

Министерство науки и высшего образования Российской Федерации
федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет промышленных технологий и дизайна»
(СПбГУПТД)

УТВЕРЖДАЮ
Директор ВШТЭ



Рабочая программа дисциплины

Б1.В.07

Машинное обучение

Учебный план: _____ ФГОС3++b010302-4_26-14.plx

Кафедра: 16 Прикладной математики и информатики

Направление подготовки:
(специальность) 01.03.02 Прикладная математика и информатика

Профиль подготовки:
(специализация) Прикладная математика и информатика

Уровень образования: бакалавриат

Форма обучения: очная

План учебного процесса

Семестр (курс для ЗАО)		Контактная работа обучающихся		Сам. работа	Контроль, час.	Трудоё мкость, ЗЕТ	Форма промежуточной аттестации
		Лекции	Практ. занятия				
7	УП	17	34	56,75	0,25	3	Зачет
	РПД	17	34	56,75	0,25	3	
Итого	УП	17	34	56,75	0,25	3	
	РПД	17	34	56,75	0,25	3	

Рабочая программа дисциплины составлена в соответствии с федеральным государственным образовательным стандартом высшего образования по направлению подготовки 01.03.02 Прикладная математика и информатика, утверждённым приказом Минобрнауки России от 10.01.2018 г. № 9

Составитель (и):

преподаватель

Коротаев П.Д.

От кафедры составителя:

Заведующий кафедрой прикладной математики и информатики

Ремизова И.В.

От выпускающей кафедры:

Заведующий кафедрой

Ремизова И.В.

Методический отдел:

Фахретдинова А. Р.

1 ВВЕДЕНИЕ К РАБОЧЕЙ ПРОГРАММЕ ДИСЦИПЛИНЫ

1.1 Цель дисциплины: Формирование у обучающихся системных знаний в области теории и практики машинного обучения, развитие навыков построения, обучения и оценки моделей машинного обучения для решения прикладных задач анализа данных, классификации, регрессии и кластеризации.

1.2 Задачи дисциплины:

- изучение основных математических моделей и алгоритмов машинного обучения, включая методы обучения с учителем и без учителя;
- освоение методов предобработки данных, отбора признаков и инженерии признаков;
- формирование практических навыков реализации алгоритмов машинного обучения с использованием современных библиотек (scikit-learn, TensorFlow, PyTorch);
- развитие умений оценки качества моделей, валидации, подбора гиперпараметров и интерпретации результатов;
- ознакомление с основами глубокого обучения и нейронных сетей.

1.3 Требования к предварительной подготовке обучающегося:

Предварительная подготовка предполагает создание основы для формирования компетенций, указанных в п. 2, при изучении дисциплин:

2 КОМПЕТЕНЦИИ ОБУЧАЮЩЕГОСЯ, ФОРМИРУЕМЫЕ В РЕЗУЛЬТАТЕ ОСВОЕНИЯ ДИСЦИПЛИНЫ

ПК-7: Способен проводить работы по совершенствованию автоматизированного документооборота в организации.
Знать: Методы автоматизации процессов обработки документов с применением алгоритмов машинного обучения; принципы классификации и кластеризации текстовых данных.
Уметь: Применять методы машинного обучения для анализа и классификации документов; автоматизировать рутинные операции документооборота с использованием интеллектуальных алгоритмов.
Владеть: Навыками разработки решений задач с помощью методов машинного обучения для задач обработки и маршрутизации документов.
ПК-8: Способен подготавливать необходимые данные и составлять технические задания на проектирование АСУП
Знать: Методы сбора, предобработки и анализа данных для проектирования АСУП; принципы формирования обучающих выборок и оценки качества данных.
Уметь: Подготавливать наборы данных для обучения моделей; формулировать требования к ключевым компонентам машинного обучения в составе АСУП; проводить разведочный анализ данных (EDA).
Владеть: Инструментами предобработки данных (pandas, NumPy, scikit-learn); методами визуализации и статистического анализа данных.
ПК-9: Способен разрабатывать объектные, структурные и документные модели АСУП
Знать: Архитектуры моделей машинного обучения; методы классификации, регрессии, кластеризации; основы нейронных сетей.
Уметь: Проектировать и обучать модели машинного обучения для решения задач АСУП; оценивать качество моделей; подбирать гиперпараметры.
Владеть: Навыками построения и развёртывания моделей машинного обучения с использованием инструментов предобработки данных (Python, scikit-learn, TensorFlow/PyTorch); методами валидации и интерпретации результатов.

