Федеральное государственное автономное образовательное учреждение высшего образования
«Московский физико-технический институт (национальный исследовательский университет)»

УТВЕРЖДЕНО
Директор физтех-школы аэрокосмических технологий
С.С. Негодяев

Рабочая программа дисциплины (модуля)

по дисциплине: Машинное обучение в науках о Земле
по направлению: Прикладные математика и физика
профиль подготовки: Геокосмические науки и технологии
Физтех-школа Аэрокосмических Технологий
кафедра термогидромеханики океана
курс: 4
квалификация: бакалавр

Семестры, формы промежуточной аттестации:
7 (осенний) - Зачет
8 (весенний) - Дифференцированный зачет

Аудиторных часов: 120 всего, в том числе:
лекции: 60 час.
семинары: 60 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 60 час.

Всего часов: 180, всего зач. ед.: 4

Программу составил: М.А. Криницкий, канд. техн. наук

Программа обсуждена на заседании кафедры термогидромеханики океана 01.06.2020

Аннотация

Машинное обучение – область науки, объединяющая теорию вероятностей, математическую статистику, инженерный подход к решению сложных задач, подходы обработки данных, искусство программирования и многие другие базовые дисциплины. За последние несколько лет многие задачи наук о Земле были решены с применением машинного обучения намного более успешно, нежели с применением классических методов до этого. С другой стороны, в фундаментальных науках практика применения машинного обучения сталкивается со вполне объяснимым академическим скепсисом, связанным с пониженной интерпретируемостью получаемых результатов. В предлагаемом курсе рассматриваются вопросы применимости методов машинного обучения в различных исследовательских и инженерных задачах наук о Земле, а также вопросы интерпретации, достоверности и неопределенностей результатов.

В настоящем курсе рассматриваются базовые понятия машинного обучения, рассматриваются наиболее часто применяемые алгоритмы классов «обучение с учителем» и «обучение без учителя». Особое внимание уделяется оценке достоверности получаемых решений и оценке неопределенностей как параметров обучаемых моделей, так и аппроксимаций целевых переменных. Изучение методов класса «обучение с учителем» начинается с одной из наиболее изученных моделей – линейной регрессии. На основе этой модели изучаются все основные свойства, присущие большинству методов, и подходы повышения точности и достоверности решений: обобщающая способность, переобучение и недообучение, подходы порождения и отбора признаков, подходы оценки качества моделей в задачах регрессии и классификации. Особое место в курсе занимает комплекс вопросов, посвященных искусственным нейронным сетям и их применению в задачах обработки и анализа данных различной природы. Изучение методов класса «обучение без учителя» включает в себя как классические (метод главных компонент, статистические методы кластеризации), так и современные методы поиска и анализа структуры данных (нейросетевые автокодировщики и порождающие модели).

Курс формирует современное понимание научной деятельности как таковой, включая постановку и проверку гипотез, постановку задач анализа данных. Курс также будет полезен при разработке новых способов натурных наблюдений, при проведении измерений и обработке результатов эксперимента в науках о Земле.

1. Цели и задачи

Цель дисциплины

- формирование базовых знаний о математических основах и общих принципах современных методов машинного обучения в применении к задачам наук о Земле;
- освоение общепринятых методик применения методов машинного обучения.

Задачи дисциплины

- дать студентам знания об общих математических принципах современных методов машинного обучения;
- научить студентов самостоятельно формулировать задачу, планировать численный эксперимент, выбирать подходящий метод решения и эффективно его реализовывать в виде программы, а также анализировать результаты и оценивать качество получаемых моделей;
- выработать у студентов навыки эффективного применения методов машинного обучения с использованием доступных языков программирования и сред исполнения программного кода;
- выработать у студентов навык адаптации существующих методов машинного обучения с учетом специфики задач и с использованием результатов новейших публикаций;
- выработать у студентов навык визуального представления данных, представления промежуточных и конечных результатов исследования.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи

критическии анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.1 Владеет методами статистической обработки и анализа научных данных
	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
	ПК-2.3 Способен представлять научные утверждения, их обоснования и доказательства, научные проблемы и их решения ясно и точно в терминах, понятных для профессиональной аудитории, в письменной и устной форме

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные понятия и математические принципы современных методов машинного обучения;
- типы задач, решаемые с применением методов машинного обучения;
- наиболее распространенные модели машинного обучения, области их применения, основные преимущества и недостатки;
- основные показатели качества моделей машинного обучения в задачах различного типа, их преимущества, недостатки, ограничения применимости и характерные значения, достигаемые по результатам современных исследований.

