МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Магнитогорский государственный технический университет им. Г.И. Носова»

УТВЕРЖДАЮ
Директор ИММиМ
_____ А.С. Савинов

15.02.2022 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ (МОДУЛЯ)

ФОРМИРОВАНИЕ ОБУЧАЮЩИХ НАБОРОВ ДАННЫХ В МЕТАЛЛУРГИИ

Направление подготовки (специальность)
22.04.02 Металлургия

Направленность (профиль/специализация) программы
Искусственный интеллект в металлургии

Уровень высшего образования - магистратура

Форма обучения
очная

Институт/ факультет	Институт металлургии, машиностроения и материалообработки
Кафедра	Литейных процессов и материаловедения
Курс	1
Семестр	2

Магнитогорск
2022 год

Рабочая программа составлена на основе ФГОС ВО - магистратура по направлению подготовки 22.04.02 Металлургия (приказ Минобрнауки России от 24.04.2018 г. № 308)

Рабочая программа рассмотрена и одобрена на заседании кафедры Литейных процессов и материаловедения
21.01.2022, протокол № 6

Зав. кафедрой _____ Н.А. Феоктистов

Рабочая программа одобрена методической комиссией ИММиМ
15.02.2022 г. протокол № 6

Председатель _____ А.С. Савинов

Рабочая программа составлена:
доцент кафедры ПМиИ, канд. пед. наук _____ Л.С. Рязанова

Рецензент:
зав. кафедрой ПЭиБЖД, канд. техн. наук _____ А.Ю. Перятинский

Лист актуализации рабочей программы

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2023 - 2024 учебном году на заседании кафедры Литейных процессов и материаловедения

Протокол от __ _____ 20__ г. № __
Зав. кафедрой _____ Н.А. Феоктистов

Рабочая программа пересмотрена, обсуждена и одобрена для реализации в 2024 - 2025 учебном году на заседании кафедры Литейных процессов и материаловедения

Протокол от __ _____ 20__ г. № __
Зав. кафедрой _____ Н.А. Феоктистов

1 Цели освоения дисциплины (модуля)

Целью показать практические аспекты технологий, связанных с хранением, обработкой, подходами к анализу больших объёмов данных в металлургической промышленности. Задачами данного курса является: - изучение источников информации на объектах металлургического производства для анализа и формирования наборов данных для моделей машинного обучения в металлургии; - приобретение теоретических и практических знаний в части сбора, обработки и хранения данных; - приобретение навыков формирования наборов данных для моделей машинного обучения в металлургии. Краткое содержание дисциплины

2 Место дисциплины (модуля) в структуре образовательной программы

Дисциплина Формирование обучающих наборов данных в металлургии входит в часть учебного плана формируемую участниками образовательных отношений образовательной программы.

Для изучения дисциплины необходимы знания (умения, владения), сформированные в результате изучения дисциплин/ практик:

Искусственные нейронные сети

Искусственный интеллект и машинное обучение

3 Компетенции обучающегося, формируемые в результате освоения дисциплины (модуля) и планируемые результаты обучения

В результате освоения дисциплины (модуля) «Формирование обучающих наборов данных в металлургии» обучающийся должен обладать следующими компетенциями:

Код индикатора	Индикатор достижения компетенции
ПК-9 Способен управлять проектами по созданию, поддержке и использованию систем бизнес-аналитики в организации со стороны заказчика	
ПК-9.1	Знает: задачи и роль систем бизнес-аналитики в поддержке принятия решений в процессе управления организацией, принципы построения систем бизнес-аналитики
ПК-9.2	Умеет: моделировать и анализировать процессы принятия управленческих решений и разрабатывать требования к системам бизнес-анализа в различных сферах деятельности
ПК-9.3	Имеет практический опыт: участия в проектах по изучению опыта создания, поддержке и использованию систем бизнес-аналитики в металлургии
ПК-10 Способен адаптировать и применять методы и алгоритмы машинного обучения для решения прикладных задач	
ПК-10.1	Знает: классы методов и алгоритмов машинного обучения
ПК-10.2	Умеет: ставить задачи и адаптировать методы и алгоритмы машинного обучения
ПК-10.3	Имеет практический опыт: участия в проектах по изучению опыта адаптации и применимости методов и алгоритмов машинного обучения для решения прикладных задач в металлургии

4. Структура, объём и содержание дисциплины (модуля)

Общая трудоемкость дисциплины составляет 4 зачетных единиц 144 акад. часов, в том числе:

– контактная работа – 48 акад. часов:

– аудиторная – 48 акад. часов;

– внеаудиторная – 0 акад. часов;

– самостоятельная работа – 96 акад. часов;

– в форме практической подготовки – 0 акад. час;

