регрессия, градиентный спуск, SVM, деревья, бустинг, нейронные сети. Уметь: применять полученные знания в процессе анализа, разработки и реализации прикладного программного обеспечения. Владеть: практическими навыками разработки, реализации и применения алгоритмов классификации, а также оценки их качества.

4. Объем, структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 6 зачетных единиц, 216 акад. часов.

№ п/п	Темы (разделы) дисциплины, их содержание	Семестр	Виды учебных занятий, включая самостоятельную работу студентов, и их трудоемкость (в академических часах)						Формы текущего контроля успеваемости Форма промежуточной аттестации *(по семестрам)*
			Контактная работа					самостоятельная работа	
			лекции	практические	лабораторные	консультации	аттестационные испытания		
1.	Введение в машинное обучение.		4					10	
	в том числе с ЭО и ДОТ							*2*	
2.	Исследование данных, их визуализация и интерпретация.		4		12			10	
	в том числе с ЭО и ДОТ							*2*	
3.	Методы классификации.		4		12			12	
	в том числе с ЭО и ДОТ							*2*	
4.	Методы числового прогнозирования.		4		12			12	
	в том числе с ЭО и ДОТ							*2*	
5.	Обнаружение закономерностей на основе ассоциативных правил.		4		12			12	
	в том числе с ЭО и ДОТ							*2*	
6.	Методы кластеризации.		4		12			12	
	в том числе с ЭО и ДОТ							*2*	
7.	Методы понижения размерности данных.		4		12			12	
	в том числе с ЭО и ДОТ							*2*	
	ИТОГО		28		72			80	Экзамен
	в том числе с ЭО и ДОТ							*14*	

Содержание разделов дисциплины:

Раздел 1. Введение в машинное обучение.
Понятия «наука о данных», «машинное обучение» (далее англ. Machine learning, ML), «интеллектуальный анализ данных». Составляющие ML: хранение данных; абстрагирование; обобщение; оценка. Этапы решения задач с использованием ML: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели. Типы входных данных. Типы алгоритмов машинного обучения. Подбор

алгоритмов по входным данным. Библиотеки Python для машинного обучения. Методология ML Ops.

Раздел 2. Исследование данных, их визуализация и интерпретация.
Преобразование данных, построение выводов по данным и оценка результатов. Структуры данных. Числовые переменные. Измерение средних значений: среднее арифметическое и медиана. Измерение разброса: квартили и пятичисловая сводка. Визуализация числовых переменных: диаграммы размаха; гистограммы (разбиения по интервалам и плотность). Интерпретация числовых данных: равномерное и нормальное распределение. Измерение разброса: дисперсия и стандартное отклонение. Категориальные переменные. Мода. Взаимосвязи между переменными. Визуализация отношений: диаграммы разброса. Исследование взаимосвязей: перекрестные таблицы.

Раздел 3. Методы классификации.
Ленивое обучение, классификация с использованием метода ближайших соседей: что такое классификация методом ближайших соседей; алгоритм k-NN; измерение степени сходства с помощью расстояния; выбор подходящего k; подготовка данных для использования в алгоритме k-NN; почему алгоритм kNN называют ленивым. Вероятностное обучение, классификация с использованием наивного байесовского классификатора: наивный байесовский классификатор; основные понятия байесовских методов; наивный байесовский алгоритм; классификация по наивному байесовскому алгоритму; Критерий Лапласа; использование числовых признаков в наивном байесовском алгоритме. Классификация с использованием деревьев решений и правил: деревья решений; выбор лучшего разделения; сокращение дерева решений.
Случайные леса. Градиентный бустинг.

Раздел 4. Методы числового прогнозирования.
Прогнозирование числовых данных, регрессионные методы: понятие регрессии; простая линейная регрессия; оценка методом наименьших квадратов; корреляции; множественная линейная регрессия.

Раздел 5. Обнаружение закономерностей на основе ассоциативных правил.
Ассоциативные правила. Типы задач, решаемых с использованием ассоциативных правил. Алгоритм Apriori для поиска ассоциативных правил, преимущества и недостатки алгоритма. Измерение интересности правила: поддержка и доверие. Построение набора правил по принципу Apriori. Выявление часто покупаемых продуктов в соответствии с ассоциативными правилами.

Раздел 6. Методы кластеризации.
Что такое кластеризация. Кластеризация как задача машинного обучения. Алгоритм кластеризации методом k-средних: преимущества и недостатки метода; использование расстояния для разбиения на кластеры и внесения изменений; выбор количества кластеров. Сегментация рынка для подростков с использованием кластеризации методом k-средних.

Раздел 7. Методы понижения размерности данных.
Для чего понижают размерность данных. Метод главных компонент, новая система координат, достоинства и ограничения метода. Использование метода главных компонент для понижения размерности данных успеваемости школьников.

5. Образовательные технологии, в том числе технологии электронного обучения и дистанционные образовательные технологии, используемые при осуществлении образовательного процесса по дисциплине

6. Перечень лицензионного и (или) свободно распространяемого программного обеспечения, используемого при осуществлении образовательного процесса по дисциплине

7. Перечень современных профессиональных баз данных и информационных справочных систем, используемых при осуществлении образовательного процесса по дисциплине (при необходимости)

8. Перечень основной и дополнительной учебной литературы, ресурсов информационно-телекоммуникационной сети «Интернет», рекомендуемых для освоения дисциплины

а) основная литература

1. Андрей Бурков. Машинное обучение без лишних слов. - Санкт-Петербург : Питер, 2020. - 192 с. - ISBN 978-5-4461-1560-0. - URL: https://www.ibooks.ru/bookshelf/367991/reading (дата обращения: 10.10.2021). - Текст: электронный.
2. Плас Дж. Вандер. Python для сложных задач: наука о данных и машинное обучение. — (Серия «Бестселлеры O'Reilly»). - Санкт-Петербург : Питер, 2021. - 576 с. - ISBN 978-54461-0914-2. - URL: https://www.ibooks.ru/bookshelf/376830/reading (дата обращения: 10.10.2021). - Текст: электронный.

