Министерство просвещения Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Ульяновский государственный педагогический университет
имени И.Н. Ульянова»
(ФГБОУ ВО «УлГПУ им. И.Н. Ульянова»)

Факультет физико-математического и технологического образования
Кафедра высшей математики

УТВЕРЖДАЮ
Проректор по учебно-методической работе
_____ С.Н. Титов
«25» июня 2021 г.

ОСНОВЫ МАШИННОГО ОБУЧЕНИЯ

Программа учебной дисциплины
модуля Основы анализа данных (Data Science)
основной профессиональной образовательной программы высшего образования – программы бакалавриата по направлению подготовки
43.03.01 Сервис,
направленность (профиль) образовательной программы
Сервис

(очная форма обучения)

Составители: Цыганов А.В.,
доцент кафедры высшей математики
Кувшинова А.Н., старший преподаватель кафедры высшей математики
Голубков А.В., ассистент кафедры высшей математики

Рассмотрено и одобрено на заседании ученого совета факультета права, экономики и управления, протокол от «21» июня 2021 г. №11

ДОКУМЕНТ ПОДПИСАН
ЭЛЕКТРОННОЙ ПОДПИСЬЮ

Сертификат: 01D77C644FC1E54000000006381D0002
Владелец: Титов Сергей Николаевич
Действителен: с 19.07.2021 до 19.07.2022

Ульяновск, 2021

Место дисциплины в образовательной программе

Дисциплина «Основы машинного обучения» включена часть, формируемоую участниками образовательных отношений Блока 1 Дисциплины (модули) вариативного модуля (модуля «Основы анализа данных (Data Science)») учебного плана основной профессиональной образовательной программы высшего образования – программы бакалавриата по направлению подготовки 43.03.01 Сервис, направленность (профиль) образовательной программы «Сервис», очной формы обучения.

Дисциплина опирается на результаты обучения, сформированные в рамках дисциплин и практик: математическая статистика, описательная статистика.

Результаты изучения дисциплины являются основой для изучения дисциплин для профессиональной деятельности и прохождения государственной итоговой аттестации.

1. Перечень планируемых результатов обучения (образовательных результатов) по дисциплине

Целью освоения дисциплины «Основы машинного обучения» является является студентов педагогов с методами и инструментами визуализации данных, онлайн сервисами визуализации данных, овладение навыками использования технологий визуализации данных при формировании образовательной среды.

Задачей освоения дисциплины является знакомство с основным понятийным аппаратом по инфографике как направлению художественно-проектной деятельности, основными концепциями развития проектирования визуальных моделей массивов числовых данных, стилевых взаимоотношений и парадигм; отработка и закрепление умения и навыки по использованию различных инструментов прикладной информатики, применяемых для сбора, обработки, анализа больших массивов информации, а также создания на их основе графических моделей - визуализации

В результате освоения программы обучающийся должен овладеть следующими результатами обучения (в таблице представлено соотнесение образовательных результатов обучения по дисциплине с индикаторами достижения компетенций):

Компетенция и индикаторы ее достижения в дисциплине	Образовательные результаты дисциплины (этапы формирования дисциплины)		
	Знает	Умеет	Владеет
УК-1. Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач.		ОР-1. анализировать источники информации с точки зрения временных и пространственных условий их возникновения; ОР-2. анализировать ранее сложившиеся в науке способы оценки информации; ОР-3. аргументировано формулировать собственное суждение и оценку информации.	

2. Объем дисциплины в зачетных единицах с указанием количества академических часов, выделенных на контактную работу обучающихся с преподавателем (по видам учебных занятий) и на самостоятельную работу обучающихся

Номер семестра	Учебные занятия						Форма промежуточной аттестации
	Всего		Лекции, час	Практические занятия, час	Лабораторные занятия, час	Самостоят. работа, час	
	Трудоемк.						
	Зач. ед.	Часы					
7	3	108	18	30	-	33	Экзамен
Итого:	3	108	18	30	-	33	Экзамен

3. Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

3.1. Указание тем (разделов) и отведенного на них количества академических часов и видов учебных занятий

Наименование раздела и тем	Количество часов по формам организации обучения			
	Лекционные занятия	Практические занятия	Лабораторные занятия	Самостоятель-ная работа
7 семестр				
Машинное обучение. Знакомство с Python. Первичный и визуальный анализ данных	2	4	-	4
Деревья решений и метод ближайших соседей	2	4	-	4
Линейные модели классификации и регрессии	2	4	-	4
Композиции: бэггинг, случайный лес	2	2	-	3
Построение и отбор признаков	2	4	-	4
Обучение без учителя: PCA, кластеризация	2	4	-	4
Обучение модели на больших объемах данных	2	2		3
Анализ временных рядов с помощью Python	2	4		4
Градиентный бустинг	2	2		3
ИТОГО:	18	30	-	33