Форма аттестации - курсовая работа, зачет

Раздел/ тема дисциплины	Семестр	Аудиторная контактная работа (в акад. часах)			Самостоятельная работа студента	Вид самостоятельной работы	Форма текущего контроля успеваемости и промежуточной аттестации	Код компетенции
		Лек.	лаб. зан.	практ. зан.				
1. Алгоритмы управления технологическим процессом. Технологические параметры металлургических процессов как источник данных для машинного обучения								
1.1 Машина непрерывного литья заготовок, прокатный стан, литейно-прокатный агрегат как объект АСУ ТП	2	2						ПК-9.1, ПК-9.2, ПК-9.3
1.2 Принципы построения функциональных схем автоматизации ТП		2						ПК-9.1, ПК-9.2, ПК-9.3
1.3 Анализ принципиальной схемы АСУ ТП машины непрерывной разливки стали				4			Текущий контроль. Оформляется отчет. Кластеризация	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
1.4 Анализ принципиальной схемы АСУ ТП прокатного стана				4			Текущий контроль. Текущий контроль. Оформляется отчет. Кластеризация	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
Итого по разделу		4		8				
2. Датчики и исполнительные механизмы. Средства автоматизации измерения физических величин								
2.1 Классификация технических средств автоматики. Сенсоры. Датчики	2	2						ПК-9.1, ПК-9.2, ПК-9.3
2.2 Исполнительные устройства. Регуляторы. Контроллеры		2						ПК-9.1, ПК-9.2, ПК-9.3

2.3 Изучение принципа сбора данных с датчиков скорости (энкодеров), месдоз, установленных на прокатном стане				4			Текущий контроль. Оформляется отчет. Дерево решений	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
2.4 Изучение принципа работы контроллера управляющего прокатным станом				4			Текущий контроль. Оформляется отчет. Классификация	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
Итого по разделу		4		8				
3. Анализ, систематизация и хранение данных технологических процессов и контроля качества готовой продукции								
3.1 Виды и источники данных, принципы разделения и объединения данных, виды шкал. Инструменты первичной обработки данных, сортировки и фильтрации данных. Методы очистки данных и заполнения пропусков, контроля диапазонов. Методы сглаживания и нормировки данных. Преобразование данных	2	2						ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
3.2 Современные подходы к обработке больших данных. Визуализация данных		2						ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
3.3 Организация хранения и доступа к данным. Виды баз данных		2						
3.4 Первичная обработка данных полученных с прокатного стана				4			Текущий контроль. Оформляется отчет. Дерево решений	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
3.5 Сортировка и фильтрация данных полученных с прокатного стана				4			Текущий контроль. Оформляется отчет. Бустинг	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
Итого по разделу		6		8				
4. Формирование наборов данных для машинного обучения								
4.1 Подготовка набора данных для машинного обучения на примере данных полученных с прокатного стана.	2	2						ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3

4.2 Формирование базы данных полученных с прокатного стана и работа с ней				4				ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
4.3 Анализ данных полученных с прокатного стана и подготовка их для создания модели машинного обучения				4				ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
4.4 Самостоятельная работа с курсовой					60	Основная литература 1		ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
4.5 Промежуточная аттестация					36	Подготовка к зачету. Основная литература 1-2. Дополнительная литература 1-2	зачет, курсовая работа	ПК-9.1, ПК-9.2, ПК-9.3, ПК-10.1, ПК-10.2, ПК-10.3
Итого по разделу		2		8	96			
Итого за семестр		16		32	96		зачёт,кр	
Итого по дисциплине		16		32	96		курсовая работа, зачет	

5 Образовательные технологии

Для реализации предусмотренных видов учебной работы в качестве образовательных технологий в преподавании дисциплины «Формирование наборов данных в металлургии» используются традиционная и модульно - компетентностная технологии.

Для изучения дисциплины «Формирование наборов данных в металлургии» предусмотрены практические занятия в компьютерном классе. В рамках интерактивного обучения применяется IT-методы (использование сетевых мультимедийных учебников разработчиков программного обеспечения, электронных образовательных ресурсов по данной дисциплине, в том числе и ЭОР кафедры); метод обучения в сотрудничестве – прохождение всех этапов и методов работы с ЭВМ; проблемное обучение; индивидуальное обучение.

Реализация компетентностного подхода предусматривает использование в учебном процессе активных и интерактивных форм проведения занятий в сочетании с внеаудиторной работой с целью формирования и развития профессиональных навыков обучающихся.

При проведении учебных занятий обеспечивается развитие у обучающихся навыков командной работы, межличностной коммуникации, принятия решений, лидерских качеств.

Используются следующие виды и формы занятий с использованием традиционных и инновационных технологий:

Практическое занятие, посвященное освоению конкретных умений и навыков по предложенному алгоритму.

Технологии проблемного обучения – организация образовательного процесса, которая предполагает постановку проблемных вопросов, создание учебных проблемных ситуаций для стимулирования активной познавательной деятельности студентов.