б) дополнительная литература

1. Пол Дейтел. Python: Искусственный интеллект, большие данные и облачные вычисления. - Санкт-Петербург : Питер, 2021. - 864 с. - ISBN 978-5-4461-1432-0. - URL: https://www.ibooks.ru/bookshelf/371701/reading (дата обращения: 10.10.2021). - Текст: электронный.

в) ресурсы сети «Интернет»

1. Электронная библиотека «Университетская библиотека online». URL: http://biblioclub.ru/.
2. Информационная система «Единое окно доступа к образовательным ресурсам». URL: http://window.edu.ru/.
3. Образовательный портал Череповецкого государственного университета. URL: https://edu.chsu.ru/.
4. Образовательная платформа Stepik, онлайн курсы: Программирование на Python: https://stepik.org/course/67/promo; Машинное обучение, URL: https://stepik.org/course/8057/promo.
5. Технологический акселератор ML START, онлайн курс. URL: https://youtube.com/playlist?list=PLrSH_ggigfrlXzHj8aLKj1cjPfwORqIxy

Приложение № 1 к рабочей программе дисциплины
«Машинное обучение»

Фонд оценочных средств
для проведения текущего контроля успеваемости
и промежуточной аттестации студентов
по дисциплине

1. Типовые контрольные задания и иные материалы,
используемые в процессе текущего контроля успеваемости

Перечень оценочных средств

Компетенции	Индикаторы достижения компетенций	Оценочные средства
ПК-3. Способен разрабатывать и тестировать программные компоненты решения задач в системах искусственного интеллекта.	ИПК3.1 Настраивает программное обеспечение и участвует в разработке программных компонентов систем искусственного интеллекта. ИПК3.2 Разрабатывает приложения систем искусственного интеллекта. ИПК3.3 Проводит тестирование систем искусственного интеллекта.	1. Задания для выполнения лабораторных работ. 2. Самостоятельная работа. 3. Вопросы к экзамену.
ПК-4. Способен разрабатывать и применять методы машинного обучения для решения задач.	ИПК4.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения. ИПК4.2 Определяет метрики оценки результатов моделирования и критерии качества построенных моделей. ИПК4.3 Принимает участие в оценке, выборе и при необходимости разработке методов машинного обучения.	1. Задания для выполнения лабораторных работ. 2. Самостоятельная работа. 3. Вопросы к экзамену.

ПК-5. Способен использовать инструментальные средства для решения задач машинного обучения.	ИПК5.1 Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи. ИПК5.2 Разрабатывает модели машинного обучения для решения задач. ИПК5.3 Создает, поддерживает и использует системы искусственного интеллекта, включающие	1. Задания для выполнения лабораторных работ. 2. Самостоятельная работа. 3. Вопросы к экзамену.
	разработанные модели и методы, с применением выбранных инструментов машинного обучения.	

5.2 Типовые контрольные задания и методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности, характеризующих этапы формирования компетенций

Образцы заданий для самостоятельной работы:

По итогам самостоятельной работы студент готовит отчет, включающий в себя ответы на вопросы и решение заданий, предполагавшихся к выполнению в ходе самостоятельной работы. Отчет сдается преподавателю в электронной форме.

Задания для самостоятельной работы по разделу дисциплины «Введение в машинное обучение»:

1. Приведите понятия «наука о данных», «машинное обучение», «большие данные», «интеллектуальный анализ данных».
2. Как Вы считаете, чем машинное обучение отличается от интеллектуального анализа данных (если эти понятия отличаются друг от друга)?
3. Приведите примеры использования методов машинного обучения.
4. Подготовьте интеллект-карту, включающую в себя представление составляющих машинного обучения: хранение данных; абстрагирование; обобщение; оценка.
5. Приведите описание этапов решения задач с использованием машинного обучения: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели.
6. Дайте описание типов входных данных, используемых при решении задач с помощью методов машинного обучения.
7. Перечислите типы алгоритмов машинного обучения.
8. Как подбирается метод машинного обучения для решения конкретной прикладной задачи? Что влияет на выбор метода?
9. Каково назначение и возможности библиотек библиотеки Python для машинного обучения (дайте заключение на основе анализа документации разработчиков библиотек).

