

Министерство науки и высшего образования Российской Федерации
НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ТОМСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ (НИ ТГУ)

Физический факультет

УТВЕРЖДАЮ:
декан физического факультета
С.Н. Филимонов

Рабочая программа дисциплины

Машинное обучение

по направлению подготовки

09.03.02 Информационные системы и технологии

Направленность (профиль) подготовки:
«Информационные системы и технологии в астрономии и космической геодезии»

Форма обучения
Очная

Квалификация
Инженер-разработчик информационных технологий

Год приема
2024

СОГЛАСОВАНО:
Руководитель ОП
О.М. Сюсина

Председатель УМК
О.М. Сюсина

Томск – 2024

1. Цель и планируемые результаты освоения дисциплины (модуля)

Целью освоения дисциплины является формирование следующей компетенции:
ОПК-5. Способен разрабатывать алгоритмы и программы, пригодные для практического применения в области информационных систем и технологий.

Результатами обучения дисциплины являются:

РООПК 5.3. Умеет разрабатывать алгоритмы и компьютерные программы для практического применения.

2. Задачи освоения дисциплины

– Освоить теоретические основы машинного обучения для решения прикладных задач, в том числе и на реальных данных.

– Научиться применять понятийную базу и передовые инструменты машинного обучения для решения научных и практических задач.

3. Место дисциплины (модуля) в структуре образовательной программы

Дисциплина относится к обязательной части образовательной программы.

4. Семестр(ы) освоения и форма(ы) промежуточной аттестации по дисциплине

Семестр 9, экзамен.

5. Входные требования для освоения дисциплины

Для успешного освоения дисциплины требуются результаты обучения по следующим дисциплинам: Основы программирования на языке Python.

6. Язык реализации

Русский

7. Объем дисциплины (модуля)

Общая трудоемкость дисциплины составляет 3 з.е., 108 часа, из которых:

– практические занятия: 32 ч.

Объем самостоятельной работы студента определен учебным планом.

8. Содержание дисциплины (модуля), структурированное по темам

Тема 1. Основы Python

Передовые инструменты машинного обучения (МО) исторически разрабатывались и разрабатываются в основном на языке Python. Данная тема является фундаментом, на котором будет строиться дальнейшая работа. Главные вопросы: основные типы и структуры данных, синтаксис, особенности языка

Тема 2. Библиотеки Python: NumPy

Библиотека Numeric Python или NumPy представляет собой набор методов и типов данных, которые используются в наукоёмкой разработке. Так же эта библиотека лежит в основе других наукоёмких библиотек. Главные вопросы: основные типы данных и приёмы работы с ними; векторные операции; стандартные процедуры.

Тема 3. Библиотеки Python: matplotlib

Библиотека matplotlib позволяет строить различные виды графиков и чертежей. Визуализация данных является важным этапом их анализа. Главные вопросы: построение сложных графиков и их оформление.

Тема 4. Библиотеки Python: pandas

Библиотека `pandas` позволяет работать с табличными данными. Главные вопросы: основные типы данных и приёмы работы с ними; статистические процедуры; связь с реляционной алгеброй.

Тема 5. Теоретические основы машинного обучения:

Постановка задачи МО в общем виде. Типы задач. Главные вопросы: задача; метрики качества; валидация; переобучение; обнаружение и обработка выбросов; метрические методы и гипотеза компактности; функции потерь; обучение с учителем и без.

Тема 6. Метрические алгоритмы машинного обучения:

Разбор основных алгоритмов МО, с точки зрения механики их работы и границ применимости. Алгоритмы: метод k -ближайших соседей; метод опорных векторов; метод опорных векторов по комбинации базисных функций; нелинейная регрессия

Тема 7. Линейная регрессия:

Разбор линейной регрессии с точки зрения МО. Постановка задачи линейной регрессии. Главные вопросы: формула в общем виде; методы поиска решений и границы применимости; линейная регрессия по комбинации базисных функций; извлечение информации из коэффициентов линейной регрессии.

Тема 8. Решающие деревья:

Решающие деревья - один из самых известных и универсальных алгоритмов МО. В данной теме рассматриваются теоретические основания работы решающих деревьев, выбора предикатов ветвления, основные гиперпараметры дерева и границы применения. Главные вопросы: принципиальная схема решающего дерева; предикаты ветвления; глубина дерева; переобучение дерева; визуализация структуры дерева.

Тема 9. Ансамблирование алгоритмов:

Ансамблирование алгоритмов - это использование вместо одной сильной модели комбинации большого количества слабых. Рассматриваются на примере решающих деревьев. Главные вопросы: бэггинг; бустинг; мудрость толпы; сравнение ансамблей с одной моделью.

Тема 10. Глубинное обучение:

Глубинное обучение - это использование глубоких искусственных нейронных сетей (ИНС) в качестве алгоритма МО. Рассматривается программная реализация в виде библиотеки для Python под названием `pytorch`. Главные вопросы: искусственные нейроны; библиотека `pytorch`; типы слоёв; функции активации; инициализация весов; оптимизаторы; нормализация данных; проектирование архитектуры; выбор функции потерь под задачу.

Тема 11. Основные архитектуры искусственных нейронных сетей:

В данной теме рассматриваются основные архитектуры, т.е. некоторые структуры слоёв, ИНС и задачи, для которых они применяются. Архитектуры: полносвязные; свёрточные; автокодировщики; рекуррентные.

Тема 12. Сохранение, передача и работа с предобученными моделями:

ИНС не обязательно обучать с состояния полного хаоса. Возможно использовать предобученные модели для схожих задач. Главные вопросы: сохранение моделей; формат `.pickle`; загрузка моделей; дообучение моделей (`transfer learning`).