3.2. Краткое описание содержания тем (разделов) дисциплины

Краткое содержание курса (7 семестр)

1. Машинное обучение. Знакомство с Python. Первичный и визуальный анализ данных

Машинное обучение. Основные задачи машинного обучения. Краткий обзор языка Python и среды разработки Jupyter Notebook. Первичный анализ данных при помощи

библиотеки Pandas. Визуальный анализ данных: библиотеки Seaborn и Plotly. Визуализация данных высокой размерности (метод t-SNE).

2. Деревья решений и метод ближайших соседей

Дерево решений. Библиотека Scikit-learn. Построение дерева решений. Работа с количественными признаками. Дерево решений в задаче регрессии. Метод ближайших соседей. Выбор параметров модели и кросс-валидация.

3. Линейные модели классификации и регрессии

Линейная регрессия: метод наименьших квадратов, метод максимального правдоподобия, разложение ошибки на смещение и разброс (bias-variance decomposition), регуляризация линейной регрессии. Логистическая регрессия: линейный классификатор, логистическая регрессия как линейный классификатор, принцип максимального правдоподобия и логистическая регрессия, L2-регуляризация логистической функции потерь. Кривые валидации и обучения. Плюсы и минусы линейных моделей в задачах машинного обучения.

4. Композиции: бэггинг, случайный лес

Бэггинг: ансамбли, бутстрэп, бэггинг. Случайный лес: алгоритм, сравнение с деревом решений и бэггингом, параметры, вариация и декорреляционный эффект, смещение, сверхслучайные деревья, схожесть с алгоритмом k-ближайших соседей, преобразование признаков в многомерное пространство. Оценка важности признаков. Плюсы и минусы случайного леса.

5. Построение и отбор признаков

Извлечение признаков (Feature Extraction) в задачах обработки текстов, изображений, геоданных и др. Преобразования признаков (Feature transformations): нормализация и изменение распределения, взаимодействия (Interactions), заполнение пропусков. Выбор признаков (Feature selection): статистические подходы, отбор с использованием моделей, перебор.

6. Обучение без учителя: PCA, кластеризация

Метод главных компонент (PCA): теория и особенности применения, примеры использования. Кластеризация: K-means, Affinity Propagation, спектральная кластеризация, агломеративная кластеризация, метрики качества кластеризации.

7. Обучение модели на больших объемах данных

Стохастический градиентный спуск и онлайн-подход к обучению.
Работа с категориальными признаками: Label Encoding, One-Hot Encoding, Hashing trick. Библиотека Vowpal Wabbit.

8. Анализ временных рядов с помощью Python

Модуль statsmodels. Скользящее окно (Rolling Window). Экспоненциальное сглаживание: простое экспоненциальное сглаживание, двойное экспоненциальное сглаживание, тройное экспоненциальное сглаживание (модель Хольта-Винтерса).

Кросс-валидация на временных рядах, подбор параметров. Эконометрический подход: стационарность, единичные корни, сезонное авторегрессионное интегрированное скользящее среднее (SARIMA).

9. Градиентный бустинг

Введение и история бустинга. GBM алгоритм: постановка ML задачи, функциональный градиентный спуск, классический GBM алгоритм, пошаговый пример работы GBM. Функции потерь: функции потерь регрессии, функции потерь классификации, веса.

4. Перечень учебно-методического обеспечения для самостоятельной работы обучающихся по дисциплине

Самостоятельная работа студентов является особой формой организации учебного процесса, представляющая собой планируемую, познавательно, организационно и методически направляемую деятельность студентов, ориентированную на достижение конкретного результата, осуществляемую без прямой помощи преподавателя. Самостоятельная работа студентов является составной частью учебной работы и имеет целью закрепление и углубление полученных знаний и навыков, поиск и приобретение новых знаний, а также выполнение учебных заданий, подготовку к предстоящим занятиям и экзамену. Она предусматривает, как правило, разработку рефератов, написание докладов, выполнение творческих, индивидуальных заданий в соответствии с учебной программой (тематическим планом изучения дисциплины). Тема для такого выступления может быть предложена преподавателем или избрана самим студентом, но материал выступления не должен дублировать лекционный материал. Реферативный материал служит дополнительной информацией для работы на практических занятиях. Основная цель данного вида работы состоит в обучении студентов методам самостоятельной работы с учебным материалом. Для полноты усвоения тем, вынесенных в практические занятия, требуется работа с первоисточниками. Курс предусматривает самостоятельную работу студентов со специальной литературой. Следует отметить, что самостоятельная работа студентов результативна лишь тогда, когда она выполняется систематически, планомерно и целенаправленно.