10. Перечислите правовые нормы и стандарты в области искусственного интеллекта, действующие в РФ.
11. Каковы этические нормы и стандарты в области искусственного интеллекта?
12. Перечислите основные международные и национальные стандарты и методологии разработки автоматизированных систем и программного обеспечения, стандарты в области информационной безопасности, подходы к управлению и фундаментальные принципы работы, развития и использования технологий искусственного интеллекта.
13. Как осуществляется поиск зарегистрированных результатов интеллектуальной деятельности и средств индивидуализации?
14. Как провести исследование результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности?
15. Назовите принципы защиты прав результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности.
16. Как осуществляется защита прав результатов интеллектуальной деятельности и средств индивидуализации при создании инновационных продуктов в профессиональной деятельности?
17. Приведите описание критериев эффективности и качества функционирования системы искусственного интеллекта: точность, релевантность, достоверность, целостность, быстрота решения задач, надежность, защищенность функционирования систем искусственного интеллекта.
18. Приведите описание методов постановки задач, проведения и анализа тестовых и экспериментальных испытаний работоспособности систем искусственного интеллекта, в том числе систем машинного обучения.
19. Перечислите методы и критерии оценки качества моделей машинного обучения.
20. Приведите содержание унифицированных и обновляемых методологии описания, сбора и разметки данных, а также механизмов контроля за соблюдением указанных методологий.
21. Что такое ML Ops? Перечислите основные этапы жизненного цикла систем машинного обучения.

Задания для самостоятельной работы по разделу дисциплины «Исследование данных, их визуализация и интерпретация»:

1. Для каких целей выполняется интерпретация данных?
2. Что такое структура данных?
3. Какие базовые наборы изменений обычно применяются в числовым данным?
4. Почему в ходе исследования данных запрашивают как средние, так и медианные значения числовых переменных?
5. Что такое «пятичисловая сводка»? Для каких целей она используется?
6. Что отображает диаграмма размаха?
7. Что отображает гистограмма?
8. Как выглядит гистограмма равномерного распределения?
9. Как выглядит кривая нормального распределения?
10. Что измеряется стандартным отклонением?

11. Что гласит правило «68–95–99,7»?
12. Что отображает таблица частотности?
13. Для каких целей строится диаграмма разброса?
14. Что показывают перекрестные таблицы (кросс-таблицы, таблицы сопряженности)?

Задания для самостоятельной работы по разделу дисциплины «Методы классификации»:

1. В чем заключается суть метода k-NN?
2. Приведите примеры задач, решаемых с использованием метода k-NN.
3. Каковы преимущества метода k-NN?
4. Каковы недостатки метода k-NN?
5. Как измеряется степень сходства между экземплярами набора данных?
6. Каким образом выбирается подходящее k?
7. Что такое «минимаксная» нормализация?
8. Каким образом выполняется стандартизация по z-оценке?
9. Что такое «фиктивное» кодирование?
10. Почему алгоритм k-NN называют ленивым?
11. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikitlearn.org/stable/) в части реализации метода k-NN.
12. Изучите пример использования метода k-NN для классификации данных (https://pythonru.com/uroki/sklearn-kmeans-i-knn).
13. Что такое «вероятностное обучение»?
14. В чем заключается суть работы наивного байесовского классификатора?
15. Приведите примеры задач, решаемых с использованием наивного байесовского классификатора.
16. Каковы преимущества наивного байесовского классификатора?
17. Каковы недостатки наивного байесовского классификатора?
18. Почему алгоритм называют наивным?
19. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikitlearn.org/stable/) в части реализации наивного байесовского классификатора.
20. Изучите пример использования наивного байесовского алгоритма для классификации данных (https://russianblogs.com/article/2703524871/).
21. Для каких целей используются методы деревьев?
22. Почему группа методов получила такое название?
23. Приведите примеры задач, решаемых с использованием деревьев.
24. Что такое «рекурсивное сегментирование»?
25. Каким образом работает алгоритм дерева решений C5.0?
26. Каким образом выбирается лучшее разделение?
27. С какой целью выполняется «сокращение» дерева решений?
28. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikitlearn.org/stable/) в части реализации деревьев решений.
29. Изучите пример использования дерева решений для классификации данных (https://www.machinelearningmastery.ru/scikit-learn-decision-trees-explained803f3812290d/).

30. В чем заключается суть работы со случайными лесами?
31. Как работает градиентный бустинг?

Задания для самостоятельной работы разделу дисциплины «Методы числового прогнозирования»:

1. Для каких целей используются методы регрессии?
2. Приведите понятие регрессии.
3. Приведите примеры задач, решаемых с использованием регрессии.
4. Как определяется простая линейная регрессия?
5. Приведите описание оценки методом наименьших квадратов.
6. Как рассчитывается коэффициент корреляции Пирсона?
7. Приведите описание множественной линейной регрессии. В чем заключаются преимущества и недостатки данного метода?
8. Изучите документацию разработчиков библиотеки Scikit-learn (https://scikitlearn.org/stable/) в части реализации линейной регрессии.
9. Изучите пример использования линейной регрессии для числового прогнозирования (https://pythonru.com/uroki/linear-regression-sklearn).

Задания для самостоятельной работы разделу дисциплины «Обнаружение закономерностей на основе ассоциативных правил»:

1. В чем заключается суть метода ассоциативных правил?
2. Какие задачи решаются с использованием данного метода?
3. К какому типу методов машинного обучения относится метод ассоциативных правил?
4. В чем заключается суть метода Apriori?
5. В каких библиотеках Python реализован метод ассоциативных правил?
6. Проанализируйте документацию разработчиков библиотек. Каким образом производится обучение модели? Какие параметры необходимо указать для запуска обучения? Как проверить эффективность модели?
7. Что необходимо сделать, чтобы повысить эффективность модели?
8. Как сохранить ассоциативные правила в файл или фрейм данных?
9. Изучите пример решения задачи с использованием метода ассоциативных правил (http://datascientist.one/apriori-algorithm/).

