

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

Кафедра компьютерной безопасности и математических методов обработки информации

УТВЕРЖДАЮ

Декан математического факультета



Нестеров П.Н.

21 мая 2024 г.

Рабочая программа дисциплины

Основы машинного обучения

Направление подготовки (специальности)
10.03.01 Информационная безопасность

Направленность (профиль)
«Безопасность компьютерных систем (в сфере информационных технологий)»

Форма обучения очная

Программа рассмотрена
на заседании кафедры
от 26 апреля 2024 г., протокол № 8

Программа одобрена НМК
математического факультета
протокол № 9 от 3 мая 2024 г.

1. Цели освоения дисциплины

Целью изучения дисциплины «Основы машинного обучения» является освоение обучающимися передовых знаний в области машинного обучения, а именно вопросов, связанных с возможностями машинного обучения, применительно к анализу больших данных. Для специальности «Информационная безопасность» актуальность данной дисциплины обусловлена тем, что задачи информационной безопасности по выявлению уязвимостей или нестандартного поведения программного обеспечения в большом объеме данных являются предметом рассмотрения машинного обучения.

Дисциплина обеспечивает приобретение знаний и умений в области использования алгоритмов машинного обучения, в том числе для решения задач защиты информации, способствует освоению принципов корректного применения современных средств и методов защиты информации.

Задачами освоения дисциплины «Основы машинного обучения» являются:

- приобретение навыков выбора и применения алгоритмов машинного обучения, программных средств и технологий для решения задач анализа данных;
- освоение методологии обнаружения в больших массивах данных доступных интерпретации и практически полезных закономерностей, необходимых для принятия решений в профессиональной деятельности.

2. Место дисциплины в структуре образовательной программы

Данная дисциплина относится к обязательной части образовательной программы.

Для освоения данной дисциплины обучающиеся должны владеть математическим аппаратом дискретной математики, теории вероятностей и математической статистики, программирования на языках высокого уровня.

Для успешного освоения дисциплины «Основы машинного обучения» ей должны предшествовать следующие дисциплины:

- «Информатика»;
- «Языки программирования»;
- «Дискретная математика»;
- «Теория вероятностей и математическая статистика»;
- «Численные методы оптимизации».

Дисциплина предшествует практике по получению профессиональных умений и опыта профессиональной деятельности, преддипломной практике и государственной итоговой аттестации.

3. Планируемые результаты обучения по дисциплине, соотнесенные с планируемыми результатами освоения образовательной программы

Процесс изучения дисциплины направлен на формирование следующих компетенций в соответствии с ФГОС ВО, ООП 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		
ОПК-2 Способен применять информационно-	????????????????	Знать: - классы методов и алгоритмов машинного обучения;

коммуникационные технологии, программные средства системного и прикладного назначения, в том числе отечественного производства, для решения задач профессиональной деятельности		- возможности современных инструментальных средств и систем программирования для решения задач машинного обучения
		Уметь: - ставить задачи и адаптировать методы и алгоритмы машинного обучения.
ОПК-11 Способен проводить эксперименты по заданной методике и обработку их результатов	????????????????	

4. Объем, структура и содержание дисциплины

Общая трудоемкость дисциплины составляет 3 зачетных единицы, 108 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						
			лекции	практические	лабораторные	консультации	аттестационные испытания	самостоятельная работа	
1	Задачи интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей	8	2		2			2	Задания для самостоятельной работы
2	Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные векторы, байесовские классификаторы	8	4		4	1		4	Задания для самостоятельной работы
3	Оценка эффективности и сравнительный анализ моделей обучения	8	8		8			10	Задания для самостоятельной работы
4	Ансамблирование классификаторов	8	4		4	1		6	Задания для самостоятельной работы
5	Методы кластерного анализа. Алгоритмы	8	4		4	1		6	Задания для самостоятельной

	машинного обучения: метод k -средних, EM, Cobweb								работы
6	Методы анализа ассоциаций и последовательностей	8	2		2	1		4	Задания для самостоятельной работы
7	Машинное обучение и большие данные для решения прикладных задач	8	8		8	1		6	
							0,3	0,7	Зачет
	ИТОГО		32		32	5	0,3	38,7	

Содержание разделов дисциплины:

Тема 1. Задачи интеллектуального анализа данных: регрессия, классификация, кластеризация, анализ ассоциаций и последовательностей, поиск аномалий, анализ связей.