Задания для самостоятельной работы предусматривают использование необходимых терминов и понятий по проблематике курса. Они нацеливают на практическую работу по применению изучаемого материала, поиск библиографического материала и электронных источников информации, иллюстративных материалов. Задания по самостоятельной работе даются по темам, которые требуют дополнительной проработки.

Общий объем самостоятельной работы студентов по дисциплине включает аудиторную и внеаудиторную самостоятельную работу студентов в течение семестра.

Аудиторная самостоятельная работа осуществляется в форме выполнения самостоятельной работы по дисциплине. Аудиторная самостоятельная работа обеспечена базой заданий для самостоятельной работы по разделам дисциплины.

Внеаудиторная самостоятельная работа осуществляется в формах:

– подготовка к практическим заданиям.

ОС-1 Самостоятельная работа

Примеры самостоятельных работ

1. Библиотеки Seaborn и Plotly. Визуализация данных высокой размерности (метод t-SNE).
2. Дерево решений. Библиотека Scikit-learn.
3. Логистическая регрессия
4. Бэггинг: ансамбли, бутстрэп, бэггинг
5. Построение и отбор признаков

6. Обучение без учителя: PCA, кластеризация
7. Работа с категориальными признаками: Label Encoding, One-Hot Encoding, Hashing trick.
8. Библиотека Vowpal Wabbit
9. Градиентный бустинг

5. Примерные оценочные материалы для проведения текущего контроля успеваемости и промежуточной аттестации обучающихся по дисциплине

Организация и проведение аттестации студента

ФГОС ВО в соответствии с принципами Болонского процесса ориентированы преимущественно не на сообщение обучающемуся комплекса теоретических знаний, но на выработку у бакалавра компетенций – динамического набора знаний, умений, навыков и личностных качеств, которые позволят выпускнику стать конкурентоспособным на рынке труда и успешно профессионально реализовываться.

В процессе оценки бакалавров необходимо используются как традиционные, так и инновационные типы, виды и формы контроля. При этом постепенно традиционные средства совершенствуются в русле компетентностного подхода, а инновационные средства адаптированы для повсеместного применения в российской вузовской практике.

Цель проведения аттестации – проверка освоения образовательной программы дисциплины-практикума через сформированность образовательных результатов.

Промежуточная аттестация осуществляется в конце семестра и завершает изучение дисциплины; помогает оценить крупные совокупности знаний и умений, формирование определенных компетенций.

Оценочными средствами текущего оценивания являются: доклад, тесты по теоретическим вопросам дисциплины, защита практических работ и т.п. Контроль усвоения материала ведется регулярно в течение всего семестра на практических (семинарских, лабораторных) занятиях.

№ п/п	**СРЕДСТВА ОЦЕНИВАНИЯ,** используемые для текущего оценивания показателя формирования компетенции	Образовательные результаты дисциплины
	Оценочные средства для текущей аттестации ОС-1. Самостоятельная работа	ОР-1. умеет анализировать источники информации с точки зрения временных и пространственных условий их возникновения; ОР-2. умеет анализировать ранее сложившиеся в науке способы оценки информации; ОР-3. умеет аргументировано формулировать собственное суждение и оценку информации
	Оценочные средства для промежуточной аттестации **зачет** ОС-2. Экзамен в форме устного собеседования	

Материалы, используемые для текущего контроля успеваемости обучающихся по дисциплине

Материалы для организации текущей аттестации представлены в п.5 программы.

Материалы, используемые для промежуточного контроля успеваемости обучающихся по дисциплине

ОС-2 Экзамен в форме устного собеседования
Примерный перечень вопросов к экзамену