Практическое занятие в форме практикума – организация учебной работы, направленная на решение комплексной учебно-познавательной задачи, требующей от студента применения как научно-теоретических знаний, так и практических навыков.

Технологии проектного обучения – организация образовательного процесса в соответствии с алгоритмом поэтапного решения проблемной задачи или выполнения учебного задания. Проект предполагает совместную учебно-познавательную деятельность группы студентов, направленную на выработку концепции, установление целей и задач, формулировку ожидаемых результатов, определение принципов и методик решения поставленных задач, планирование хода работы, поиск доступных и оптимальных ресурсов, поэтапную реализацию плана работы, презентацию результатов работы, их осмысление и рефлексию.

Интерактивные технологии – организация образовательного процесса, которая предполагает активное и нелинейное взаимодействие всех участников, достижение на этой основе личностно значимого для них образовательного результата. Наряду со специализированными технологиями такого рода принцип интерактивности прослеживается в большинстве современных образовательных технологий. Интерактивность подразумевает субъект-субъектные отношения в ходе образовательного процесса и, как следствие, формирование саморазвивающейся информационно-ресурсной среды.

6 Учебно-методическое обеспечение самостоятельной работы обучающихся

Представлено в приложении 1.

7 Оценочные средства для проведения промежуточной аттестации

Представлены в приложении 2.

8 Учебно-методическое и информационное обеспечение дисциплины (модуля)

а) Основная литература:

1. Галушкин, А.И. Нейронные сети: основы теории [Электронный ресурс] / А.И. Галушкин. — Электрон. дан. — Москва : Горячая линия-Телеком, 2017. — 496 с. — Режим доступа: https://e.lanbook.com/book/111043. — Загл. с экрана. (13.03.2019).

2. Паттерсон, Д. Глубокое обучение с точки зрения практика / Д. Паттерсон, А. Гибсон. — Москва : ДМК Пресс, 2018. — 418 с. — ISBN 978-5-97060-481-6. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/116122 (дата обращения: 09.10.2021). — Режим доступа: для авториз. пользователей.

б) Дополнительная литература:

1. Ростовцев, В. С. Искусственные нейронные сети : учебник / В. С. Ростовцев. — Санкт-Петербург : Лань, 2019. — 216 с. — ISBN 978-5-8114-3768-9. — Текст : электронный // Лань : электронно-библиотечная система. — URL: https://e.lanbook.com/book/122180.

2. Антонио, Д. Библиотека Keras – инструмент глубокого обучения. Реализация нейронных сетей с помощью библиотек Theano и TensorFlow / Д. Антонио, П.Суджит; перевод с английского А. А. Слинкин. — Москва : ДМК Пресс, 2018. — 294 с. — ISBN 978-5-97060-573-8. — Текст : электронный // Лань: электронно-библиотечная система. — URL: https://e.lanbook.com/book/111438 (дата обращения: 09.10.2021). — Режим доступа: для авториз. пользователей.

в) Методические указания:

г) Программное обеспечение и Интернет-ресурсы:

Программное обеспечение

Наименование ПО	№ договора	Срок действия лицензии
MS Office 2007 Professional	№ 135 от 17.09.2007	бессрочно
7Zip	свободно распространяемое ПО	бессрочно
Kaspersky Endpoint Security для бизнеса-Стандартный	Д-162-21 от 26.03.2021	26.03.2023
MS Visual Studio Code	свободно распространяемое ПО	бессрочно
MS Visual Studio 2017 Community Edition	свободно распространяемое ПО	бессрочно
Браузер Mozilla Firefox	свободно распространяемое ПО	бессрочно
Браузер Yandex	свободно распространяемое ПО	бессрочно

Профессиональные базы данных и информационные справочные системы

Название курса	Ссылка
Электронная база периодических изданий East View Information Services, ООО «ИВИС»	https://dlib.eastview.com/

Национальная информационно-аналитическая система – Российский индекс	URL: https://elibrary.ru/project_risc.asp
Поисковая система Академия Google (Google Scholar)	URL: https://scholar.google.ru/
Информационная система - Единое окно доступа к информационным	URL: http://window.edu.ru/
Российская Государственная библиотека. Каталоги	https://www.rsl.ru/ru/4readers/catalogues/
Электронные ресурсы библиотеки МГТУ им. Г.И. Носова	https://magtu.informsystema.ru/Marc.html?locale=ru

9 Материально-техническое обеспечение дисциплины (модуля)

Материально-техническое обеспечение дисциплины включает:

Основное оборудование, стенды, макеты, компьютерная техника, предустановленное программное обеспечение, используемое для различных видов занятий:

- Самостоятельная работа студента: персональные компьютеры с пакетом MS Office, выходом в Интернет и с доступом в электронную информационно- образовательную среду университета. доска, проектор, компьютер, стеллажи для хранения учебно-наглядных пособий и учебно-методической документации;