Data Mining. OLAP и Data Mining. Задача регрессии. Обучение с учителем. Задача классификации. Обучение без учителя. Задача кластеризации. Задача анализа ассоциаций и последовательностей. Программное обеспечение для интеллектуального анализа данных. Большие данные. Масштабируемые алгоритмы.

Тема 2. Методы классификации. Алгоритмы машинного обучения: деревья решений, опорные векторы, байесовские классификаторы.

Индукция деревьев решений. Информационный выигрыш. Предредукция и постредукция. Решающие правила. Алгоритм «случайный лес». Алгоритмы ограниченного перебора. Метод опорных векторов. Байесовская и наивная байесовская классификация.

Тема 3. Оценка эффективности и сравнительный анализ моделей обучения.

Подготовка данных. Выбор значимых признаков. Наборы данных. Типы данных. Шкалы измерений. Форматы хранения данных. Качество данных. Очистка данных. Снижение размерности данных. Интеграция данных. Визуализация данных. Методы отбора значимых признаков. Фильтры. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, специфичность. Обучающее множество. Независимое тестовое множество. Подтверждающее множество. Проблема переобучения. Метод удержания. Метод перекрестной проверки. Матрица стоимостей ошибок. Диаграмма выигрыша. Диаграмма роста. Кривая ошибок. AUC. Изолинии точности.

Тема 4. Ансамблирование классификаторов.

Ансамбли моделей. Бэггинг. Бэггинг с рандомизацией. Последовательно обучающиеся классификаторы. Бустинг ансамбля классификаторов. Стэкинг.

Тема 5. Методы кластерного анализа.

Алгоритмы машинного обучения: метод k -средних, EM, Cobweb. Типологический и таксономический анализ. Статистические методы кластеризации. Метод k -средних. Меры расстояний. Иерархические методы кластеризации. Визуализация кластеров. Дендрограммы. Диаграммы рассеивания. Самоорганизующиеся карты Кохонена. Графовые методы кластеризации. Выделение связанных компонент. Нечеткая кластеризация.

Тема 6. Методы анализа ассоциаций и последовательностей.

Поиск часто встречающихся наборов элементов. Меры интересности: поддержка, достоверность. Количественные и нечеткие ассоциативные правила. Поиск последовательных шаблонов.

Тема 7. Машинное обучение и большие данные для решения прикладных задач.

Контекстная реклама, совместная фильтрация, анализ тональности, чат-боты, большие языковые модели.

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

В процессе обучения используются следующие образовательные технологии:

Академическая лекция (или лекция общего курса) – последовательное изложение материала, осуществляемое преимущественно в виде монолога преподавателя. Требования к академической лекции: современный научный уровень и насыщенная информативность, убедительная аргументация, доступная и понятная речь, четкая структура и логика, наличие ярких примеров, научных доказательств, обоснований, фактов.

Проблемная лекция – изложение материала, предполагающее постановку проблемных и дискуссионных вопросов, освещение различных научных подходов, авторские комментарии, связанные с различными моделями интерпретации изучаемого материала. Проблемная лекция начинается с вопросов, с постановки проблемы, которую в ходе изложения материала необходимо решить. В лекции сочетаются проблемные и информационные начала. При этом процесс познания студентов в сотрудничестве и диалоге с преподавателем приближается к поисковой, исследовательской деятельности. Содержание проблемы раскрывается путем организации поиска ее решения или суммирования и анализа традиционных и современных точек зрения.

Лабораторная работа – организация учебной работы с реальными материальными и информационными объектами.