1. Основы машинного обучения. Обзор языка Python и среды разработки Jupyter Notebook.
2. Визуальный анализ данных: библиотеки Seaborn и Plotly. Визуализация данных высокой размерности (метод t-SNE).
3. Дерево решений. Библиотека Scikit-learn. Построение дерева решений. Работа с количественными признаками.
4. Дерево решений в задаче регрессии. Метод ближайших соседей. Выбор параметров модели и кросс-валидация
5. Линейные модели классификации и регрессии
6. Бэггинг: ансамбли, бутстрэп, бэггинг.
7. Случайный лес: алгоритм, сравнение с деревом решений и бэггингом, параметры, вариация и декорреляционный эффект, смещение, сверхслучайные деревья, схожесть с алгоритмом k-ближайших соседей, преобразование признаков в многомерное пространство.
8. Извлечение признаков (Feature Extraction) в задачах обработки текстов, изображений, геоданных и др. Преобразования признаков.
9. Метод главных компонент (PCA). Кластеризаци я
10. Стохастический градиентный спуск и онлайн-подход к обучению.
11. Модуль statsmodels. Скользящее окно (Rolling Window). Экспоненциальное сглаживание: простое экспоненциальное сглаживание, двойное экспоненциальное сглаживание, тройное экспоненциальное сглаживание (модель Хольта-Винтерса).
12. Градиентный бустинг.

В конце изучения дисциплины подводятся итоги работы студентов на лекционных и практических занятиях путем суммирования заработанных баллов в течение семестра.

Критерии оценивания знаний обучающихся по дисциплине

Формирование балльно-рейтинговой оценки работы обучающихся

		Посещение лекций	Посещение практических занятий	Работа на практических занятиях	Экзамен
7 семестр	Разбалловка по видам работ	9x1=9 баллов	15x1=15 баллов	212 баллов	64 балла
	Суммарный макс. балл	9 баллов max	24 балла max	236 баллов Max	300 баллов max

Критерии оценивания работы обучающегося по итогам 7 семестра

Оценка	**Баллы (3 ЗЕ)**
«отлично»	271-300
«хорошо»	211-270
«удовлетворительно»	151-210
«неудовлетворительно»	150 и менее

6. Методические указания для обучающихся по освоению дисциплины

Успешное изучение курса требует от обучающихся посещения лекций, активной работы на практических занятиях, выполнения всех учебных заданий преподавателя, ознакомления с основной и дополнительной литературой.

Запись **лекции** – одна из форм активной самостоятельной работы обучающихся, требующая навыков и умения кратко, схематично, последовательно и логично фиксировать основные положения, выводы, обобщения, формулировки. В конце лекции преподаватель оставляет время (5 минут) для того, чтобы обучающиеся имели возможность задать уточняющие вопросы по изучаемому материалу. Из-за недостаточного количества аудиторных часов некоторые темы не удается осветить в полном объеме, поэтому преподаватель, по своему усмотрению, некоторые вопросы выносит на самостоятельную работу студентов, рекомендуя ту или иную литературу. Кроме этого, для лучшего освоения материала и систематизации знаний по дисциплине, необходимо постоянно разбирать материалы лекций по конспектам и учебным пособиям. В случае необходимости обращаться к преподавателю за консультацией.

Подготовка к практическим занятиям.

При подготовке к практическим занятиям студент должен изучить теоретический материал по теме занятия (использовать конспект лекций, изучить основную литературу, ознакомиться с дополнительной литературой, при необходимости дополнить конспект, делая в нем соответствующие записи из литературных источников). В случае затруднений, возникающих при освоении теоретического материала, студенту следует обращаться за консультацией к преподавателю. Идя на консультацию, необходимо хорошо продумать вопросы, которые требуют разъяснения.

В начале практического занятия преподаватель знакомит студентов с темой, оглашает план проведения занятия, выдает задания. В течение отведенного времени на выполнение работы студент может обратиться к преподавателю за консультацией или разъяснениями. В конце занятия проводится прием выполненных заданий, собеседование со студентом.

Результаты выполнения практических зданий оцениваются в баллах, в соответствии с балльно-рейтинговой системой университета.

Планы практических занятий
7 семестр

1. Машинное обучение. Знакомство с Python. Первичный и визуальный анализ данных

Машинное обучение. Основные задачи машинного обучения. Краткий обзор языка Python и среды разработки Jupyter Notebook. Первичный анализ данных при помощи библиотеки Pandas. Визуальный анализ данных: библиотеки Seaborn и Plotly. Визуализация данных высокой размерности (метод t-SNE).

2. Деревья решений и метод ближайших соседей

Дерево решений. Библиотека Scikit-learn. Построение дерева решений. Работа с количественными признаками. Дерево решений в задаче регрессии. Метод ближайших соседей. Выбор параметров модели и кросс-валидация.

3. Линейные модели классификации и регрессии

Линейная регрессия: метод наименьших квадратов, метод максимального правдоподобия, разложение ошибки на смещение и разброс (bias-variance decomposition), регуляризация линейной регрессии. Логистическая регрессия: линейный классификатор, логистическая регрессия как линейный классификатор, принцип максимального правдоподобия и логистическая регрессия, L2-регуляризация

логистической функции потерь. Кривые валидации и обучения. Плюсы и минусы линейных моделей в задачах машинного обучения.