уметь:

- определить тип задачи с точки зрения машинного обучения, сформулировать задачу в терминах методов машинного обучения, описать исходные данные и целевые переменные, подобрать подходящий тип модели, сформулировать и обосновать метрики качества получаемого решения;
- реализовывать в виде программы цепь обработки данных и тренировки модели машинного обучения в рамках решения поставленной задачи;
- идентифицировать явления недообучения и переобучения для различных типов задач и для различных конкретных видов моделей машинного обучения, руководствуясь метриками качества решения и диагностическими показателями процесса обучения; принимать меры для купирования эффектов недообучения и переобучения;
- проводить исследование чувствительности модели к значениям гиперпараметров;
- проводить оптимизацию гиперпараметров;
- исследовать исходные данные в аспекте сформулированной задачи;
- оценивать границы применимости и возможные причины смещенности полученного решения.

владеть:

- навыками самостоятельной реализации алгоритмов машинного обучения по материалам современных исследований, изложенных в научных статьях;
- навыками адаптации существующих алгоритмов машинного обучения с учетом особенностей сформулированной задачи;
- навыками оптимизации процессов предобработки исходных данных и постобработки результатов численных экспериментов.

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

4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение	2	2		2
2	Классификация задач и методов машинного обучения	4	4		4
3	Технические средства анализа данных	4	4		4
4	Линейная регрессия и принципы машинного обучения	4	4		4
5	Общая схема решения задач машинного обучения	4	4		4
6	Задачи классификации и логистическая регрессия	4	4		4
7	Оптимизация моделей машинного обучения и настройка гиперпараметров	4	4		4
8	Метод опорных векторов	4	4		4
9	Деревья решений	4	4		4
10	Композиции и ансамбли	4	4		4
11	Искусственные нейронные сети	4	4		5
12	Технические средства конструирования и обучения ИНС	4	4		5
13	Сверточные нейронные сети	4	5		4
14	Рекуррентные нейронные сети	4	5		4
15	Задачи типа «обучение без учителя»	6	4		4
Итого часов		60	60		60
Подготовка к экзамену		0 час.			
Общая трудоёмкость		180 час., 4 зач.ед.			

4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 7 (Осенний)

1. Введение

Машинное обучение и искусственный интеллект. Исторический обзор, начальные определения, взаимосвязь понятий. Обзор языков программирования и инструментов для проведения исследований с применением методом машинного обучения. Обзор современных приложений в академической науке и в технике.

Машинное обучение как способ выявления неявных закономерностей в данных.

Машинное обучение как способ решения задач на основе натурных данных.

Машинное обучение как способ аппроксимации распределения данных.

2. Классификация задач и методов машинного обучения

Типы задач машинного обучения: «обучение с учителем», «обучение без учителя», «обучение с частичным привлечением учителя» и др. Задачи классификации и регрессии. Примеры в задачах наук о Земле.

Краткий обзор методов обучения «с учителем». Линейная регрессия, логистическая регрессия, наивный байесовский классификатор, метод K ближайших соседей, решающие деревья, композиционные методы, метод опорных векторов, искусственные нейронные сети. Примеры в задачах наук о Земле.

Краткий обзор методов обучения «без учителя». Метод главных компонент и другие методы сокращения размерности. Методы кластеризации: метод K средних; метод моделирования гауссовой смеси; агломеративная и дивизивная иерархическая кластеризация; DBSCAN и HDBSCAN. Методы обнаружения аномалий. Нейросетевые автокодировщики. Генеративные состязательные сети. Примеры в задачах наук о Земле.

3. Технические средства анализа данных

Python как язык программирования для анализа данных. Динамическая типизация и полная интроспекция. Парадигмы программирования, реализованные в Python. Особенности многопотоковой обработки данных. Библиотеки обработки данных и библиотеки визуализации данных. Специальные библиотеки для задач наук о Земле Basemap и NetCDF4. Библиотека анализа двумерных данных OpenCV.

Инструментарий для обработки, визуализации и анализа данных с использованием Python.

Интерактивная среда разработки JetBrains PyCharm. Клиент-серверная интерактивная среда исполнения кода и визуализации Jupyter. Особенности исполнения программ в среде Jupyter. Построение документов в среде Jupyter с применением разметки и LaTeX.

4. Линейная регрессия и принципы машинного обучения

Вероятностная постановка задачи обучения по прецедентам. Статистические основы модели линейной регрессии. Варианты решения задачи линейной регрессии. Ограничения линейной регрессии. Проблема мультиколлинеарности признаков. Генерация и отбор признаков, спрямляющие пространства. Однослойный перцептрон.

Принципы машинного обучения в примерах. Принцип близости похожих событий в пространстве представлений. Принцип оптимизации функционала потерь. Регуляризация моделей. Разделимость и отделимость событий в задачах классификации. Интерпретируемость моделей машинного обучения. Принцип композиции алгоритмов.