Задания для самостоятельной работы разделу дисциплины «Методы кластеризации»:

1. Что такое «кластеризация»? Чем кластеризация отличается от классификации?
2. Какие задачи решаются с использованием методов кластеризации?
3. Перечислите известные Вам методы кластеризации.
4. В чем заключатся суть метода k-средних?
5. Перечислите достоинства и недостатки метода k-средних 6. В каких библиотеках Python реализован метод k-средних?

7. Проанализируйте документацию разработчиков библиотек. Каким образом производится обучение модели? Какие параметры необходимо указать для запуска обучения? Как проверить эффективность модели?
8. Что необходимо сделать, чтобы повысить эффективность модели?
9. Изучите пример решения задачи с использованием метода k-средних (https://coderlessons.com/tutorials/python-technologies/uznaite-mashinnoe-obuchenie-spython/ml-algoritm-klasterizatsii-k-srednikh).

Задания для самостоятельной работы разделу дисциплины «Методы понижения размерности данных»:

1. В чем заключается принцип работы алгоритма понижения размерности данных tSNE?
2. Какие задачи решаются с использованием данного алгоритма?
3. В каких библиотеках Python реализован данный алгоритм?
4. Изучите документацию разработчиков по оценщику TSHE, реализующему алгоритм понижения размерности данных t-SNE (https://scikitlearn.org/stable/modules/manifold.html#t-sne).
5. Каким образом можно выполнить визуализацию результата работы оценщика TSHE? Проанализируйте информацию разработчиков средств визуализации.

Образцы заданий для лабораторных работ:

По итогам выполнения лабораторной работы студент демонстрирует результаты работы программы преподавателю, предварительно разработав тестовые случаи, а также сдает в электронном виде отчет, содержащий порядок выполнения работы.

Лабораторная работа «Исследование данных, их визуализация и интерпретация».

Изучите документацию разработчиков библиотек Pandas, Matplotlib и выполните представленные ниже задания:

1. загрузите данные из файла usedcars.csv в dataframe usedcars;
2. отобразите структуру usedcars;
3. запросите статистику по всем числовым переменным usedcars;
4. посчитайте средние значения для всех числовых переменных usedcars;
5. посчитайте медианы для всех числовых переменных usedcars;
6. изучите пятичисловую сводку для переменных price и mileage;
7. постройте диаграммы размаха для переменных price и mileage;
8. постройте гистограмму для данных о цене и пробеге подержанных автомобилей;
9. вычислить дисперсию и стандартное отклонение по векторам price и mileage;
10. постройте таблицу частотности для данных о подержанном автомобиле;
11. вычислите моду переменных year, model и color;
12. ответьте на вопрос о соотношении цены и пробега, построив диаграмму разброса;
13. ответьте на вопрос о том, существует ли связь между моделью и цветом, построив кросс-таблицу.

Лабораторная работа «Классификация методом k-NN»

Обычный скрининг рака позволяет диагностировать и вылечить это заболевание до того, как появятся заметные симптомы. Процесс раннего выявления включает в себя исследование ткани на наличие аномальных уплотнений или новообразований. Если такое уплотнение обнаружится, то выполняется аспирационная биопсия с использованием полой тонкой иглы, которой из этого новообразования извлекают небольшое количество клеток. Затем врач рассматривает клетки под микроскопом и определяет, злокачественное это новообразование или доброкачественное. Интеллектуальная система, позволяющая автоматизировать идентификацию раковых клеток, принесла бы значительную пользу системе здравоохранения. Автоматизированные процессы, очевидно, повысят эффективность процесса выявления рака, что сократит время диагностики и позволит уделять больше внимания лечению заболевания. Интеллектуальная программа скрининга могла бы также обеспечить большую точность диагностики, исключив из процесса субъективный человеческий фактор. Напишите программу для выявления рака, применив алгоритм k-NN к исследованиям клеток, полученных при биопсии.

Лабораторная работа «Классификация с использованием наивного байесовского алгоритма»

По мере роста популярности мобильных телефонов во всем мире появились новые возможности для распространения рекламы по почте, используемые недобросовестными маркетологами. Такие рекламодатели используют короткие текстовые сообщения (СМС), чтобы привлечь потенциальных потребителей нежелательной рекламой, известной как СМСспам. Этот тип спама является особенно опасным, поскольку, в отличие от почтового спама, СМС может причинить больше ущерба из-за широкого использования мобильных телефонов. Разработка интеллектуальной программы классификации, которая бы фильтровала СМС-спам, стала бы полезным инструментом для операторов сотовой связи. Поскольку наивный байесовский алгоритм успешно применялся для фильтрации спама в электронной почте, вполне вероятно, что он также может быть применен к СМС-спаму. Однако в отличие от спама в электронной почте СМС-спам создает дополнительные проблемы для автоматических фильтров. Размер СМС часто ограничен 160 символами, что сокращает объем текста, по которому можно определить, является ли сообщение нежелательным. Такое ограничение привело к тому, что сформировался своеобразный сокращенный СМС-язык, что еще больше стирает грань между обычными сообщениями и спамом. Напишите программу для фильтрации СМС-спама, используя наивный байесовский алгоритм.