Консультации – вид учебных занятий, являющийся одной из форм контроля самостоятельной работы студентов. На консультациях по просьбе студентов рассматриваются наиболее сложные моменты при освоении материала дисциплины, преподаватель отвечает на вопросы студентов, которые возникают у них в процессе самостоятельной работы.

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

В процессе осуществления образовательного процесса используются:

для формирования материалов для текущего контроля успеваемости и проведения промежуточной аттестации, для формирования методических материалов по дисциплине:

- программы Microsoft Office;
- издательская система LaTeX;
- Adobe Acrobat Reader;

при проведении практических занятий используется программное обеспечение:

- Python;
- PyTorch.

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

В процессе осуществления образовательного процесса по дисциплине используется:

- Автоматизированная библиотечно-информационная система «БУКИ-NEXT»

http://www.lib.uniyl.ac.ru/opac/bk_cat_find.php

- Электронная библиотечная система «Лань» <https://e.lanbook.com>

- Электронная библиотечная система «Юрайт» <https://urait.ru>

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

а) основная литература

1. Д. Келлехер, Б. Тирни Наука о данных: базовый курс – Москва: Альпина Паблишер, 2020.
2. Чубукова И. А. Data Mining: учебное пособие — Москва: ИНТУИТ, 2016.
<https://www.studentlibrary.ru/ru/book/ISBN9785947748192.html>

б) дополнительная литература

1. Маккинли, Уэс Python и анализ данных — Москва: ДМК Пресс, 2015.
<https://www.studentlibrary.ru/ru/book/ISBN9785970603154.html>

в) ресурсы сети «Интернет» (при необходимости)

1. <http://github.com/>
2. <http://habr.com/>
3. <http://www.kdnuggets.com/>
4. Python, Свободное ПО - <https://www.python.org/>

9. Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине

Материально-техническая база, необходимая для осуществления образовательного процесса по дисциплине включает в свой состав специальные помещения:

- учебные аудитории для проведения занятий лекционного типа;
- учебные аудитории для проведения лабораторных работ, оснащенные средствами вычислительной техники, с установленным программным обеспечением;
- учебные аудитории для проведения групповых и индивидуальных консультаций;
- учебные аудитории для проведения текущего контроля и промежуточной аттестации;
- помещения для самостоятельной работы;
- помещения для хранения и профилактического обслуживания технических средств обучения.

Помещения для самостоятельной работы обучающихся оснащены компьютерной техникой с возможностью подключения к сети «Интернет» и обеспечением доступа к электронной информационно-образовательной среде ЯрГУ.

Автор(ы):