5. Общая схема решения задач машинного обучения

Определение типа задачи и постановка задачи. Исследование или/и формирование массива исходных данных, визуализация данных. Адаптация алгоритмов машинного обучения и алгоритмов их настройки под сформулированную задачу. Предобработка данных для выбранных алгоритмов. Оптимизация (обучение) моделей. Оценка качества и оптимизация гиперпараметров. Применение модели и построение выводов по результатам.

6. Задачи классификации и логистическая регрессия

Примеры задач классификации в науках о Земле. Статистические основы модели логистической регрессии. Формулировка модели логистической регрессии и логистическая функция ошибки. Обучение модели логистической регрессии. Ограничения логистической регрессии. Проблема мультиколлинеарности признаков, генерация и отбор признаков, спрямляющие пространства. Однослойный перцептрон с произвольной функцией активации. Виды функции активации.

7. Оптимизация моделей машинного обучения и настройка гиперпараметров

Формулировка задачи оптимизации. Примеры оптимизационных задач. Задача выпуклого программирования. Общая задача нелинейного программирования. Понятие ландшафта функции потерь. Проблемы невыпуклого ландшафта функции потерь и методы оптимизации, решающие эту проблему. Градиентные методы оптимизации первого и второго порядка. Модификации градиентных методов оптимизации первого порядка.

Явление переобучения и недообучения моделей. Понятие VC-размерности, сложность модели. Баланс между смещением и разбросом. Настройка гиперпараметров модели. Подход скользящего контроля. Стратегии скользящего контроля.

8. Метод опорных векторов

Линейно разделимая выборка и разделяющая гиперплоскость. Геометрическая интерпретация задачи. Модель линейного метода опорных векторов. Функция потерь метода опорных векторов. Варианты оптимизация моделей в методе опорных векторов. Проблема неразделимости выборки. Ядра и спрямляющие пространства. Примеры ядер. Метод опорных векторов как двуслойный перцептрон. Метод опорных векторов в задачах восстановления регрессии.

Семестр: 8 (Весенний)

9. Деревья решений

Взаимная информация, информационная энтропия, кросс-энтропия, дивергенция Кульбака-Лейблера и принцип оптимизации правдоподобия. Метод наименьших квадратов в задачах линейной регрессии как частный случай принципа максимизации правдоподобия.

Алгоритмы построения решающих деревьев ID3 и C4.5. Решающие деревья в задачах классификации и восстановления регрессии.

10. Композиции и ансамбли

Комитеты и композиции алгоритмов. Бутстрэп и бэггинг. Случайные леса. Градиентные метаалгоритмы. Бустинг над произвольным семейством алгоритмов. AdaBoost. XGBoost, LightGBM и CatBoost в задачах классификации и регрессии.

11. Искусственные нейронные сети

Исторический обзор развития искусственного интеллекта. «Восходящее» и «нисходящее» направления. Коннективизм и принцип ассоциативности.

Перцептрон. Варианты однослойного перцептрона. Способы обучения перцептрона. Искусственная нейронная сеть как универсальный аппроксиматор. Многослойный перцептрон.

Алгоритмы обучения ИНС. Алгоритм обратного распространения ошибки. Ландшафт функции потерь. Сходимость обучения ИНС. Регуляризации и эвристики оптимизации ИНС. Пакетная нормализация. Прореживание.

Переобучение, недообучение и обобщающая способность ИНС.

12. Технические средства конструирования и обучения ИНС

Обзор средств и библиотек для программной реализации искусственных нейронных сетей. Numpy, Keras, Tensorflow, Theano, PyTorch.

Реализация многослойного перцептрона и процедуры оптимизации. Особенности вычислений на графических сопроцессорах. Организация порождения обучающих данных и цепи вычислений процесса оптимизации ИНС.

13. Сверточные нейронные сети

Краткий исторический обзор. Когнитрон и неокогнитрон, LeNet и более поздние архитектуры. Характерные задачи, решаемые СНС.

Принцип локальности признаков. Принцип оценки корреляции с шаблоном. Принцип общих параметров. Математические основы сверточных нейронных сетей. Обратное распространение градиента функции потерь. Рецептивное поле. Субдискретизация. Соединения быстрого доступа.

Свойства СНС.

Виды задач, решаемые с применением ИНС и СНС. Современные архитектуры СНС.

14. Рекуррентные нейронные сети

Краткий исторический обзор. Принцип локальности признаков. Кодирование последовательностей. One-hot, Word2Vec, GloVe, fastText.

Рекуррентные нейронные сети: основные принципы. LSTM, двунаправленный LSTM, GRU.

Типы задач, решаемые РНС. Классификация и регрессия, порождение последовательности на базе последовательности.

15. Задачи типа «обучение без учителя»

Задача сокращения размерности. Метод главных компонент. t-SNE. Самообучающиеся карты Кохонена. Нейросетевой автокодировщик и его разновидности.