- Лекция: персональные компьютеры с пакетом MS Office, выходом в Интернет и с доступом в электронную информационно- образовательную среду университета. доска, проектор, компьютер, стеллажи для хранения учебно-наглядных пособий и учебно-методической документации;

Зачет, диф. зачет: персональные компьютеры с пакетом MS Office, выходом в Интернет и с доступом в электронную информационно- образовательную среду университета. доска, проектор, компьютер, стеллажи для хранения учебно-наглядных пособий и учебно-методической документации;

-Практические занятия и семинары: персональные компьютеры с пакетом MS Office, выходом в Интернет и с доступом в электронную информационно- образовательную среду университета. доска, проектор, компьютер, стеллажи для хранения учебно-наглядных пособий и учебно-методической документации.

Приложение 1

6 Учебно-методическое обеспечение самостоятельной работы обучающихся

Самостоятельная работа с учебными материалами, разбор тем, изученных на лекциях и практических занятиях, разбор решенных заданий.

Обучающиеся изучают источники из списка основной и дополнительной литературы. Предполагается самостоятельное изучение ресурсов по предметной области курса в сети интернет.

Методические рекомендации по подготовке к практическим занятиям представлены в приложении к рабочей программе дисциплины.

Выполнение заданий в рамках портфолио. Обучающиеся решают практические задачи, входящие в портфолио. Методические рекомендации по выполнению домашнего задания представлены в приложении к рабочей программе дисциплины.

Выполнение и защита итогового задания. Обучающиеся выбирают темы итогового задания – самостоятельно (обязательно согласование с преподавателем) или из списка предложенных тем. Используют полученные знания для разведочного анализа данных, проводят выбор способа предобработки данных, выбор способа решения поставленной задачи, проводят оптимизацию гиперпараметров. Методические рекомендации по самостоятельному изучению теоретического материала представлены в приложении к рабочей программе дисциплины.

Приложение 2

7 Оценочные средства для проведения промежуточной аттестации

а) Планируемые результаты обучения и оценочные средства для проведения промежуточной аттестации:

Структурный элемент компетенции	Планируемые результаты обучения	Оценочные средства
ПК-8 Способен адаптировать и применять методы и алгоритмы машинного обучения для решения прикладных задач в различных предметных областях		
ПК-8.1	ставит задачи по адаптации или совершенствованию методов и алгоритмов для решения комплекса задач предметной области: знает: классы методов и алгоритмов машинного обучения; классы методов и алгоритмов машинного обучения; умеет: ставить задачи и адаптировать методы и алгоритмы машинного обучения; ставить задачи и адаптировать методы и алгоритмы машинного обучения; имеет практический опыт: участия в проектах по изучению опыта адаптации и применимости методов и алгоритмов машинного обучения для решения прикладных задач в металлургии; постановки задач по адаптации или совершенствованию методов и алгоритмов для решения комплекса задач предметной области	**Вопросы для зачета:** 1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач. 2. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения. 3. Метрики качества алгоритм регрессии и классификации. 4. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out. 5. Деревья решений. Методы построения деревьев. Их регуляризация. 6. Композиции алгоритмов. Разложение ошибки на смещение и разброс. 7. Случайный лес, его особенности. 8. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов 9. Нейронные сети. Метод обратного распространения ошибок. Свёрточные сети. 10. Кластеризация. Алгоритм K-Means. **Перечень примерных практических заданий:** 1. Кластеризация. По представленному набору данных студенты производят разбиение множества на группы 2. Кластеризация. По представленному набору данных

		студенты производят разбиение множества на группы оптимизируя метрики 3. Классификация. Студенты разрабатывают рекомендательную систему. В качестве набора данных используется набор «МНЛЗ» 4. Классификация. Студенты разрабатывают рекомендательную систему. В качестве набора данных используется набор «МНЛЗ» 5. Деревья решений. Студенты сравнивают эффективности деревьев решений на представленных наборах данных 6. Деревья решений. Студенты сравнивают эффективности деревьев решений на представленных наборах данных 7. Бустинг. Разработка системы согласования экспертных оценок по представленному набору данных. 8. Бустинг. Разработка системы согласования экспертных оценок по представленному набору данных.

б) Порядок проведения промежуточной аттестации, показатели и критерии оценивания:

Контроль качества освоения образовательной программы осуществляется в соответствии с Положением о балльно-рейтинговой системе оценивания результатов учебной деятельности обучающихся.