3 СОДЕРЖАНИЕ ДИСЦИПЛИНЫ

Наименование и содержание разделов, тем и учебных занятий	Семестр (курс для ЗАО)	Контактная работа		СР (часы)	Форма текущего контроля
		Лек. (часы)	Пр. (часы)		
Раздел 1. Основы машинного обучения	7				О
Тема 1. Введение в машинное обучение. Определение, задачи и типы машинного обучения (обучение с учителем, без учителя, с подкреплением). Основные термины: признаки, целевая переменная, модель, функция потерь, обобщающая способность. Обзор жизненного цикла ML-проекта и этические аспекты (bias, fairness, прозрачность).		2	4	7	
Тема 2. Предобработка данных и инженерия признаков Типы данных и первичный анализ: пропуски, дубликаты, выбросы, описательная статистика. Методы очистки и заполнения пропусков, кодирование категориальных переменных (One-Hot, Label Encoding). Масштабирование и нормализация, создание новых признаков, отбор признаков на основе статистических метрик.		2	4	7	
Тема 3. Линейные модели регрессии и классификации Линейная регрессия: метод наименьших квадратов, градиентный спуск, регуляризация (L1, L2, Elastic Net). Логистическая регрессия: сигмоида, интерпретация вероятностей, граница принятия решений. Метрики качества: MSE, MAE, R^2 для регрессии; Accuracy, Precision, Recall, F1-score и ROC-AUC для классификации.		2	4	7	
Тема 4. Деревья решений и ансамблевые методы Деревья решений: принцип построения, энтропия и критерий Джини, проблема переобучения и стрижка (pruning). Бэггинг и Random Forest: уменьшение дисперсии, out-of-bag-оценка, важность признаков. Бустинг: AdaBoost, градиентный бустинг (GBM, XGBoost, LightGBM) — отличие от бэггинга и борьба с переобучением.		2	4	7	
Раздел 2. Продвинутое методы и практика ML					О

Тема 5. Кластеризация и снижение размерности Кластеризация: K-Means, DBSCAN, иерархическая кластеризация - принципы работы, выбор числа кластеров, метрики качества (силуэт, индекс Дэвиса-Болдина). Метод главных компонент (PCA): математическая идея, собственные векторы и числа, выбор числа компонент по explained variance. t-SNE и UMAP: нелинейное снижение размерности для визуализации, отличия от PCA, настройка перплексии, интерпретация результата.		2	4	7	
Тема 6. Валидация моделей и подбор гиперпараметров. Стратегии валидации: hold-out, k-fold, стратифицированная и групповая кросс-валидация, временные ряды. Борьба с переобучением и недообучением: кривые обучения, bias-variance tradeoff, анализ ошибок модели. Методы подбора гиперпараметров: Grid Search, Random Search, Bayesian Optimization (Optuna/Hyperopt), вложенная кросс-валидация.		2	4	7	
Тема 7. Основы глубокого обучения Устройство нейронной сети: нейрон, слои, функции активации (ReLU, сигмоида, tanh, softmax), прямое и обратное распространение ошибки. Базовые архитектуры: полносвязные сети для табличных данных, свёрточные сети для изображений (CNN), рекуррентные сети для последовательностей (RNN, LSTM). Перенос обучения (Transfer Learning) и fine-tuning: использование предобученных моделей, заморозка слоёв, дообучение на своих данных.		3	6	7,75	
Тема 8. Практика применения ML Пайплайны и автоматизация: построение конвейеров обработки данных и обучения (на примере sklearn Pipeline / ColumnTransformer), сохранение и загрузка моделей (pickle, ONNX). Интерпретация предсказаний модели: SHAP-значения, LIME, визуализация важности признаков, частичная зависимость (Partial Dependence Plots). Деплой модели в production: REST API для модели (FastAPI / Flask), упаковка в Docker, базовые принципы MLOps и мониторинг качества модели в эксплуатации.		2	4	7	
Итого в семестре (на курсе для ЗАО)		17	34	56,75	
Консультации и промежуточная аттестация (Зачет)		0,25			
Всего контактная работа и СР по дисциплине		51,25		56,75	