9. Текущий контроль по дисциплине

Целью текущего контроля по дисциплине является оценка освоения студентами программы дисциплины. Текущий контроль осуществляется следующими средствами:

1. Контроль посещаемости занятий
2. Выступление у доски на практических занятиях
3. Защита домашних работ

Контроль посещаемости указывает на пропущенные студентом темы, пробелы в которых выявляются путём дополнительных вопросов.

Выступление у доски на практических занятиях строится по схеме:

- Преподаватель ставит задачу на пару
- Студент садится за компьютер и начинает её решать.
- В случае затруднения обращается за помощью к одногруппникам.
- В случае затруднения одногруппников в процесс включается преподаватель.
- Если остаётся время после решения основной задачи, то решаются дополнительные задачи. Например, сравнение моделей или построение аналитических графиков.

Защита домашних работ проводится теми же способами и с теми же целями, что и выступления у доски.

10. Порядок проведения и критерии оценивания промежуточной аттестации

Зачет с оценкой проводится в устной форме по билетам. Билет содержит два вопроса: общий и про конкретные алгоритмы (или модели)

Примеры общих вопросов:

1. Постановка задачи машинного обучения.
2. Функции потерь. Требования к функциям потерь. Функции потерь для задач классификации.
3. Функции потерь. Требования к функциям потерь. Функции потерь для задач регрессии.
4. Метрики качества. Требования к метрикам качества. Метрики качества для задач классификации.
5. Метрики качества. Требования к метрикам качества. Метрики качества для задач регрессии
6. Переобучение. Обнаружение и устранение.
7. Кросс-валидация и k-Fold.
8. Матрица ошибок. Ошибки первого и второго рода. Производные метрики качества.
9. Выбор порога классификация. Кривая точность-полнота и ROC-кривая.
10. Параллельное и последовательное ансамблирование моделей.
11. Метод градиентного спуска и метод обратного распространения ошибки.
12. Математическая модель искусственного нейрона.

Примеры вопросов про алгоритмы и модели:

1. Метод k-ближайших соседей.
2. Метод опорных векторов.
3. Непараметрическая регрессия. Формула Надарая-Уотсона.
4. Линейная регрессия.
5. Логистическая регрессия.
6. Линейная регрессия по комбинации базисных функций.
7. Решающие деревья.
8. Random Forest
9. Градиентный бустинг над решающими деревьями
10. Полносвязные ИНС
11. Свёрточные ИНС
12. Автокодировщики

На экзамене проверяется формирование ОПК-5 в соответствии с индикатором ИОПК-5.3.

Для получения оценки: отлично - студент должен уверенно ответить на оба вопроса из билета, а также на дополнительные вопросы; хорошо - студент должен ответить на оба вопроса из билета, но дать неполный ответ на дополнительных вопросах. Оценка удовлетворительно ставится, если на один из вопросов билета студент ответить не смог.

11. Учебно-методическое обеспечение

- а) Электронный учебный курс по дисциплине в электронном университете «Moodle»
- б) Оценочные материалы текущего контроля и промежуточной аттестации по дисциплине.
- в) План практических занятий по дисциплине.

12. Перечень учебной литературы и ресурсов сети Интернет

1. а) основная литература:
2. – Ярушкина Н. Г. Интеллектуальный анализ временных рядов: Учебное пособие / Н.Г. Ярушкина, Т.В. Афанасьева, И.Г. Перфильева. - М.: ИД ФОРУМ: ИНФРА-М, 2012. - 160 с.:
3. – Информационные аналитические системы [Электронный ресурс] : учебник / Т. В. Алексеева, Ю. В. Амириди, В. В. Дик и др.; под ред. В. В. Дика. - М.: МФПУ Синергия, 2013. - 384 с. - (Университетская серия). - ISBN 978-5-4257-0092-6.
- 4.
5. б) дополнительная литература:
6. – Технология Data Mining: Интеллектуальный анализ данных, Степанов, Роман Григорьевич, 2009г
7. – Барсегян, А. А. Анализ данных и процессов: учеб. пособие / А. А. Барсегян, М. С. Куприянов, И. И. Холод, М. Д. Тесс, С. И. Елизаров. 3-е изд., перераб. и доп. СПб.: БХВ-Петербург, 2009. 512 с.: ил. + CD-ROM (Учебная литература для вузов)
- 8.
9. в) ресурсы сети Интернет:
10. – Введение в машинное обучение. Coursera. К.В. Воронцов
<https://www.coursera.org/learn/vvedenie-mashinnoe-obuchenie>
11. – NumPy Documentation <https://numpy.org/doc/>
12. – Pytorch Documentation <https://pytorch.org/docs/stable/index.html>

13. Перечень информационных технологий

- а) лицензионное и свободно распространяемое программное обеспечение:
– Python 3.7+ и модули: jupyter, numpy, matplotlib, pandas, scikit-learn, pytorch

14. Материально-техническое обеспечение

Аудитории для проведения занятий лекционного типа. Компьютерный класс для практических занятий. Аудитории для проведения индивидуальных и групповых консультаций, текущего контроля и промежуточной аттестации.

Помещения для самостоятельной работы, оснащенные компьютерной техникой и доступом к сети Интернет, в электронную информационно-образовательную среду и к информационным справочным системам.

15. Информация о разработчиках

Красавин Дмитрий Сергеевич, Томский государственный университет, м.н.с.

Галушина Татьяна Юрьевна, к.ф.-м.н., Томский государственный университет, доцент