Доцент кафедры КБиММОИ, канд. физ.-мат. наук

Д.М. Мурин

**Приложение № 1 к рабочей программе дисциплины
«Основы машинного обучения»**

**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**

**1.1 Контрольные задания и иные материалы,
используемые в процессе текущей аттестации**

Задания для самостоятельной работы

1. Может ли в методе k ближайших соседей при $k = 2$ получиться лучший результат, чем при $k = 1$? Отказы от классификации тоже считать ошибками
2. Покажите, что с ростом размерности пространства признаков при равномерном распределении точек в кубе $[0; 1]^d$ вероятность попасть в куб $[0; 0,99]^d$ стремится к нулю. Это одна из иллюстраций проклятия размерностей (dimension curse). Попробуйте придумать или найти еще какую-нибудь иллюстрацию к этому явлению и кратко изложить. В чем по-вашему суть проклятия размерности и какое это имеет значение для задач машинного обучения?
3. Покажите асимптотическую эквивалентность энтропийного и статистического критериев информативности.
4. Какая стратегия поведения в листьях решающего дерева приводит к меньшей вероятности ошибки: отвечать тот класс, который преобладает в листе, или отвечать случайно с тем же распределением классов, что и в листе?
5. Unsupervised решающие деревья можно было бы применить для кластеризации выборки или оценки плотности, но проблема построения таких деревьев заключается в введении меры информативности. В одной статье предлагался следующий подход: давайте использовать тот же Information Gain, а энтропию множества S теперь оценивать по формуле: $H(S) = 0,5 \ln((2\pi e) n |\Sigma|)$ Здесь Σ — оцененная по множеству матрица ковариаций, т.е. не имея других сведений, мы по умолчанию считаем, что скопления точек можно приближенно считать распределенными нормально. Убедитесь, что это выражение в самом деле задает энтропию многомерного нормального распределения
6. Как выглядит бинарный линейный классификатор?
7. Что такое отступ алгоритма на объекте? Какие выводы можно сделать из знака отступа?
8. Как классификаторы вида $a(x) = \text{sign}(\langle w, x \rangle - w_0)$ сводятся к классификаторам вида $a(x) = \text{sign}(\langle w, x \rangle)$?
9. Как выглядит запись функционала эмпирического риска через отступы? Какое значение он должен принимать для «наилучшего» алгоритма классификации?
10. Если в функционале эмпирического риска всюду написаны строгие неравенства ($M_i < 0$) можете ли вы сразу придумать параметр w для алгоритма классификации $a(x) = \text{sign}(\langle w, x \rangle)$, минимизирующий такой функционал?
11. Запишите функционал аппроксимированного эмпирического риска, если выбрана функция потерь $L(M)$.

12. Что такое функция потерь, зачем она нужна? Как обычно выглядит ее график?
13. В чем практический смысл квадратичной функции потерь? Почему может быть полезна функция потерь, принимающая большие значения для большого положительного отступа?
14. Приведите пример негладких и немонотонных функций потерь.
15. Что такое регуляризация? Какие регуляризаторы вы знаете?
16. Как связаны переобучение и обобщающая способность алгоритма. Как влияет регуляризация на обобщающую способность?
17. Как связаны острые минимумы функционала аппроксимированного эмпирического риска с проблемой переобучения?
18. Что делает регуляризация с аппроксимированным риском как функцией параметров алгоритма?
19. Для какого алгоритма классификации функционал аппроксимированного риска будет принимать большее значение на обучающей выборке: для построенного с регуляризацией или без нее? Почему?
20. Как построить ROC-кривую (нужен алгоритм), если например, у вас есть правильные ответы к домашнему заданию про фамилии и ваши прогнозы?
21. Придумайте ядро, которое позволит линейному классификатору с помощью Kernel Trick построить в исходном пространстве признаков разделяющую поверхность $(x_1)^2 + 2(x_2)^2 = 3$. Какой будет размерность спрямляющего пространства?
22. Чему будет равна размерность минимального спрямляющего пространства для ядра $K(w, x) = (\langle w, x \rangle + 1)^2$?
23. Покажите (хотя бы на уровне «создания очевидности»), чему будет равна минимальная размерность спрямляющего пространства для радиального (гауссовского) ядра.

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

Список вопросов к зачету:

1. Data Mining.
2. OLAP и Data Mining.
3. Задача регрессии.
4. Обучение с учителем. Задача классификации.
5. Обучение без учителя. Задача кластеризации.
6. Задача анализа ассоциаций и последовательностей.
7. Индукция деревьев решений.
8. Предредукция и постредукция.
9. Решающие правила.
10. Алгоритм «случайный лес».
11. Алгоритмы ограниченного перебора.
12. Метод опорных векторов.
13. Байесовская и наивная байесовская классификация.
14. Метрики качества: правильность, полнота, точность, F-мера, чувствительность, специфичность.
15. Обучающее множество. Независимое тестовое множество. Подтверждающее множество.
16. Метод удержания.
17. Метод перекрестной проверки.
18. Матрица стоимостей ошибок.
19. Ансамбли моделей.