Задача кластеризации. Постановка задачи. Мера близости событий и проклятие размерности в задачах кластеризации.

Виды алгоритмов кластеризации. Графовые и эвристические алгоритмы. DBSCAN и HDBSCAN. Статистические алгоритмы. Метод разделения гауссовой смеси, метод K средних. Иерархические алгоритмы. Метод Ланса-Уильямса. Вариации метода Ланса-Уильямса. Свойства иерархических алгоритмов.

Нейросетевые генеративные модели. Принцип и статистические основы генеративных состязательных сетей. DCGAN, LSGAN, WGAN.

5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система), интерактивная доска.

6.Перечень рекомендуемой литературы

Основная литература

1. Машинное обучение [Текст]/Х. Бринк, Дж. Ричардс, М. Феверолф, Real-World Machine Learning, -СПб., Питер, 2017

1. Флах П. «Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных.» / Флах П. М.: ДМК Пресс, 2015. - 400 с.
2. Гудфеллоу Я., Бенджио И., Курвилль А. «Глубокое обучение.» / М.: ДМК Пресс, 2017. - 652 с.
3. Николенко С. И., Кадурин А. А., Архангельская Е. О. «Глубокое обучение.» / СПб.: Питер. 2019. - 480 с.

Дополнительная литература

1. Hastie T., Tibshirani R., Friedman J. «The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition» / T. Hastie, R. Tibshirani, J. Friedman, 2-е изд., New York: Springer-Verlag, 2009.
2. Bishop C. «Pattern Recognition and Machine Learning» / C. Bishop, New York: Springer-Verlag, 2006.

7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

http://lib.mipt.ru/ – электронная библиотека Физтеха

http://benran.ru –библиотека по естественным наукам Российской академии наук.

http://ml4es.ru/ - веб-сайт автора курса, посвященный этому курсу

8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

Интерактивная среда разработки программ PyCharm (производитель JetBrains, версия classroom)
Язык программирования Python 3, клиент-серверная среда исполнения кода Jupyter

9. Методические указания для обучающихся по освоению дисциплины (модуля)

Основной упор курса сделан на приобретение практических навыков применения методов машинного обучения в реальных задачах геофизики при условии понимания математических оснований этих методов. Большинство рассматриваемых тем сопровождаются примерами задач, разбираемыми на семинарах. Кроме того, в процессе изучения курса студентам на основании полученных знаний предлагается выполнить ряд заданий по анализу данных – как синтетических, так и полученных в ходе различных экспедиций и экспериментов. Для закрепления получаемых знаний и навыков студенты получают как теоретические, так и практические домашние задания.

По итогу освоения курса студентам предлагается выполнить курсовую работу. Тему курсовой работы студент выбирает, руководствуясь собственными интересами в геофизике или выбирает из предложенных преподавателем. Для уточнения темы проводится консультация с преподавателем.

Успешное освоение курса возможно лишь при условии большой самостоятельной работы студента.
Самостоятельная работа включает в себя:
– решение задач при выполнении домашних заданий;
– чтение рекомендованной литературы;
– проработку учебного материала (по конспектам, учебной и научной литературе);
– самостоятельное освоение современных методов машинного обучения, не затронутых в курсе, на основе предложенного учебного материала.

Руководство и контроль за самостоятельной работой студента осуществляется анализом итогов домашних заданий, курсовой работы, а также индивидуальных консультаций.
Показателем владения материалом служит умение ставить и решать задачи. Для формирования умения применять знания на практике студенту необходимо решать как можно больше задач.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Прикладные математика и физика
профиль подготовки: Геокосмические науки и технологии
Физтех-школа Аэрокосмических Технологий
кафедра термогидромеханики океана
курс: 4
квалификация: бакалавр

Семестры, формы промежуточной аттестации:
7 (осенний) - Зачет
8 (весенний) - Дифференцированный зачет

Разработчик: М.А. Криницкий, канд. техн. наук

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
УК-1 Способен осуществлять поиск, критический анализ и синтез информации, применять системный подход для решения поставленных задач	УК-1.1 Анализирует задачу, выделяя этапы ее решения, действия по решению задачи
	УК-1.2 Находит, критически анализирует и выбирает информацию, необходимую для решения поставленной задачи
	УК-1.3 Рассматривает различные варианты решения задачи, оценивает их преимущества и недостатки
	УК-1.4 Грамотно, логично, аргументированно формирует собственные суждения и оценки
	УК-1.5 Определяет и оценивает практические последствия возможных вариантов решения задачи
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
	ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области
ПК-2 Способен анализировать полученные в ходе научно-исследовательской работы данные и делать научные выводы (заключения)	ПК-2.1 Владеет методами статистической обработки и анализа научных данных
	ПК-2.2 Умеет находить ключевые параметры, определяющие изучаемое явление, и производить численные оценки по порядку величины
	ПК-2.3 Способен представлять научные утверждения, их обоснования и доказательства, научные проблемы и их решения ясно и точно в терминах, понятных для профессиональной аудитории, в письменной и устной форме