4 КУРСОВОЕ ПРОЕКТИРОВАНИЕ

Курсовое проектирование учебным планом не предусмотрено

5 БАЛЛЬНО-РЕЙТИНГОВАЯ СИСТЕМА ОЦЕНИВАНИЯ

Использование балльно-рейтинговой системы не предусмотрено.

6. ОПИСАНИЕ ПОКАЗАТЕЛЕЙ, КРИТЕРИЕВ, СИСТЕМЫ И СРЕДСТВ ОЦЕНИВАНИЯ РЕЗУЛЬТАТОВ ОБУЧЕНИЯ

6.1 Показатели оценивания

Код компетенции	Показатели оценивания результатов обучения
ПК-7	1. Демонстрирует понимание методов автоматизации процессов обработки документов с применением алгоритмов машинного обучения; ориентируется в принципах классификации и кластеризации текстовых данных. 2. Применяет методы машинного обучения для анализа и классификации документов; автоматизирует рутинные операции документооборота с использованием интеллектуальных алгоритмов. 3. Использует инструментарий разработки решений на базе машинного обучения для задач обработки и маршрутизации документов организации.
ПК-8	1. Ориентируется в методах сбора, предобработки и анализа данных для проектирования АСУП; понимает принципы формирования обучающих выборок и оценки качества исходных данных. 2. Подготавливает наборы данных для обучения моделей; формулирует требования к ключевым компонентам машинного обучения в составе АСУП; выполняет разведочный анализ данных (EDA). 3. Использует инструменты предобработки данных (pandas, NumPy, scikit-learn); применяет методы визуализации и статистического анализа данных.
ПК-9	1. Демонстрирует понимание архитектур моделей машинного обучения, методов классификации, регрессии, кластеризации; ориентируется в основах нейронных сетей. 2. Проектирует и обучает модели машинного обучения для решения задач АСУП; оценивает их качество; подбирает гиперпараметры. 3. Использует инструменты построения и развёртывания моделей машинного обучения (Python, scikit-learn, TensorFlow/PyTorch); применяет методы валидации и интерпретации результатов.

6.2 Система и критерии оценивания

Шкала оценивания	Критерии оценивания сформированности компетенций	
	Устное собеседование	Письменная работа
Зачтено	Обучающийся демонстрирует знание основных понятий, методов и алгоритмов машинного обучения; уверенно отвечает на вопросы по теории и практике; способен обосновать выбор модели и метрик качества; допускает не более незначительных неточностей.	Задания выполнены верно (либо с несущественными ошибками); представленный код работоспособен и воспроизводит заявленные результаты; полученные результаты корректно интерпретированы; оформление соответствует требованиям.
Не зачтено	Обучающийся не знает базовых понятий машинного обучения; не может объяснить принципы работы алгоритмов и интерпретировать результаты; ответы не соответствуют поставленным вопросам.	Задания не выполнены или выполнены с грубыми ошибками; код не работает или не воспроизводит результаты; интерпретация результатов отсутствует или некорректна.

6.3 Типовые контрольные задания и/или иные материалы, используемые для оценки результатов обучения

6.3.1 Типовые контрольные вопросы

№ п/п	Формулировки вопросов
Семестр 7	
1	Дайте определение машинного обучения. Перечислите основные задачи ML и приведите примеры.
2	В чём отличие обучения с учителем от обучения без учителя и обучения с подкреплением?
3	Опишите типовой цикл решения ML-задачи: от постановки до внедрения модели.
4	Перечислите основные метрики качества для задач классификации и регрессии. В каких случаях каждая из них предпочтительна?
5	Какие методы обработки пропущенных значений существуют? В каких ситуациях их применяют?