Лабораторная работа «Классификация с использованием деревьев решений»

Мировой финансовый кризис 2007–2008 годов показал, как важна прозрачность и строгость в принятии банковских решений. Когда кредиты стали менее доступными, банки ужесточили систему кредитования и обратились к машинному обучению для более точного определения рискованных кредитов. Благодаря высокой точности и возможности формулировать статистическую модель на понятном человеку языке деревья решений широко применяются в банковской сфере. Поскольку правительства многих стран тщательно следят за справедливостью кредитования, руководители банков должны быть в состоянии объяснить, почему одному заявителю было отказано в получении займа, в то время как другому одобрили выдачу кредита. Эта информация полезна и для клиентов, желающих узнать, почему их кредитный рейтинг оказался неудовлетворительным. Автоматические модели оценки кредитоспособности используются для рассылок по кредитным картам и мгновенных онлайн-процессов одобрения кредитов. Разработайте простую модель принятия решения о предоставлении кредита с использованием

алгоритма построения деревьев решений. Настройте параметры модели, чтобы свести к минимуму ошибки, которые могут привести к финансовым потерям.

Лабораторная работа «Прогнозирование числовых данных, регрессия»

Для того чтобы медицинская страховая компания могла зарабатывать деньги, необходимо, чтобы сумма ежегодных взносов превышала расходы на медицинское обслуживание бенефициаров. Следовательно, страховщики вкладывают много времени и денег в разработку моделей, которые точно прогнозируют медицинские расходы застрахованного населения. Медицинские расходы трудно оценить, поскольку самые дорогостоящие случаи происходят редко и кажутся случайными. Тем не менее некоторые ситуации являются более распространенными для определенных слоев населения. Например, рак легких чаще встречается у курильщиков, чем у некурящих, а от болезней сердца чаще страдают тучные люди. Целью этого анализа является использование данных о пациентах для прогнозирования средних расходов на медицинское обслуживание для подобных групп населения. Эти оценки могут быть использованы для создания страховых таблиц, согласно которым сумма ежегодных взносов устанавливается выше или ниже в зависимости от ожидаемых затрат на лечение. Используя регрессию, напишите программу, дающую прогноз стоимости медицинской страховки для конкретного клиента.

Лабораторная работа «Ассоциативные правила»

Анализ потребительской корзины применяется рекомендательными системами, используемыми во многих обычных и интернет-магазинах. Выявленные ассоциативные правила указывают на сочетания товаров, которые часто покупаются вместе. Знание этих паттернов позволяет создать новые способы оптимизации товаров в сети продуктовых магазинов, рекламных акций или раскладки товаров в магазине. Например, если покупатели часто приобретают на завтрак кофе или апельсиновый сок вместе с выпечкой, то, возможно, удастся повысить прибыль, если разместить выпечку поближе к кофе и сокам. Однако эти методы можно применять ко многим другим типам задач, от рекомендаций фильмов до обнаружения опасных зависимостей между лекарствами. При этом алгоритм Apriori способен эффективно обрабатывать потенциально большие наборы ассоциативных правил. Выполните анализ потребительской корзины на основе данных о транзакциях продуктового магазина.

Лабораторная работа «Кластеризация методом k-средних»

Общение с друзьями в социальных сетях, таких как Facebook, ВКонтакте, Instagram и др. стало для подростков всего мира обычным делом. Имея достаточное количество наличных денег, подростки являются желанной социально-демографической группой для компаний, которые продают закуски, напитки, электронику и средства гигиены. Миллионы подростков, посещающих такие сайты, привлекли внимание маркетологов, стремящихся найти свою нишу на все более высококонкурентном рынке. Один из способов найти такую нишу — выявление среди подростков групп, имеющих схожие вкусы, чтобы клиенты, не заинтересованные в этих товарах, не получали рекламу, ориентированную на подростков. Например, скорее всего, будет трудно продать спортивную одежду тем подросткам, которые не интересуются спортом. Исходя из информации на страницах подростков в социальных сетях, можно выделить группы с общими интересами, такими как спорт или музыка. Кластеризация может автоматизировать процесс обнаружения естественных сегментов в этой социально-возрастной группе. Однако только нам решать, насколько эти кластеры интересны и как их можно использовать для рекламы. Используя алгоритм кластеризации kсредних, напишите программу, выполняющую сегментацию рынка для подростков.

Лабораторная работа «Понижение размерности данных. Метод главных компонент»

В наборе данных содержится информация о 200 школьниках в США: их поле, этнической принадлежности, социально-экономическом статусе, типе школы, программе обучения и оценкам по пяти предметам (чтение, письмо, математика, естественные науки и социальные науки).

```
##    id female race ses schtyp prog read write math science socst
## 1  70      0    4   1      1    1   57    52   41      47    57
## 2 121      1    4   2      1    3   68    59   53      63    61
## 3  86      0    4   3      1    1   44    33   54      58    31
## 4 141      0    4   3      1    3   63    44   47      53    56
## 5 172      0    4   2      1    2   47    52   57      53    61
```

Постройте парные диаграммы рассеяния для предметов, как скоррелированы оценки между собой? Примените метод главных компонент, передав в него оценки по пяти предметам. Что описывает первая главная компонента? Какой вклад вносят предметы в первую главную компоненту? Что представляет собой вторая главная компонента? Проанализируйте связь успеваемости с категориальными переменными.