2. Показатели оценивания компетенций

В результате изучения дисциплины «Машинное обучение в науках о Земле» обучающийся должен:

знать:

- основные понятия и математические принципы современных методов машинного обучения;
- типы задач, решаемые с применением методов машинного обучения;
- наиболее распространенные модели машинного обучения, области их применения, основные преимущества и недостатки;
- основные показатели качества моделей машинного обучения в задачах различного типа, их преимущества, недостатки, ограничения применимости и характерные значения, достигаемые по результатам современных исследований.

уметь:

• определить тип задачи с точки зрения машинного обучения, сформулировать задачу в терминах методов машинного обучения, описать исходные данные и целевые переменные, подобрать подходящий тип модели, сформулировать и обосновать метрики качества получаемого решения;
• реализовывать в виде программы цепь обработки данных и тренировки модели машинного обучения в рамках решения поставленной задачи;
• идентифицировать явления недообучения и переобучения для различных типов задач и для различных конкретных видов моделей машинного обучения, руководствуясь метриками качества решения и диагностическими показателями процесса обучения; принимать меры для купирования эффектов недообучения и переобучения;
• проводить исследование чувствительности модели к значениям гиперпараметров;
• проводить оптимизацию гиперпараметров;
• исследовать исходные данные в аспекте сформулированной задачи;
• оценивать границы применимости и возможные причины смещенности полученного решения.

владеть:

- навыками самостоятельной реализации алгоритмов машинного обучения по материалам современных исследований, изложенных в научных статьях;
- навыками адаптации существующих алгоритмов машинного обучения с учетом особенностей сформулированной задачи;
- навыками оптимизации процессов предобработки исходных данных и постобработки результатов численных экспериментов.

3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

Текущий контроль осуществляется в форме домашних заданий по каждой теме.

Каждое домашнее задание оценивается согласно указанному весу. В рамках каждого домашнего задания предлагаются задачи, каждая из которых оценивается согласно указанному количеству баллов. Суммарное количество баллов за каждое домашнее задание - 100. Кроме того, для студентов, желающих более глубоко освоить материал, предлагаются опциональные задачи. Суммарное количество баллов по домашнему заданию с учетом веса этого домашнего задания идет в общий зачет за семестр.

За осенний семестр максимальное количество взвешенных баллов – 500 без учета опциональных заданий. Условие получения зачета – набор не менее 350 взвешенных баллов за семестр.

За весенний семестр максимальное количество взвешенных баллов – 800 без учета опциональных заданий. Условие допуска к дифференцированному зачету – набор не менее 600 взвешенных баллов за семестр.

Домашнее задание по теме «Классификация задач и методов машинного обучения»

вес = 0.4

1. (100 баллов) Определить тип задачи машинного обучения для предложенных проблем.

Домашнее задание по теме «Технические средства анализа данных»

вес = 0.6

1. (20 баллов) Установить комплект программного обеспечения и дополнительных библиотек на домашний/рабочий компьютер. Проверить функциональность интерактивной среды исполнения кода Jupyter;
2. (20 баллов) По предложенной тетрадке-шаблону проверить функциональность библиотек Basemap, NetCDF4 и OpenCV;
3. (40 баллов) Используя тетрадку-шаблон, заполнить недостающие фрагменты кода в предложенных задачах и проверить корректность исполнением.
4. (20 баллов) Оформить отчет о выполненных проверках и заданиях в виде тетрадки (одной или нескольких) Jupyter с использованием ячеек разметки и с применением LaTeX;

Домашнее задание по теме «Линейная регрессия и принципы машинного обучения»

вес = 1.5

1. (50 баллов) Решить задачу оценки балла общей облачности по предложенным данным с использованием модели линейной регрессии. Реализация линейной регрессии может быть написана с использованием библиотеки Numpy. Варианты готовых реализаций модели (напр., из пакета scikit-learn) в этом задании использовать запрещается.
2. (Опционально, +30 баллов) решить задачу классификации рукописных цифр из набора данных MNIST методом регрессии с использованием модели однослойного перцептрона. Варианты готовых реализаций модели (напр., из пакета scikit-learn) в этом задании использовать запрещается.
3. (20 баллов) Используя тетрадку-шаблон, заполнить недостающие участки кода в предложенных задачах визуализации данных.
4. (30 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Задачи классификации и логистическая регрессия»
вес = 1.0
1. (60 баллов) Решить задачу классификации оптических снимков небосвода в отношении состояния диска Солнца с использованием модели логистической регрессии. Реализация логистической регрессии может быть написана с использованием библиотеки Numpy. Библиотеку scikit-learn и другие варианты готовых реализаций модели в этом задании использовать запрещается.
2. (Опционально, +40 баллов) решить задачу оценки балла общей облачности в подходе классификации с использованием модели однослойного перцептрона. Варианты готовых реализаций модели (напр., из пакета scikit-learn) в этом задании использовать запрещается.
3. (40 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Метод опорных векторов»
вес = 1.5
1. (70 баллов) Реализовать самостоятельно метод опорных векторов. Решить с помощью этой модели одну из предложенных задач классификации:
a. оценки балла общей облачности на снимке небосвода
b. определения состояния диска Солнца на снимке небосвода
c. определения рукописных цифр из набора данных MNIST
d. бинарной классификации мезоциклонов Южного полушария
e. классификации состояний стратосферного полярного вихря

