Программа разработана экспертами
Федерального учебно-методического объединения
высшего образования по укрупненной группе
специальностей и направлений подготовки
45.00.00 Языкознание и литературоведение

Утверждена на заседании ФУМО
25 мая 2021 года

Примерная программа учебной дисциплины

МАШИННОЕ ОБУЧЕНИЕ

Уровень высшего образования:

БАКАЛАВРИАТ

Направление подготовки:

45.03.03 «ФУНДАМЕНТАЛЬНАЯ И ПРИКЛАДНАЯ ЛИНГВИСТИКА»

Москва 2021

Раздел 1. Характеристики учебных занятий

1.1 Цели и задачи учебных занятий

Целью данного курса является ознакомление студентов с основными задачами и методами обработки текстов на естественном языке. Результатом занятий должно стать приобретение студентами навыков работы в области автоматической обработки текстов на естественном языке: применения методов машинного обучения к языковым данным и тестирования готовых систем.

1.2 Место дисциплины (модуля) в структуре образовательной программы, связь с другими дисциплинами (модулями) программы

Относится к вариативной части ОПОП ВО.

1.3 Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)

Устанавливаются образовательной организацией.

1.4 Перечень результатов обучения

По окончании курса студент должен знать основные методы автоматической обработки языковых данных и методологию тестирования готовых систем; уметь самостоятельно подбирать метод для решения лингвистической задачи, реализовывать тестирование системы, подбирая адекватный набор данных, выполняя эталонную разметку и выбирая подходящую меру качества; владеть навыками применения методов машинного обучения к языковым данным и методами расчета мер качества при тестировании.
Набор компетенций, соотнесенных с результатами обучения, определяется образовательной организацией.

1.5 Перечень рекомендуемых образовательных технологий

В преподавании дисциплины «Машинное обучение» используются разнообразные образовательные технологии как традиционного, так и инновационного характера, учитывающие смешанный, теоретико- и практикоориентированный характер дисциплины:

- лекции;
- практические занятия;
- дискуссии;
- выступления с докладами и сообщениями;
- аудиторные контрольные работы;
- внеаудиторные контрольные работы;
- тестирование.

Степень необходимости образовательной среды и ее выбор определяется образовательной организацией. Формы текущей аттестации определяются образовательной организацией.

1.6 Объем дисциплины (модуля) в зачетных единицах

2 з.е.

Раздел 2. Организация, структура и содержание учебных занятий

2.1 Организация учебных занятий

Предусмотрены учебные занятия с использованием дистанционных технологий.

2.2 Краткая аннотация содержания дисциплины (модуля)

Наименование темы (раздела, части)	Вид учебных занятий	Кол-во часов
1. Машинное обучение как раздел искусственного интеллекта. Основные задачи машинного обучения.	Практические занятия	2
2. Задачи автоматической обработки текстов, которые могут быть представлены как задачи классификации. Наивный байесовский классификатор. Методы оценки качества машинного обучения. Обучающая и тестовая выборка. Процедура скользящего контроля (cross-validation). Программирование алгоритмов.	Лекции (2) Практические занятия (2)	4
3. Методы кластеризации. Иерархическая кластеризация. Расстояние между кластерами. Методы его пересчёта. Метод k-средних, его сходимость. Программирование алгоритмов.	Практические занятия	2
4. Регрессия. Линейная регрессия: метод наименьших квадратов, метод максимального правдоподобия, регуляризация линейной регрессии. Логистическая регрессия как линейный классификатор. Программирование алгоритмов.	Практические занятия	2
5. Методы автоматической классификации текстов. Метод Байеса для автоматической классификации текстов. Векторная модель представления документов. Метод опорных векторов. Применение машинного обучения в зависимости от размера обучающей коллекции. Программирование алгоритмов.	Лекции (2) Практические занятия (2)	4
6. Нейронные сети. Устройство нейронов, пороговые функции (функции активации). Основная идея обучения. Векторные представления слов (word embeddings) на основе нейронных сетей. Программа Word2vec и ее применения. Программирование алгоритмов.	Лекции (2) Практические занятия (2)	4
ИТОГО		18

Раздел 3. Обеспечение учебных занятий

3.1 Методические указания по освоению дисциплины

Преподавание дисциплины осуществляется в форме лекций и практических занятий. Во время занятий обучающиеся выполняют практические задания, иллюстрирующие основные принципы автоматической обработки текстов. Для закрепления пройденного материала предлагаются домашние задания по каждой из тем. Успешное овладение содержанием дисциплины «Машинное обучение» предполагает работу обучающихся в группах в аудитории, а также их самостоятельную работу.
Дополнительные методические указания устанавливаются образовательной организацией.

3.2 Примерный перечень учебно-методического обеспечения самостоятельной работы обучающихся по дисциплине (модулю), в том числе примерный перечень учебной литературы и ресурсов информационно-телекоммуникационной сети «Интернет», необходимых для освоения дисциплины (модуля)

Самостоятельная работа студентов должна включать усвоение теоретического материала, подготовку к практическим занятиям, выполнение творческих заданий, работу с электронным учебно-методическим комплексом, подготовку к текущему контролю знаний, к промежуточной аттестации (зачету).