6	Сравните методы кодирования категориальных признаков: One-Hot, Target, Ordinal. Когда применять каждый?
7	В чём разница между нормализацией и стандартизацией признаков? Какие алгоритмы чувствительны к масштабу?
8	Опишите подходы к отбору признаков (фильтры, обёртки, встроенные методы).
9	Сформулируйте задачу линейной регрессии. Выведите формулу метода наименьших квадратов.
10	Объясните идею регуляризации L1 и L2. В чём разница между Lasso и Ridge?
11	Опишите принцип работы логистической регрессии. Как она обобщается на многоклассовый случай?
12	Объясните принцип работы метода опорных векторов (SVM). Что такое ядро?
13	Опишите алгоритм построения дерева решений. Какие критерии расщепления вы знаете?
14	В чём идея случайного леса? Как он снижает дисперсию модели?
15	Сравните бэггинг и бустинг. В чём принципиальные отличия Gradient Boosting, XGBoost, LightGBM, CatBoost?
16	Сформулируйте задачу кластеризации. Опишите алгоритм K-means и его ограничения.
17	В чём отличие DBSCAN от K-means? Какие данные подходят каждому алгоритму?
18	Опишите метод главных компонент (PCA). Для чего применяется снижение размерности?
19	Что такое кросс-валидация? Опишите k-fold и stratified k-fold.
20	Объясните понятие переобучения. По каким признакам его можно диагностировать?
21	Опишите подход bias–variance tradeoff. Как он влияет на выбор модели?
22	Сравните методы подбора гиперпараметров: grid search, random search, байесовская оптимизация.
23	Опишите устройство искусственного нейрона. Какие функции активации вы знаете и для чего они нужны?
24	Объясните алгоритм обратного распространения ошибки.
25	В чём отличие свёрточной нейронной сети от полносвязной? Где применяются CNN?
26	Сравните оптимизаторы SGD, Momentum, Adam. Какие их особенности?
27	Опишите этапы полного цикла ML-проекта на примере задачи классификации документов.
28	Какие задачи машинного обучения возникают при автоматизации документооборота? Приведите примеры.
29	Что такое объяснимое машинное обучение (XAI)? Как работают SHAP и LIME?
30	Опишите проблемы воспроизводимости в ML. Какие инструменты (MLflow, DVC) применяются для их решения?

6.3.2 Типовые тестовые задания

Не предусмотрено.

6.3.3 Типовые практико-ориентированные задания (задачи, кейсы)

Задание 1. Классификация служебных документов организации.

Задан размеченный корпус из 5 000 служебных документов организации, отнесённых к одной из 6 категорий (договоры, акты, счета-фактуры, приказы, служебные записки, прочее). Требуется: (а) провести разведочный анализ данных и предобработку текстов (токенизация, лемматизация, удаление стоп-слов); (б) построить векторное представление документов (TF-IDF и/или эмбединги); (в) обучить и сравнить две модели классификации (например, логистическую регрессию и градиентный бустинг); (г) оценить качество моделей по метрикам ассигасу и маско-F1 на отложенной выборке; (д) обосновать выбор итоговой модели для встраивания в контур документооборота.

Задание 2. Прогнозирование значения показателя АСУП.

Дан временной ряд значений целевого показателя АСУП за 24 месяца (продажи, нагрузка на склад, число обращений) и набор сопутствующих признаков. Требуется: (а) выполнить EDA и оценить наличие тренда/сезонности; (б) подготовить признаки (лаги, скользящие средние, календарные признаки); (в) обучить две модели регрессии (Ridge/Lasso и градиентный бустинг); (г) оценить качество прогноза по MAE, RMSE, MAPE на временной кросс-валидации; (д) визуализировать прогноз и доверительный интервал, дать рекомендации по применению модели.

Задание 3. Кластеризация клиентов / объектов АСУП.