Варианты готовых реализаций модели SVM (напр., из пакета scikit-learn) в этом задании использовать запрещается.
2. (Опционально, +40 баллов) Решить задачу классификации рукописных цифр из набора данных MNIST с применением модели SVM. Разрешается использовать SVM с ядром. В этом задании разрешается использовать готовые реализации модели, например, библиотеку scikit-learn.
3. (30 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Деревья решений»
вес = 0.5
1. (40 баллов) Используя тетрадку-шаблон, решить предложенные теоретические задачи. Оформить решение в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.
2. (60 баллов) Решить задачу (на выбор):
a. оценки балла общей облачности на снимке небосвода в подходе классификации
b. определения состояния диска Солнца на снимке небосвода в подходе классификации
c. определения рукописных цифр из набора данных MNIST
d. бинарной классификации мезоциклонов Южного полушария
e. классификации состояний стратосферного полярного вихря

с использованием метода решающих деревьев. Исследовать полученную модель на чувствительность к особенностям данных. Исследовать модель на чувствительность к значениям гиперпараметров. Настроить гиперпараметры.

Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Композиции и ансамбли»

вес = 0.8

1. (65 баллов) Решить задачу (на выбор):

a. оценки балла общей облачности на снимке небосвода в подходе классификации

b. определения состояния диска Солнца на снимке небосвода в подходе классификации

c. определения рукописных цифр из набора данных MNIST

d. бинарной классификации мезоциклонов Южного полушария

e. классификации состояний стратосферного полярного вихря

с использованием одного из композиционных методов на базе решающих деревьев. Исследовать полученную модель на чувствительность к особенностям данных. Исследовать модель на чувствительность к значениям гиперпараметров. Настроить гиперпараметры.

2. (35 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Искусственные нейронные сети»

вес = 2.0

1. (65 баллов) Реализовать самостоятельно многослойный перцептрон (MLP) с обучением методом обратного распространения ошибки и L2-регуляризацией весов. С использованием этой реализации решить одну из предложенных задач классификации или регрессии. Реализация MLP может быть написана с использованием библиотеки Numpy или других низкоуровневых математических библиотек. Варианты готовых реализаций модели и ее компонент (напр., из пакетов keras, tensorflow, pytorch etc.) в этом задании использовать запрещается.

2. (35 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Технические средства конструирования и обучения ИНС»

вес = 0.7

1. (60 баллов) Решить одну из предложенных задач классификации или регрессии с применением MLP. Исследовать модель на чувствительность к особенностям данных. Исследовать модель на чувствительность к значениям гиперпараметров. Оптимизировать гиперпараметры. В этом задании разрешается использовать библиотеки построения и обучения ИНС.

2. (40 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Сверточные нейронные сети»

вес = 1.3

1. (30 баллов) На базе предложенной тетрадки-шаблона выполнить задание, заполнив недостающие участки кода:

a. самостоятельно реализовать операцию дискретной свертки;

b. подобрать ядра свертки, выполняющие операции детектора границ, повышения контрастности изображения и проч. Проверить функциональность на произвольном изображении;

2. (50 баллов) Реализовать сверточную нейронную сеть одной из предложенных архитектур. С применением реализованной сети решить одну из предложенных задач:

a. оценка балла общей облачности по широкоугольным оптическим снимкам видимой полусферы небосвода. Допускается решение в подходе классификации, регрессии, упорядоченной регрессии и проч.

b. определение состояния диска Солнца по широкоугольным оптическим снимкам видимой полусферы небосвода;

c. классификация или детектирование полярных мезомасштабных циклонов по данным спутниковых мозаик в инфракрасном и микроволновом диапазонах;

d. сегментация рукописных цифр в синтетически порожденных изображениях на основе набора данных MNIST.