2. Список вопросов и (или) заданий для проведения промежуточной аттестации

Вопросы к экзамену:

1. Понятия «наука о данных», «машинное обучение» (далее *англ*. machine learning, ML), «большие данные», «интеллектуальный анализ данных».
2. Составляющие ML: хранение данных; абстрагирование; обобщение; оценка.
3. Этапы решения задач с использованием ML: сбор данных; исследование и подготовка данных; обучение модели; оценка модели; улучшение модели.
4. Типы входных данных.
5. Типы алгоритмов машинного обучения.
6. Подбор алгоритмов по входным данным.
7. Библиотека Scikit-Learn.
8. Методология ML Ops. Этапы жизненного цикла систем машинного обучения.
9. Преобразование данных, построение выводов по данным и оценка результатов.
10. Структуры данных. Числовые переменные.
11. Измерение средних значений: среднее арифметическое и медиана.
12. Измерение разброса: квартили и пятичисловая сводка.
13. Визуализация числовых переменных: диаграммы размаха; гистограммы (разбиения по интервалам и плотность).
14. Интерпретация числовых данных: равномерное и нормальное распределение.
15. Измерение разброса: дисперсия и стандартное отклонение.
16. Категориальные переменные. Мода.
17. Взаимосвязи между переменными.
18. Визуализация отношений: диаграммы разброса.
19. Исследование взаимосвязей: перекрестные таблицы.
20. Ленивое обучение, классификация с использованием метода ближайших соседей.

21. Вероятностное обучение, классификация с использованием наивного байесовского классификатора.
22. Классификация с использованием деревьев решений и правил.
23. Прогнозирование числовых данных, регрессионные методы.
24. Ассоциативные правила. Типы задач, решаемых с использованием ассоциативных правил.
25. Алгоритм Apriori для поиска ассоциативных правил, преимущества и недостатки алгоритма.
26. Измерение интересности правила: поддержка и доверие.
27. Построение набора правил по принципу Apriori.
28. Кластеризация как задача машинного обучения.
29. Алгоритм кластеризации методом k-средних.
30. Понижение размерности данных. Метод главных компонент, новая система координат, достоинства и ограничения метода.

Уровни оценки компетенций следующие: базовый – 55-69 баллов, повышенный – 70-100 баллов. Преподаватель проводит систематический контроль знаний студентов, ориентируясь на перечень вопросов для проведения зачета/экзамена.

Критерии оценки лабораторных работ /самостоятельной работы студента

− ***5 баллов*** выставляется студенту, если работа выполнена самостоятельно и полностью верно; представлен отчет, содержащий результаты выполнения заданий работы и ответы на вопросы для подготовки/защиты лабораторной работы; студент анализирует результаты, полученные в ходе выполнения работы, делает выводы.

− ***4 балла*** выставляется студенту, если работа выполнена самостоятельно, в целом правильно, но имеются некоторые неточности в выполнении заданий или ответах на контрольные вопросы; представлен отчет, содержащий результаты выполнения заданий и ответы на вопросы для подготовки/защиты лабораторной работы; студент анализирует результаты, полученные в ходе выполнения работы, делает выводы.

− ***3 балла*** выставляется студенту, если работа выполнена самостоятельно, в целом правильно, но имеются некоторые неточности в выполнении заданий или ответах на контрольные вопросы; представлен отчет, содержащий результаты выполнения заданий лабораторной работы и ответы на вопросы для подготовки/защиты лабораторной работы; студент испытывает затруднения при проведении анализа результатов, полученных в ходе выполнения лабораторной работы, и формулировке выводов.

− ***2 балла*** выставляется студенту, если студент не до конца справился с заданием, не совсем верно ответил на вопросы для подготовки/защиты лабораторной работы, однако оформил отчет по результатам работы.

− ***1 балл*** выставляется студенту, если студент не до конца справился с заданием, не совсем верно ответил на вопросы для подготовки/защиты лабораторной работы, не оформил отчет по результатам работы.

− ***0 баллов*** выставляется студенту, если студент не справился с заданием, неверно ответил на вопросы для подготовки/защиты лабораторной работы.

Критерии оценивания устного ответа студента на экзамене

Ответ на экзамене оценивается исходя из 40 баллов (максимум). Билет содержит теоретический вопрос и практическое задание, преподаватель может задавать дополнительные вопросы. Полный ответ на основной вопрос оценивается максимум в 20 баллов, предполагает свободное изложение (не чтение) всего необходимого материала,

ответы студента на уточняющие вопросы, если они есть. Правильный ответ на дополнительный вопрос оценивается максимум в 5 баллов. Правильное выполнение практического задания оценивается в 20 баллов.

Шкала оценивания компетенций:

Оценка в 100-балльной шкале	Оценка в 5-ти балльной шкале	Уровень сформированности компетенций
0-54 баллов	неудовлетворительно (не зачтено)	недостаточный
55-69 баллов	удовлетворительно (зачтено)	базовый
70-85 баллов	хорошо (зачтено)	повышенный
86-100 баллов	отлично (зачтено)	

Критерии оценивания компетенций:

Индикаторы достижения компетенций	**Критерии оценивания компетенций**		
	Недостаточный уровень	**Базовый уровень**	**Повышенный уровень**
ИПК3.1 Настраивает программное обеспечение и участвует в разработке программных компонентов систем искусственного интеллекта.	Не знает основные программные платформы и компоненты систем искусственного интеллекта: механизмы логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов, принципы Data Ops и Dev Ops. Не умеет настраивать основные	Знает основные программные платформы и компоненты систем искусственного интеллекта: механизмы логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов, принципы Data Ops и Dev Ops. Умеет настраивать основные	Демонстрирует свободное владение основными программными платформы и компонентами систем искусственного интеллекта: механизмы логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов, принципы Data Ops и

	программные платформы и компоненты систем искусственного интеллекта: механизмов логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов на особенности проблемной области, участвует в их разработке.	программные платформы и компоненты систем искусственного интеллекта: механизмов логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальных интерфейсов на особенности проблемной области, участвует в их разработке.	Dev Ops. Полностью верно и самостоятельно настраивает основные программные платформы и компоненты систем искусственного интеллекта: механизмы логического вывода (рассуждений), объяснений, приобретения знаний, интеллектуальные интерфейсы на особенности проблемной области, участвует в их разработке.