20. Бэггинг.
21. Бэггинг с рандомизацией.
22. Последовательно обучающиеся классификаторы.
23. Бустинг ансамбля классификаторов.
24. Стэкинг.
25. Метод k-средних.
26. Типологический и таксономический анализ.
27. Статистические методы кластеризации.
28. Меры расстояний.
29. Иерархические методы кластеризации.
30. Визуализация кластеров.
31. Дендрограммы.
32. Самоорганизующиеся карты Кохонена.
33. Графовые методы кластеризации.
34. Нечеткая кластеризация.

3. Правила выставления оценки на зачете.

В процессе зачета требуется ответить на один из приведенных выше вопросов. На подготовку к ответу дается не менее 1 академического часа.

По итогам зачета выставляется одна из оценок: «зачтено», «не зачтено».

Оценка «Зачтено» выставляется студенту, который демонстрирует владение содержанием материала и понятийным аппаратом машинного обучения; умеет связывать теорию с практикой. В ответе могут допускаться отдельные неточности (несущественные ошибки), которые исправляются самим студентом после дополнительных и (или) уточняющих вопросов экзаменатора. На часть дополнительных вопросов студент может не дать ответ или дать неверный ответ.

Оценка «Не зачтено» выставляется студенту, который демонстрирует разрозненные, бессистемные знания; беспорядочно и неуверенно излагает материал; не умеет выделять главное и второстепенное, не умеет соединять теоретические положения с практикой; допускает грубые ошибки при определении понятий, вследствие непонимания их существенных и несущественных признаков и связей; дает неполные ответы, логика и последовательность изложения которых имеют существенные и принципиальные нарушения, в ответах отсутствуют выводы. Дополнительные и уточняющие вопросы экзаменатора не приводят к коррекции ответов студента. На основную часть дополнительных вопросов студент затрудняется дать ответ или дает неверные ответы.

Оценка «Не зачтено» выставляется также студенту, который взял экзаменационный билет, но отказался дать на него ответ.

Приложение № 2 к рабочей программе дисциплины «Основы машинного обучения»

Методические указания для студентов по освоению дисциплины

Учебным планом на изучение дисциплины «Основы машинного обучения» отводится один семестр. В качестве итогового контроля предусмотрен зачет. В процессе изучения дисциплины проводятся лабораторные работы, выполняются домашние задания.

Дисциплина «Основы машинного обучения» является предшествующей для прохождения производственной и преддипломной практики и итоговой государственной аттестации обучающихся и призвана по возможности подготовить их к практической деятельности.

Для успешного освоения дисциплины важно, чтобы обучающийся уделит особенное внимание выполнению лабораторных работ. Теоретические основы, необходимые для выполнения лабораторных работ, подробно разбираются на лекционных занятиях. Основная цель выполнения лабораторных работ – дать обучающимся представление о возможном применении методов машинного обучения на практике, в том числе в сфере информационной безопасности. Для успешного выполнения лабораторных работ необходимо знать и понимать лекционный материал. Поэтому в процессе изучения дисциплины рекомендуется регулярное повторение пройденного материала, чему способствуют регулярные задания для самостоятельной работы. Материал, законспектированный на лекциях, необходимо дома еще раз прорабатывать и при необходимости дополнять информацией, полученной на консультациях, практических занятиях или из учебной литературы.

В качестве заданий для самостоятельной работы дома обучающимся предлагаются математические или практические упражнения, которые должны позволить обучающемуся лучше изучить понятия и методы, применяемые им для решения типовых задач из соответствующих разделов дисциплины. Решения задач должны быть подготовлены, оформлены и представлены в установленные сроки.

По окончании семестра изучения дисциплины обучающиеся сдают зачет. Зачет принимается по билетам, каждый из которых включает в себя два теоретических вопроса. На самостоятельную подготовку к зачету выделяется 2 дня.

Опыт преподавания дисциплины «Основы машинного обучения» говорит о сложности ее самостоятельного изучения для обучающегося, несмотря на наличие достаточно качественных учебных пособий. Это связано с насыщенностью изучаемого материала и большим числом лабораторных работ, необходимых для приобретения навыков практического использования изучаемого материала. Поэтому посещение всех аудиторных занятий является настоятельно рекомендуемым.