Критерии оценки: правильность отчета - все верно 5 баллов, есть незначительные ошибки - 4 балла, есть грубые ошибки , но логика расчета верна – 3 балла, расчет сдан но не верен 2 балла; Оформление - все таблицы, рисунки и расчет выполнены в соответствии с ГОСТ 7.32-2017 - 5 баллов, за каждую ошибку снимается по 0,5 баллов. Сдача в срок: в семестре - 5 баллов, на сессии- 3 балла, работа сдана позже - 1 балл; Защита работы: защита работы – это объяснение процесса расчета, выполненного в домашних условиях. - ответил на 5 вопросов преподаватели - 5 баллов, за каждый неправильный ответ минус один балл.

Курсовые работы: при оценивании результатов мероприятия используется балльно-рейтинговая система оценивания результатов учебной деятельности обучающихся. Менее 12 баллов, неудовлетворительно, 12-15 - удовлетворительно, 16-17 - хорошо, 18-20 – отлично.

Зачет: на зачете происходит оценивание учебной деятельности обучающихся по дисциплине на основе полученных оценок за контрольно-рейтинговые мероприятия текущего контроля. Если сумма набранных баллов за мероприятия текущего контроля больше 60, то выставляется зачет. Если баллов недостаточно проводится письменный опрос. Студент получает случайный билет с двумя вопросами. Подготавливает письменный ответ по билету. Время подготовки 30 минут. В случае необходимости устное обсуждение ответов. При оценивании результатов мероприятия используется баллнорейтинговая система оценивания результатов учебной деятельности обучающихся. Правильный ответ на вопрос соответствует 20 баллам. Правильный ответ с небольшими ошибками соответствует 15 баллам. Правильный ответ с грубыми ошибками соответствует 5 баллам. Неправильный ответ на вопрос соответствует 0 Максимальное количество баллов – 40.

Приложение 3

Методические рекомендации по выполнению заданий

Введение в машинное обучение

1.1 Введение

Благодаря машинному обучению программист не обязан писать инструкции, учитывающие все возможные проблемы и содержащие все решения. Вместо этого в компьютер (или отдельную программу) закладывают алгоритм самостоятельного нахождения решений путём комплексного использования статистических данных, из которых выводятся закономерности и на основе которых делаются прогнозы.

Технология машинного обучения на основе анализа данных берёт начало в 1950 году, когда начали разрабатывать первые программы для игры в шашки. За прошедшие десятилетий общий принцип не изменился. Зато благодаря взрывному росту вычислительных мощностей компьютеров многократно усложнились закономерности и прогнозы, создаваемые ими, и расширился круг проблем и задач, решаемых с использованием машинного обучения.

Чтобы запустить процесс машинного обучение, для начала необходимо загрузить в компьютер Дата сет(некоторое количество исходных данных), на которых алгоритм будет учиться обрабатывать запросы. Например, могут быть фотографии собак и котов, на которых уже есть метки, обозначающие к кому они относятся. После процесса обучения, программа уже сама сможет распознавать собак и котов на новых изображениях без содержания меток. Процесс обучения продолжается и после выданных прогнозов, чем больше данных мы проанализировали программой, тем более точно она распознает нужные изображения.

Благодаря машинному обучению компьютеры учатся распознавать на фотографиях и рисунках не только лица, но и пейзажи, предметы, текст и цифры. Что касается текста, то и здесь не обойтись без машинного обучения: функция проверки грамматики сейчас присутствует в любом текстовом редакторе и даже в телефонах. Причем учитывается не только написание слов, но и контекст, оттенки смысла и другие тонкие лингвистические аспекты. Более того, уже существует программное обеспечение, способное без участия человека писать новостные статьи (на тему экономики и, к примеру, спорта).

1.2 Типы задач машинного обучения

Все задачи, решаемые с помощью ML, относятся к одной из следующих категорий.

1) Задача регрессии – прогноз на основе выборки объектов с различными признаками. На выходе должно получиться вещественное число (2, 35, 76.454 и др.), к примеру цена квартиры, стоимость ценной бумаги по прошествии полугода, ожидаемый доход магазина на следующий месяц, качество вина при слепом тестировании.

2) Задача классификации – получение категориального ответа на основе набора признаков. Имеет конечное количество ответов (как правило, в формате «да» или «нет»): есть ли на фотографии кот, является ли изображение человеческим лицом, болен ли пациент раком.

3) Задача кластеризации – распределение данных на группы: разделение всех клиентов мобильного оператора по уровню платёжеспособности, отнесение космических объектов к той или иной категории (планета, звёзда, чёрная дыра и т. п.).

4) Задача уменьшения размерности – сведение большого числа признаков к меньшему (обычно 2–3) для удобства их последующей визуализации (например, сжатие данных).

5) Задача выявления аномалий – отделение аномалий от стандартных случаев. На первый взгляд она совпадает с задачей классификации, но есть одно существенное отличие:

аномалии – явление редкое, и обучающих примеров, на которых можно натаскать машинно обучающуюся модель на выявление таких объектов, либо исчезающе мало, либо просто нет, поэтому методы классификации здесь не работают. На практике такой задачей является, например, выявление мошеннических действий с банковскими картами.