ИПК3.2 Разрабатывает приложения систем искусственного интеллекта.	Не знает современные языки программирования, библиотеки и программные платформы для функционального, логического, объектно-ориентированного программирования приложений систем искусственного интеллекта. Не умеет разрабатывать программные приложения систем искусственного интеллекта с использованием современных языков программирования, библиотек и программных платформ функционального, логического, объектно-ориентированного программирования.	Знает современные языки программирования, библиотеки и программные платформы для функционального, логического, объектно-ориентированного программирования приложений систем искусственного интеллекта. Умеет разрабатывать программные приложения систем искусственного интеллекта с использованием современных языков программирования, библиотек и программных платформ функционального, логического, объектно-ориентированного программирования.	Демонстрирует свободное владение современными языками программирования, библиотеками и программными платформами для функционального, логического, объектно-ориентированного программирования приложений систем искусственного интеллекта. Полностью верно и самостоятельно разрабатывает программные приложения систем искусственного интеллекта с использованием современных языков программирования, библиотек и программных платформ функционального, логического, объектно-ориентированного программирования.

ИПК3.3 Проводит тестирование систем искусственного интеллекта.	Не знает основные критерии качества систем искусственного интеллекта, методы и инструментальные средства тестирования работоспособности и качества функционирования систем искусственного интеллекта. Не умеет проводить тестирование работоспособности и качества функционирования систем искусственного интеллекта и проверять выполнение требований к системам искусственного интеллекта со стороны пользователя.	Знает основные критерии качества систем искусственного интеллекта, методы и инструментальные средства тестирования работоспособности и качества функционирования систем искусственного интеллекта. Умеет проводить тестирование работоспособности и качества функционирования систем искусственного интеллекта и проверять выполнение требований к системам искусственного интеллекта со стороны пользователя.	Глубоко знает и понимает основные критерии качества систем искусственного интеллекта, методы и инструментальные средства тестирования работоспособности и качества функционирования систем искусственного интеллекта. Полностью верно и самостоятельно проводит тестирование работоспособности и качества функционирования систем искусственного интеллекта и проверяет выполнение требований к системам искусственного интеллекта со стороны пользователя.
ИПК4.1 Проводит анализ требований и определяет необходимые классы задач машинного обучения.	Не знает: принципы и методы машинного обучения, типы и классы задач машинного обучения, методологию ML Ops; статистические методы анализа данных. Не умеет: сопоставить задачам предметной области классы задач машинного обучения; использовать статистические	Знает: принципы и методы машинного обучения, типы и классы задач машинного обучения, методологию ML Ops; статистические методы анализа данных. Умеет: сопоставить задачам предметной области классы задач машинного обучения; использовать статистические методы анализа данных при	Демонстрирует глубокое знание и понимание: принципов и методов машинного обучения, типов и классов задач машинного обучения, методологии ML Ops; статистических методов анализа данных. Полностью верно и самостоятельно: сопоставляет задачам предметной области классы задач машинного обучения;

	методы анализа данных при решении	решении задач машинного обучения.	

	задач машинного обучения.		использует статистические методы анализа данных при решении задач машинного обучения.
ИПК4.2 Определяет метрики оценки результатов моделирования и критерии качества построенных моделей.	Не знает методы и критерии оценки качества моделей машинного обучения. Не умеет определять критерии и метрики оценки результатов моделирования при построении систем искусственного интеллекта в исследуемой области.	Знает методы и критерии оценки качества моделей машинного обучения. Умеет определять критерии и метрики оценки результатов моделирования при построении систем искусственного интеллекта в исследуемой области.	Глубоко знает и понимает методы и критерии оценки качества моделей машинного обучения. Полностью верно и самостоятельно определяет критерии и метрики оценки результатов моделирования при построении систем искусственного интеллекта в исследуемой области.

ИПК4.3 Принимает участие в оценке, выборе и при необходимости разработке методов машинного обучения.	Не знает классические методы и алгоритмы машинного обучения: предиктивные — обучение с учителем, дескриптивные — обучение без учителя. Не умеет проводить сравнительный анализ и осуществлять выбор, настройку, при необходимости разработку методов и алгоритмов для решения задач машинного обучения.	Знает классические методы и алгоритмы машинного обучения: предиктивные — обучение с учителем, дескриптивные — обучение без учителя. Умеет проводить сравнительный анализ и осуществлять выбор, настройку, при необходимости разработку методов и алгоритмов для решения задач машинного обучения.	Демонстрирует глубокое знание и понимание классических методов и алгоритмов машинного обучения: предиктивные — обучение с учителем, дескриптивные — обучение без учителя. Полностью верно и самостоятельно проводит сравнительный анализ и осуществляет выбор, настройку, при необходимости разработку методов и алгоритмов для решения задач машинного обучения.
ИПК5.1 Осуществляет оценку и выбор инструментальных средств для решения поставленной задачи.	Не знает возможности современных инструментальных средств и систем программирования для решения задач анализа данных и машинного обучения.	Знает возможности современных инструментальных средств и систем программирования для решения задач анализа данных и машинного обучения. Умеет проводить	Демонстрирует свободное владение возможностями современных инструментальных средств и систем программирования для решения задач анализа данных и

	Не умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения.	сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения.	машинного обучения. Умеет проводить сравнительный анализ и осуществлять выбор инструментальных средств для решения задач машинного обучения, в том числе в новой или нестандартной ситуации.