Список рекомендованной литературы

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Баранов А. Н. 2017. Введение в прикладную лингвистику: [учебник]. Московский государственный университет им. М.В. Ломоносова. Изд. 5-е. Москва: URSS, Москва: ЛЕНАНД.

Введение в науку о языке. / А. Е. Кибрик и др.; под ред. О. В. Федоровой и С. Г. Татевосова. М.: Буки Веди, 2019.

Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. 2009. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ.

Маннинг К.Д. и др. 2011. Введение в информационный поиск (пер. с англ.). Издательский дом «Вильямс».

Описание материально-технической базы, рекомендуемой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория с мультимедийным комплексом.

Описание материально-технической базы (в т.ч. программного обеспечения), рекомендуемой для адаптации электронных и печатных образовательных ресурсов для обучающихся из числа инвалидов и лиц с ОВЗ

Устанавливается образовательной организацией.

3.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания

Для контроля усвоения данной дисциплины предусмотрен зачет. Мероприятия по текущему контролю знаний обучающихся проводятся в часы, отведенные для изучения дисциплины.
В течение семестра студентами выполняются практические и контрольные работы.
Порядок проведения зачета определяется ВУЗом.

3.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)

Примерные вопросы для самоконтроля:

1. Основные задачи машинного обучения.
2. Задачи автоматической обработки текстов, которые могут быть представлены как задачи классификации.
3. Что такое обучающая и тестовая выборка?
4. Основные методы автоматической рубрикации текстов.
5. Применение метода Байеса для автоматического порождения рефератов.
6. Векторная модель представления документов.
7. Линейные классификаторы. Метод опорных векторов.
8. Применение машинного обучения в зависимости от размера обучающей коллекции.
9. Нейронные сети. Устройство нейронов, пороговые функции (функции активации).
10. Векторные представления слов (word embeddings) на основе нейронных сетей.

Примерные практические задания:

1. Даны документы и их классы C1 и C2

D1=(X1, X2, X3)	C1
D2=(X1, X3, X5)	C1
D3=(X2, X4, X6)	C2

Определить класс документа с помощью Байесовского классификатора: D4 (X2, X4, X5)

Примерный перечень вопросов к зачету (экзамену) по всему курсу:

1. Каковы основные задачи машинного обучения? Каковы их особенности? Задачи автоматической обработки текстов, которые могут быть представлены как задачи классификации.
2. Методы оценки качества машинного обучения. Обучающая и тестовая выборка. Процедура скользящего контроля (cross-validation).
3. Иерархическая кластеризация. Расстояние между кластерами. Методы его пересчёта.
4. Метод k-средних, его сходимость.
5. Линейная регрессия: метод наименьших квадратов, метод максимального правдоподобия, регуляризация линейной регрессии.
6. Логистическая регрессия как линейный классификатор.
7. Основные методы автоматической рубрикации (= тематической классификации) текстов.
8. Метод Байеса для автоматической классификации текстов.
9. Применение метода Байеса для автоматического порождения рефератов.
10. Векторная модель представления документов.
11. Метод Roccio автоматической классификации текстов.
12. Линейные классификаторы. Метод опорных векторов.
13. Проблемы ручного рубрицирования, автоматического рубрицирования на основе инженерного метода и на основе машинного обучения.
14. Применение машинного обучения в зависимости от размера обучающей коллекции.
15. Нейронные сети. Устройство нейронов, пороговые функции (функции активации). Основная идея обучения.
16. Векторные представления слов (word embeddings) на основе нейронных сетей. Программа Word2vec. Применения.

3.5 Материально-техническое обеспечение

Минимально необходимый для реализации курса перечень материально-технического обеспечения включает лекционные аудитории (с компьютерным и видеопроекционным

оборудованием для презентаций, средствами звуковоспроизведения и экраном, с выходом в Интернет). Количество индивидуальных рабочих станций должно соответствовать количеству студентов.

3.6 Информационное обеспечение

Рекомендуемая обязательная литература

Jurafsky D. & Martin J. H. 2000. Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition. Upper Saddle River, N.J: Prentice Hall.

Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д. 2009. Онтологии и тезаурусы: модели, инструменты, приложения. Изд-во ИНТУИТ.

Маннинг К.Д. и др. 2011. Введение в информационный поиск (пер. с англ.). Издательский дом «Вильямс».

Рекомендуемая дополнительная литература

Баранов А. Н. 2017. Введение в прикладную лингвистику: [учебник]. Московский государственный университет им. М.В. Ломоносова. Изд. 5-е. Москва: URSS, Москва: ЛЕНАНД.

Введение в науку о языке. / А. Е. Кибрик и др.; под ред. О. В. Федоровой и С. Г. Татевосова. М.: Буки Веди, 2019.

Рекомендуемый перечень иных информационных источников

1. Портал материалов по машинному обучению machinelearning.ru
2. https://habr.com/ru/company/ods/blog/322626/
3. https://habr.com/ru/company/skillfactory/blog/504882/

Раздел 4. Разработчики программы

Лукашевич Наталья Валентиновна, доктор технических наук, профессор; Сорокин Алексей Андреевич, кандидат физико-математических наук.

Рабочая группа ФУМО 45.00.00 по проблемам искусственного интеллекта в языкознании и литературоведении.