1.3 Основные виды машинного обучения

Основная масса задач, решаемых при помощи методов машинного обучения, относится к двум разным видам: обучение с учителем (supervised learning) либо без него (unsupervised learning). Однако этим учителем вовсе не обязательно является сам программист, который стоит над компьютером и контролирует каждое действие в программе. «Учитель» в терминах машинного обучения – это само вмешательство человека в процесс обработки информации. В обоих видах обучения машине предоставляются исходные данные, которые ей предстоит проанализировать и найти закономерности. Различие лишь в том, что при обучении с учителем есть ряд гипотез, которые необходимо опровергнуть или подтвердить. Эту разницу легко понять на примерах.

Машинное обучение с учителем

Предположим, в нашем распоряжении оказались сведения о десяти тысячах московских квартир: площадь, этаж, район, наличие или отсутствие парковки у дома, расстояние от метро, цена квартиры и т. п. Нам необходимо создать модель, предсказывающую рыночную стоимость квартиры по её параметрам. Это идеальный пример машинного обучения с учителем: у нас есть исходные данные (количество квартир и их свойства, которые называются признаками) и готовый ответ по каждой из квартир – её стоимость. Программе предстоит решить задачу регрессии.

Ещё пример из практики: подтвердить или опровергнуть наличие рака у пациента, зная все его медицинские показатели. Выяснить, является ли входящее письмо спамом, проанализировав его текст. Это всё задачи на классификацию.

Машинное обучение без учителя

В случае обучения без учителя, когда готовых «правильных ответов» системе не предоставлено, всё обстоит ещё интереснее. Например, у нас есть информация о весе и росте какого-то количества людей, и эти данные нужно распределить по трём группам, для каждой из которых предстоит пошить рубашки подходящих размеров. Это задача кластеризации. В этом случае предстоит разделить все данные на 3 кластера (но, как правило, такого строгого и единственно возможного деления нет).

Если взять другую ситуацию, когда каждый из объектов в выборке обладает сотней различных признаков, то основной трудностью будет графическое отображение такой выборки. Поэтому количество признаков уменьшают до двух или трёх, и становится возможным визуализировать их на плоскости или в 3D. Это – задача уменьшения размерности.

1.4 Основные алгоритмы моделей машинного обучения

1.4.1. Дерево принятия решений

Это метод поддержки принятия решений, основанный на использовании древовидного графа: модели принятия решений, которая учитывает их потенциальные последствия (с расчётом вероятности наступления того или иного события), эффективность, ресурсозатратность.

Для бизнес-процессов это дерево складывается из минимального числа вопросов,

предполагающих однозначный ответ — «да» или «нет». Последовательно дав ответы на все эти вопросы, мы приходим к правильному выбору. Методологические преимущества дерева принятия решений – в том, что оно структурирует и систематизирует проблему, а итоговое решение принимается на основе логических выводов.

1.4.2. Наивная байесовская классификация

Наивные байесовские классификаторы относятся к семейству простых вероятностных классификаторов и берут начало из теоремы Байеса, которая применительно к данному случаю рассматривает функции как независимые (это называется строгим, или наивным, предположением). На практике используется в следующих областях машинного обучения:

- определение спама, приходящего на электронную почту;
- автоматическая привязка новостных статей к тематическим рубрикам;
- выявление эмоциональной окраски текста;
- распознавание лиц и других паттернов на изображениях.

1.4.3. Метод наименьших квадратов

Всем, кто хоть немного изучал статистику, знакомо понятие линейной регрессии. К вариантам её реализации относятся и наименьшие квадраты. Обычно с помощью линейной регрессии решают задачи по подгонке прямой, которая проходит через множество точек. Вот как это делается с помощью метода наименьших квадратов: провести прямую, измерить расстояние от неё до каждой из точек (точки и линию соединяют вертикальными отрезками), получившуюся сумму перенести наверх. В результате та кривая, в которой сумма расстояний будет наименьшей, и есть искомая (эта линия пройдёт через точки с нормально распределённым отклонением от истинного значения).

Линейная функция обычно используется при подборе данных для машинного обучения, а метод наименьших квадратов – для сведения к минимуму погрешностей путем создания метрики ошибок.

1.4.4. Логистическая регрессия

Логистическая регрессия – это способ определения зависимости между переменными, одна из которых категориально зависима, а другие независимы. Для этого применяется логистическая функция (аккумулятивное логистическое распределение). Практическое значение логистической регрессии заключается в том, что она является мощным статистическим методом предсказания событий, который включает в себя одну или несколько независимых переменных. Это востребовано в следующих ситуациях:

- кредитный скоринг;
- замеры успешности проводимых рекламных кампаний;
- прогноз прибыли с определённого товара;
- оценка вероятности землетрясения в конкретную дату.