ИПК5.2 Разрабатывает модели машинного обучения для решения задач.	Не знает: функциональные возможности современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения; принципы проведения машинного эксперимента, проблемы переобучения и недообучения модели, требования к обучающей, тестовой и валидационной выборкам для решения задач анализа данных и машинного обучения. Не умеет: применять современные инструментальные средства и системы программирования для разработки моделей машинного обучения; планировать и выполнять машинные эксперименты, оценивать точность и качество построенных моделей.	Знает: функциональные возможности современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения; принципы проведения машинного эксперимента, проблемы переобучения и недообучения модели, требования к обучающей, тестовой и валидационной выборкам для решения задач анализа данных и машинного обучения. Умеет: применять современные инструментальные средства и системы программирования для разработки моделей машинного обучения; планировать и выполнять машинные эксперименты, оценивать точность и качество построенных моделей.	Демонстрирует глубокое знание и понимание: функциональных возможностей современных инструментальных средств и систем программирования в области создания моделей и методов машинного обучения; принципов проведения машинного эксперимента, проблем переобучения и недообучения модели, требований к обучающей, тестовой и валидационной выборкам для решения задач анализа данных и машинного обучения. Полностью верно и самостоятельно: применяет современные инструментальные средства и системы программирования для разработки моделей машинного обучения; планирует и выполняет машинные эксперименты, оценивает точность и качество построенных моделей.

ИПК5.3 Создает, поддерживает и использует системы искусственного интеллекта, включающие разработанные модели и методы, с применением выбранных инструментов машинного обучения.	Не знает принципы построения систем искусственного интеллекта, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта с применением машинного обучения. Не умеет решать задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования системы искусственного интеллекта с применением машинного обучения.	Знает принципы построения систем искусственного интеллекта, методы и подходы к планированию и реализации проектов по созданию систем искусственного интеллекта с применением машинного обучения. Умеет решать задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования системы искусственного интеллекта с применением машинного обучения.	Демонстрирует свободное владение принципами построения систем искусственного интеллекта, методами и подходами к планированию и реализации проектов по созданию систем искусственного интеллекта с применением машинного обучения. Полностью верно и самостоятельно решает задачи по выполнению коллективной проектной деятельности для создания, поддержки и использования системы искусственного интеллекта с применением машинного обучения.

Приложение № 2 к рабочей программе дисциплины
«Машинное обучение»

Методические указания для студентов по освоению дисциплины

На реализацию проекта студенты делятся на команды по 1-4 человека. В команде студенты сами решают как распределить задачи между участниками. При реализации проекта перед студентами одновременно встает ряд задач: найти базу данных для классификации, разобраться с форматом базы и методами вычисления числовых признаков, придумать или выбрать алгоритм для классификации, реализовать вывод результатов и оценить качество работы. Для организации эффективной командной работы над проектом необходимо использовать системы контроля версий и, например, интернет-сервис Bitbucket (https://bitbucket.org), предоставляющий серверное хранилище исходных кодов.

Учебно-методическое обеспечение самостоятельной работы студентов по дисциплине
Для самостоятельной работы особенно рекомендуется использовать учебную литературу, указанную в разделе № 7 данной рабочей программы.
Также для подбора учебной литературы рекомендуется использовать широкий спектр интернет-ресурсов:
1. Электронно-библиотечная система «Университетская библиотека online» (www.biblioclub.ru) - электронная библиотека, обеспечивающая доступ к наиболее востребованным материалам-первоисточникам, учебной, научной и художественной литературе ведущих издательств (*регистрация в электронной библиотеке – только в сети университета. После регистрации работа с системой возможна с любой точки доступа в Internet.).
2. Для самостоятельного подбора литературы в библиотеке ЯрГУ рекомендуется использовать:
1. Личный кабинет (http://lib.uniyar.ac.ru/opac/bk_login.php) дает возможность получения on-line доступа к списку выданной в автоматизированном режиме литературы, просмотра и копирования электронных версий изданий сотрудников университета (учеб. и метод. пособия, тексты лекций и т.д.) Для работы в «Личном кабинете» необходимо зайти на сайт Научной библиотеки ЯрГУ с любой точки, имеющей доступ в Internet, в пункт меню «Электронный каталог»; пройти процедуру авторизации, выбрав вкладку «Авторизация», и заполнить представленные поля информации.
2. Электронная библиотека учебных материалов ЯрГУ (http://www.lib.uniyar.ac.ru/opac/bk_cat_find.php) содержит более 2500 полных текстов учебных и учебно-методических материалов по основным изучаемым дисциплинам, изданных в университете. Доступ в сети университета, либо по логину/паролю.
3. Электронная картотека «Книгообеспеченность» (http://www.lib.uniyar.ac.ru/opac/bk_bookreq_find.php) раскрывает учебный фонд научной библиотеки ЯрГУ, предоставляет оперативную информацию о состоянии книгообеспеченности дисциплин основной и дополнительной литературой, а также цикла дисциплин и специальностей. Электронная картотека «Книгообеспеченность» доступна в сети университета и через Личный кабинет.