Имеется таблица из 10 000 объектов (клиенты, поставщики или единицы оборудования) с 15 числовыми и 5 категориальными признаками. Требуется: (а) подготовить данные (масштабирование, кодирование, обработка пропусков); (б) применить понижение размерности (PCA или UMAP) для визуализации; (в) выполнить кластеризацию двумя методами (K-means и DBSCAN) и подобрать число кластеров; (г) оценить качество разбиения по silhouette и Davies–Bouldin; (д) дать содержательную интерпретацию полученных кластеров и предложить рекомендации для бизнес-процесса.

6.4 Методические материалы, определяющие процедуры оценивания результатов обучения

6.4.1 Условия допуска обучающегося к промежуточной аттестации и порядок ликвидации академической задолженности

Проведение промежуточной аттестации регламентировано локальным нормативным актом СПбГУПТД «Положение о проведении текущего контроля успеваемости и промежуточной аттестации обучающихся»

6.4.2 Форма проведения промежуточной аттестации по дисциплине

Устная ☒ Письменная ☒ Компьютерное тестирование ☐ Иная ☐

6.4.3 Особенности проведения промежуточной аттестации по дисциплине

Промежуточная аттестация по дисциплине проводится в форме зачёта в смешанном устно-письменном формате. Зачёт состоит из двух частей: (1) устное собеседование по теоретическим вопросам из перечня, утверждённого фондом оценочных средств; (2) письменная (практическая) часть — выполнение типового практико-ориентированного задания с использованием персонального компьютера, среды Python и библиотек scikit-learn / TensorFlow или PyTorch. Время на устную часть — до 15 минут на одного обучающегося; время на письменную-практическую часть — до 60 минут. Для получения зачёта обучающийся должен дать удовлетворительные ответы на теоретические вопросы и продемонстрировать работоспособное решение практической задачи с корректной интерпретацией результатов. Использование справочной литературы и документации к библиотекам в практической части допускается.

7. УЧЕБНО-МЕТОДИЧЕСКОЕ И ИНФОРМАЦИОННОЕ ОБЕСПЕЧЕНИЕ ДИСЦИПЛИНЫ

7.1 Учебная литература

Автор	Заглавие	Издательство	Год издания	Ссылка
7.1.1 Основная учебная литература				
Протодьяконов, А. В., Пылов, П. А., Садовников, В. Е.	Алгоритмы Data Science и их практическая реализация на Python	Москва, Вологда: Инфра-Инженерия	2022	https://www.iprbookshop.ru/124000.html
Павлова, А. И.	Искусственные нейронные сети	Москва: Ай Пи Ар Медиа	2021	http://www.iprbookshop.ru/108228.html
7.1.2 Дополнительная учебная литература				
Закревская, Н. С., Ковалевский, А. П.	Специальные главы математического анализа. Теория вероятностей	Новосибирск: Новосибирский государственный технический университет	2022	https://www.iprbookshop.ru/126597.html
Веретельникова, Е. Л.	Теория вероятностей. Случайные события	Новосибирск: Новосибирский государственный технический университет	2022	https://www.iprbookshop.ru/127547.html

7.2 Перечень профессиональных баз данных и информационно-справочных систем

Электронно-библиотечная система IPRbooks [Электронный ресурс]. URL: <http://www.iprbookshop.ru/>
Электронная библиотека ВШТЭ СПб ГУПТД [Электронный ресурс]. URL: <http://nizrp.narod.ru>
Электронно-библиотечная система «Айбукс» [Электронный ресурс]. URL: <https://www.ibooks.ru/>
Информационная система «Единое окно доступа к образовательным ресурсам. Раздел. Информатика и информационные технологии» [Электронный ресурс].

7.3 Перечень лицензионного и свободно распространяемого программного обеспечения

MicrosoftOfficeProfessional 2013

7.4 Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине

Аудитория	Оснащение
-----------	-----------

Компьютерный класс	Мультимедийное оборудование, компьютерная техника с возможностью подключения к сети «Интернет» и обеспечением доступа в электронную информационно-образовательную среду
Лекционная аудитория	Мультимедийное оборудование, специализированная мебель, доска