3. (20 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Рекуррентные нейронные сети»

вес = 0.7

1. (70 баллов) С применением рекуррентных нейронных сетей решить одну из предложенных задач:

a. оценка балла общей облачности по широкоугольным оптическим снимкам видимой полусферы небосвода с учетом временного ряда снимков;

b. определение состояния диска Солнца по широкоугольным оптическим снимкам видимой полусферы небосвода с учетом временного ряда снимков;

c. прогноз потоков приходящей коротковолновой и длинноволновой радиации над океаном с заданной заблаговременностью на основании данных широкоугольных оптических снимков видимой полусферы небосвода с учетом временного ряда снимков.

2. (30 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

Домашнее задание по теме «Задачи типа «обучение без учителя»

вес = 2.0

1. (40 баллов) Используя тетрадку-шаблон, заполнить недостающие фрагменты программного кода для выполнения практических заданий:

a. самостоятельная реализация метода главных компонент на предложенных синтетических данных или реальных экспедиционных данных;

b. визуализация данных в главных компонентах

c. применение метода t-SNE для визуализации данных;

d. применение самообучающихся карт Кохонена для визуализации данных

2. (25 баллов) Реализовать один из вариантов нейросетевого автокодировщика с использованием высокоуровневых библиотек конструирования и обучения ИНС. С применением автокодировщика исследовать один из наборов данных, предложенных на выбор:

a. данные рукописных цифр MNIST;

b. широкоугольные снимки видимой полусферы небосвода;

c. геофизические поля, описывающие состояние стратосферного полярного вихря;

3. (25 баллов) Решить задачу кластеризации (опционально с применением автокодировщика, см. задание 2) в одной из задач, предложенных на выбор:

a. рукописных цифр из набора данных MNIST;

b. кластеризация широкоугольных оптических снимков видимой полусферы небосвода;

c. кластеризация состояний стратосферного полярного вихря;

d. кластеризация крупномасштабных условий циркуляции атмосферы, сопровождающих зарождение полярного мезомасштабного циклона.

4. (10 баллов) Прокомментировать полученные результаты в виде отчета, оформленного с применением ячеек разметки и с использованием LaTeX.

4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

Промежуточная аттестация по дисциплине «Машинное обучение для решения исследовательских и инженерных задач в науках о Земле» проводится в форме зачёта в осеннем семестре и в виде дифференцированного зачёта (устного) в весеннем.

Примеры экзаменационных билетов:

Для практических заданий дифференцированного зачета подготовлены следующие наборы данных:

- широкоугольные снимки видимой полусферы небосвода с заранее рассчитанными статистическими характеристиками, одновременными показаниями наблюдателя относительно визуально оцениваемых метеоролгических характеристик, а также одновременными показаниями радиометров. По данным, собранным автором курса в ряде экспедиций в Атлантическом и Индийском океанах;
- данные геофизических полей, описывающие состояние стратосферного полярного вихря в зимний период. По данным реанализа JRA-55 с 1959г. по н.в.
- данные геофизических полей и спутниковых мозаик (предоставлены AMSRC, USA) Южного полушария с частичной разметкой, касающейся времени, расположения и размеров мезомасштабных циклонов Южного океана;
- публично доступный набор данных рукописных цифр MNIST.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 1

1. Теория: Машинное обучение и искусственный интеллект: основные понятия, определения, основные исторические события. Современные достижения в приложениях МО и ИИ. Современные доситжения МО и ИИ в геофизике.
2. Практика: Реализация понижающего нейросетевого автокодировщика. Обучение на данных MNIST. Визуализация скрытых представлений с использованием t-SNE.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 2

1. Теория: Классификация задач МО с примерами задач геофизики.
2. Практика: Решение задачи классификации рукописных цифр из набора данных MNIST методом опорных векторов (опционально) с предварительным обучением нейросетевого генератора признаков с применением понижающего автокодировщика.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 3

1. Теория: Линейная регрессия как статистическая модель. Линейная регрессия как однослойный перцептрон.
2. Практика: Кластеризация данных - на выбор из предложенных вариантов. Выбор метода – на выбор студента.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 4

1. Теория: Принципы машинного обучения. Общая схема решения задач обучения по прецедентам.
2. Практика: Постановка и решение задачи классификации по предлагаемым наборам данных. Выбор набора данных и метода – на выбор студента.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 5

1. Теория: Логистическая регрессия как статистическая модель. Логистическая регрессия как однослойный перцептрон.
2. Практика: Классификация состояний стратосферного полярного вихря с применением сверточной нейронной сети.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 6

1. Теория: Методы оптимизации в МО. Ландшафт функции потерь. Методы первого и второго порядков.
2. Практика: Исследование и визуализация характеристик статистических признаков данных широкоугольных снимков видимой полусферы небосвода.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 7

1. Теория: Переобучение и недообучение моделей МО. Смещение и разброс. Подходы балансирования смещения и разброса.
2. Практика: Визуализация данных, описывающих состояние стратосферного полярного вихря, с применением t-SNE и самоорганизующихся карт Кохонена.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 8

1. Теория: Метод опорных векторов. Метод опорных векторов с ядром. Метод опорных векторов как двуслойный перцептрон.
2. Практика: Постановка и решение задачи классификации по предлагаемым наборам данных. Выбор набора данных и метода – на выбор студента. (Опционально) с предварительным обучением нейросетевого генератора признаков с применением понижающего вариационного автокодировщика.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 9

1. Теория: Деревья решений. Композиционные алгоритмы на основе деревьев решений.