1.4.5. Метод опорных векторов (SVM)

Это целый набор алгоритмов, необходимых для решения задач на классификацию и регрессионный анализ. Исходя из того, что объект, находящийся в N-мерном пространстве, относится к одному из двух классов, метод опорных векторов строит гиперплоскость с мерностью (N – 1), чтобы все объекты оказались в одной из двух групп. На бумаге это можно изобразить так: есть точки двух разных видов, и их можно линейно разделить. Кроме сепарации точек, данный метод генерирует гиперплоскость таким образом, чтобы она была максимально удалена от самой близкой точки каждой группы.

SVM и его модификации помогают решать такие сложные задачи машинного обучения, как сплайсинг ДНК, определение пола человека по фотографии, вывод рекламных баннеров на сайты.

1.4.6. Метод ансамблей

Он базируется на алгоритмах машинного обучения, генерирующих множество классификаторов и разделяющих все объекты из вновь поступающих данных на основе их усреднения или итогов голосования. Изначально метод ансамблей был частным случаем байесовского усреднения, но затем усложнился и оброс дополнительными алгоритмами:

• бустинг (boosting) – преобразует слабые модели в сильные посредством формирования ансамбля классификаторов (с математической точки зрения это является улучшающим пересечением);

• бэггинг (bagging) – собирает усложнённые классификаторы, при этом параллельно обучая базовые (улучшающее объединение);

• корректирование ошибок выходного кодирования.

Метод ансамблей – более мощный инструмент по сравнению с отдельно стоящими моделями прогнозирования, поскольку:

• он сводит к минимуму влияние случайностей, усредняя ошибки каждого базового классификатора;

• уменьшает дисперсию, поскольку несколько разных моделей, исходящих из разных гипотез, имеют больше шансов прийти к правильному результату, чем одна отдельно взятая;

• исключает выход за рамки множества: если агрегированная гипотеза оказывается вне множества базовых гипотез, то на этапе формирования комбинированной гипотезы оно расширяется при помощи того или иного способа, и гипотеза уже входит в него.

1.4.7. Алгоритмы кластеризации

Кластеризация заключается в распределении множества объектов по категориям так, чтобы в каждой категории – кластере – оказались наиболее схожие между собой элементы.

Кластеризировать объекты можно по разным алгоритмам. Чаще всего используют следующие:

• на основе центра тяжести треугольника;

• на базе подключения;

• сокращения размерности;

• плотности (основанные на пространственной кластеризации);

• вероятностные;

• машинное обучение, в том числе нейронные сети.

Алгоритмы кластеризации используются в биологии (исследование взаимодействия генов в геноме, насчитывающем до нескольких тысяч элементов), социологии (обработка результатов социологических исследований методом Уорда, на выходе дающим кластеры с минимальной дисперсией и примерно одинакового размера) и информационных технологиях.

1.4.8. Метод главных компонент (PCA)

Метод главных компонент, или PCA, представляет собой статистическую операцию по ортогональному преобразованию, которая имеет своей целью перевод наблюдений за переменными, которые могут быть как-то взаимосвязаны между собой, в набор главных компонент – значений, которые линейно не коррелированы.

Практические задачи, в которых применяется PCA, – визуализация и большинство процедур сжатия, упрощения, минимизации данных для того, чтобы облегчить процесс обучения. Однако метод главных компонент не годится для ситуаций, когда исходные данные слабо упорядочены (то есть все компоненты метода характеризуются высокой дисперсией). Так что его применимость определяется тем, насколько хорошо изучена и описана предметная область.

1.4.9. Сингулярное разложение

В линейной алгебре сингулярное разложение, или SVD, определяется как разложение прямоугольной матрицы, состоящей из комплексных или вещественных чисел. Так, матрицу M размерностью [m*n] можно разложить таким образом, что $M = U\Sigma V$, где U и V будут унитарными матрицами, а Σ – диагональной.

Одним из частных случаев сингулярного разложения является метод главных компонент. Самые первые технологии компьютерного зрения разрабатывались на основе SVD и PCA и работали следующим образом: вначале лица (или другие паттерны, которые предстояло найти) представляли в виде суммы базисных компонент, затем уменьшали их размерность, после чего производили их сопоставление с изображениями из выборки.

Современные алгоритмы сингулярного разложения в машинном обучении, конечно, значительно сложнее и изощрённее, чем их предшественники, но суть их в целом нем изменилась.

1.4.10. Анализ независимых компонент (ICA)

Это один из статистических методов, который выявляет скрытые факторы, оказывающие влияние на случайные величины, сигналы и пр. ICA формирует порождающую модель для баз многофакторных данных. Переменные в модели содержат некоторые скрытые переменные, причем нет никакой информации о правилах их смешивания. Эти скрытые переменные являются независимыми компонентами выборки и считаются негауссовскими сигналами.