4. Композиции: бэггинг, случайный лес

Бэггинг: ансамбли, бутстрэп, бэггинг. Случайный лес: алгоритм, сравнение с деревом решений и бэггингом, параметры, вариация и декорреляционный эффект, смещение, сверхслучайные деревья, схожесть с алгоритмом k-ближайших соседей, преобразование признаков в многомерное пространство. Оценка важности признаков. Плюсы и минусы случайного леса.

5. Построение и отбор признаков

Извлечение признаков (Feature Extraction) в задачах обработки текстов, изображений, геоданных и др. Преобразования признаков (Feature transformations): нормализация и изменение распределения, взаимодействия (Interactions), заполнение пропусков. Выбор признаков (Feature selection): статистические подходы, отбор с использованием моделей, перебор.

6. Обучение без учителя: PCA, кластеризация

Метод главных компонент (PCA): теория и особенности применения, примеры использования. Кластеризация: K-means, Affinity Propagation, спектральная кластеризация, агломеративная кластеризация, метрики качества кластеризации.

7. Обучение модели на больших объемах данных

Стохастический градиентный спуск и онлайн-подход к обучению.
Работа с категориальными признаками: Label Encoding, One-Hot Encoding, Hashing trick. Библиотека Vowpal Wabbit.

8. Анализ временных рядов с помощью Python

Модуль statsmodels. Скользящее окно (Rolling Window). Экспоненциальное сглаживание: простое экспоненциальное сглаживание, двойное экспоненциальное сглаживание, тройное экспоненциальное сглаживание (модель Хольта-Винтерса).
Кросс-валидация на временных рядах, подбор параметров. Эконометрический подход: стационарность, единичные корни, сезонное авторегрессионное интегрированное скользящее среднее (SARIMA).

9. Градиентный бустинг

Введение и история бустинга. GBM алгоритм: постановка ML задачи, функциональный градиентный спуск, классический GBM алгоритм, пошаговый пример работы GBM. Функции потерь: функции потерь регрессии, функции потерь классификации, веса.

7. Перечень основной и дополнительной учебной литературы, Интернет-ресурсов, необходимых для освоения дисциплины

Основная литература

1.Разработка мультимедийных приложений с использованием библиотек OpenCV и IPP: курс : учебное пособие / А. Бовырин, П. Дружков, В. Ерухимов [и др.]. – 2-е изд., исправ. –

Москва : Национальный Открытый Университет «ИНТУИТ», 2016. – 516 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=429192 (дата обращения: 03.12.2021). – Библиогр. в кн. – Текст : электронный.
2.Введение в разработку мультимедийных приложений с использованием библиотек OpenCV и IPP: курс : учебное пособие / А. Бовырин, П. Дружков, В. Ерухимов [и др.]. – 2-е изд., исправ. – Москва : Национальный Открытый Университет «ИНТУИТ», 2016. – 382 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=429234 (дата обращения: 03.12.2021). – Библиогр. в кн. – Текст : электронный.
3.Воронина, В. В. Теория и практика машинного обучения : учебное пособие / В. В. Воронина. — Ульяновск : УлГТУ, 2017. — 290 с. — ISBN 978-5-9795-1712-4. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/165053 (дата обращения: 03.12.2021). — Режим доступа: для авториз. пользователей.

Дополнительная литература

4.Чубукова, И. А. Data Mining : учебное пособие : [16+] / И. А. Чубукова. – 2-е изд., испр. – Москва : Интернет-Университет Информационных Технологий (ИНТУИТ) : Бином. Лаборатория знаний, 2008. – 383 с. – (Основы информационных технологий). – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=233055 (дата обращения: 03.12.2021). – ISBN 978-5-94774-819-2. – Текст : электронный.

5.Бутаков, Н. А. Обработка больших данных с Apache Spark : учебно-методическое пособие : [16+] / Н. А. Бутаков, М. В. Петров, Д. Насонов. – Санкт-Петербург : Университет ИТМО, 2019. – 52 с. : ил. – Режим доступа: по подписке. – URL: https://biblioclub.ru/index.php?page=book&id=566771 (дата обращения: 03.12.2021). – Библиогр. в кн. – Текст : электронный.

Интернет-ресурсы

1. Открытый курс машинного обучения. URL: https://habr.com/ru/company/ods/blog/322626/
2. Материалы открытого курса OpenDataScience и Mail.Ru Group по машинному обучению и новый запуск. URL: https://habr.com/ru/company/mailru/blog/344044/