2. Практика: Постановка и решение задачи кластеризации по предлагаемым наборам данных. Выбор набора данных и метода – на выбор студента. Визуализация результатов (при необходимости) с применением методов сокращения размерности.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 10

1. Теория: Искусственные нейронные сети. Математическая модель искусственного нейрона. Однослойный перцептрон. Многослойный перцептрон. Обучение ИНС методом обратного распространения ошибки. Обобщающая способность и методы регуляризации ИНС.

2. Практика: Постановка и решение задачи оценки балла общей облачности по данным широкоугольных оптических снимков видимой полусферы небосвода. Метод – композиционный на базе деревьев решений, на выбор.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 11

3. Теория: Сверточные нейронные сети. Принципы, лежащие в основе СНС. Операция дискретной одномерной и двумерной свертки. Основные блоки СНС. Современные архитектуры СНС. Задачи, решаемые с применением СНС.

4. Практика: Исследование синтетического набора данных. Постановка и решение задачи кластеризации на этом наборе данных. Анализ решения и свойств полученной модели. Вид модели – на выбор студента.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 12

5. Теория: Задача сокращения размерности. Метод главных компонент как статистическая модель. Самообучающиеся карты Кохонена. Понижающий нейросетевой автокодировщик.

6. Практика: Постановка и решение задачи классификации на синтетических данных с известными свойствами. Анализ решения и свойств полученной модели. Вид модели – многослойный перцептрон.

ЭКЗАМЕНАЦИОННЫЙ БИЛЕТ № 13

7. Теория: Задача кластеризации. Виды методов кластеризации. Отдельные методы кластеризации.

8. Практика: Постановка и решение задачи восстановления регрессии на синтетических данных с известными свойствами. Анализ решения и свойств полученной модели. Вид модели – многослойный перцептрон.

Критерии оценивания

оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины при выполнении курсовой работы, домашних заданий, ответе экзаменационного билета и ответе на вопросы по программе дисциплины;

оценка «отлично (9)» выставляется студенту, показавшему систематизированные, глубокие знания учебной программы дисциплины при выполнении курсовой работы, домашних заданий, ответе экзаменационного билета и ответе на вопросы по программе дисциплины;

оценка «отлично (8)» выставляется студенту, показавшему систематизированные, знания учебной программы дисциплины при выполнении курсовой работы, домашних заданий, ответе экзаменационного билета и ответе на вопросы по программе дисциплины;

оценка «хорошо (7)» выставляется студенту, если он твердо знает материал экзаменационного билета, грамотно и по существу излагает его, демонстрирует умение применять полученные знания на практике при выполнении курсовой работы и домашних заданий, но допускает в ответе или в решении задач некоторые неточности;

оценка «хорошо (6)» выставляется студенту, если он знает материал экзаменационного билета, по существу излагает его, демонстрирует умение применять полученные знания на практике при выполнении курсовой работы и домашних заданий, но допускает в ответе или в решении задач много неточностей;

оценка «хорошо (5)» выставляется студенту, если он знает материал экзаменационного билета, излагает его, демонстрирует умение применять полученные знания на практике при выполнении курсовой работы и домашних заданий, не допускает в ответе грубых ошибок;

оценка «удовлетворительно (4)» выставляется студенту, если во время ответа экзаменационного билета, при выполнении курсовой работы и домашних заданий он показал фрагментарный, характер знаний, недостаточно правильные формулировки базовых понятий, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения;

оценка «удовлетворительно (3)» выставляется студенту, если во время ответа экзаменационного билета, при выполнении курсовой работы и домашних заданий он показал разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушение логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

оценка «неудовлетворительно (2-1)» выставляется студенту, если во время ответа экзаменационного билета, он показал, что не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач.

5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности

Оценка за промежуточную аттестацию в виде зачёта выставляется по результатам текущего контроля успеваемости.

При проведении устного дифференцированного зачёта обучающемуся предоставляется 60 минут на подготовку. Опрос обучающегося по билету не должен превышать двух астрономических часов.

Во время проведения дифференцированного зачёта при подготовке ответов на билеты, обучающиеся могут пользоваться программой дисциплины, конспектами и любой другой литературой.

Во время проведения дифференцированного зачёта при ответе обучающегося на вопросы по билету или по программе дисциплины, он не может пользоваться конспектами и любой другой литературой.