В отличие от анализа главных компонент, который связан с данным методом, анализ независимых компонент более эффективен, особенно в тех случаях, когда классические подходы оказываются бессильны. Он обнаруживает скрытые причины явлений и благодаря этому нашёл широкое применение в самых различных областях – от астрономии и медицины до распознавания речи, автоматического тестирования и анализа динамики финансовых показателей.

1.5 Примеры применения в реальной жизни

Пример 1. Диагностика заболеваний

Пациенты в данном случае являются объектами, а признаками – все наблюдающиеся у них симптомы, анамнез, результаты анализов, уже предпринятые лечебные меры (фактически вся история болезни, формализованная и разбитая на отдельные критерии). Некоторые признаки – пол, наличие или отсутствие головной боли, кашля, сыпи и иные – рассматриваются как бинарные. Оценка тяжести состояния (крайне тяжёлое, средней тяжести и др.) является порядковым признаком, а многие другие – количественными: объём лекарственного препарата, уровень гемоглобина в крови, показатели артериального давления и пульса, возраст, вес. Собрав информацию о состоянии пациента, содержащую много таких признаков, можно загрузить её в компьютер и с помощью программы, способной к машинному обучению, решить следующие задачи:

- провести дифференциальную диагностику (определение вида заболевания);
- выбрать наиболее оптимальную стратегию лечения;
- спрогнозировать развитие болезни, её длительность и исход;
- просчитать риск возможных осложнений;
- выявить синдромы – наборы симптомов, сопутствующие данному заболеванию или нарушению.

Ни один врач не способен обработать весь массив информации по каждому пациенту мгновенно, обобщить большое количество других подобных историй болезни и сразу же выдать чёткий результат. Поэтому машинное обучение становится для врачей незаменимым помощником.

Пример 2. Поиск мест залегания полезных ископаемых

В роли признаков здесь выступают сведения, добытые при помощи геологической разведки: наличие на территории местности каких-либо пород (и это будет признаком бинарного типа), их физические и химические свойства (которые раскладываются на ряд количественных и качественных признаков).

Для обучающей выборки берутся 2 вида прецедентов: районы, где точно присутствуют месторождения полезных ископаемых, и районы с похожими характеристиками, где эти ископаемые не были обнаружены. Но добыча редких полезных ископаемых имеет свою специфику: во многих случаях количество признаков значительно превышает число объектов, и методы традиционной статистики плохо подходят для таких ситуаций. Поэтому при машинном обучении акцент делается на обнаружение закономерностей в уже собранном массиве данных. Для этого определяются небольшие и наиболее информативные совокупности признаков, которые максимально показательны для ответа на вопрос исследования – есть в указанной местности то или иное ископаемое

или нет. Можно провести аналогию с медициной: у месторождений тоже можно выявить свои синдромы. Ценность применения машинного обучения в этой области заключается в том, что полученные результаты не только носят практический характер, но и представляют серьёзный научный интерес для геологов и геофизиков.

Пример 3. Оценка надёжности и платёжеспособности кандидатов на получение кредитов

С этой задачей ежедневно сталкиваются все банки, занимающиеся выдачей кредитов. Необходимость в автоматизации этого процесса назрела давно, ещё в 1960–1970-е годы, когда в США и других странах начался бум кредитных карт.

Лица, запрашивающие у банка заём, – это объекты, а вот признаки будут отличаться в зависимости от того, физическое это лицо или юридическое. Признаковое описание частного лица, претендующего на кредит, формируется на основе данных анкеты, которую оно заполняет. Затем анкета дополняется некоторыми другими сведениями о потенциальном клиенте, которые банк получает по своим каналам. Часть из них относятся к бинарным признакам (пол, наличие телефонного номера), другие — к порядковым (образование, должность), большинство же являются количественными (величина займа, общая сумма задолженностей по другим банкам, возраст, количество членов семьи, доход, трудовой стаж) или номинальными (имя, название фирмы-работодателя, профессия, адрес).

Для машинного обучения составляется выборка, в которую входят кредитополучатели, чья кредитная история известна. Все заёмщики делятся на классы, в простейшем случае их 2 – «хорошие» заёмщики и «плохие», и положительное решение о выдаче кредита принимается только в пользу «хороших».

Более сложный алгоритм машинного обучения, называемый кредитным скорингом, предусматривает начисление каждому заёмщику условных баллов за каждый признак, и решение о предоставлении кредита будет зависеть от суммы набранных баллов. Во время машинного обучения системы кредитного скоринга вначале назначают некоторое количество баллов каждому признаку, а затем определяют условия выдачи займа (срок, процентную ставку и остальные параметры, которые отражаются в кредитном договоре). Но существует также и другой алгоритм обучения системы – на основе прецедентов.