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

ВЫПИСКА ИЗ ПРОТОКОЛА № 8

заседания учебно-методического совета от 09 апреля 2021 года.

ПОВЕСТКА:

Рассмотрение дополнительных общеобразовательных и профессиональных программ.

Проректор по учебной работе А. А. Воронов

СЛУШАЛИ: директора по внутреннему контролю и аудиту Е. Г. Евсеева о представлении дополнительных общеобразовательных и профессиональных программ. (ЦДПО).

ПОСТАНОВИЛИ:
Рекомендовать к утверждению в установленном порядке дополнительную профессиональную программу повышения квалификации «Методы анализа данных и машинного обучения».

Решение принято единогласно заочным голосованием.

Форма проведения заседания: заочная

Председатель УМС МФТИ А.А. Воронов

Ученый секретарь УМС МФТИ М.В. Березникова

Министерство науки и высшего образования Российской Федерации

федеральное государственное автономное образовательное учреждение высшего образования

«Московский физико-технический институт
(национальный исследовательский университет)»

УТВЕРЖДАЮ

Ректор МФТИ

д-р физ.-мат. наук, профессор,

член-корреспондент РАН

_____ Н.Н. Кудрявцев

«___» _____ 2021 г.

Дополнительная профессиональная программа

повышения квалификации

«Методы анализа данных и машинного обучения»

Код ОКВЭД 63.11 - Деятельность по обработке данных, предоставление услуг по размещению информации и связанная с этим деятельность
Код ОКВЭД 62.01 Разработка компьютерного программного обеспечения

УГСН 01.00.00 Математика и механика

ФГОС 01.03.02 «Прикладная математика и информатика (уровень бакалавриата)»

Москва 2021

1. Общая характеристика программы

1.1 Цель реализации программы

Курс рассчитан на слушателей, которые хотят ознакомиться с современными методами и инструментами машинного обучения, одной из самых востребованных областей в ИТ. Будут рассмотрены основные методы прикладной статистики, например, A/B - тестирование; методы классического машинного обучения (линейные модели, деревья решения, ансамбли алгоритмов, бустинги), задачи анализа временных рядов, обработки текстов, основы нейронных сетей. Для успешного завершения курса слушателю необходимо выполнить собственный проект, где он решит прикладную задачу с использованием реальных данных и изученных методов машинного обучения.

Курс подойдет тем, кто хочет стать начинающим специалистом в data science.

Целевая аудитория программы: ИТ-специалисты, желающие повысить свои компетенции в Анализе Данных, или специалисты, желающие переквалифицироваться в разрабстчиков, занимающихся Анализом Данных со знанием основных на сегодняшний день инструментов и библиотек.

Задачи программы: совершенствование и (или) получение новых компетенций, необходимых для осуществления профессиональной деятельности, повышение профессионального уровня в рамках имеющейся квалификации ИТ-специалистов.

Совершенствуемые компетенции

Компетенции формируемые и совершенствуемые в результате обучения представлены в таблицах 1 и 2.

Таблица 1

№	Компетенция в соответствии с направлением подготовки	«Прикладная математика и информатика (уровень бакалавриата)» код: 01.03.02
1.	Способность демонстрации общенаучных базовых знаний естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой	ПК-1
2.	Способность понимать и применять в исследовательской и прикладной деятельности современный математический аппарат	ПК-2
3.	Способность критически переосмысливать накопленный опыт, изменять при необходимости вид и характер своей профессиональной деятельности	ПК-3
4.	Способность применять в профессиональной деятельности современные языки программирования, библиотеки и пакеты программ, а также уверенно владеть методами машинного обучения	ПК-4

Таблица 2

№	Универсальные компетенций (УК), общекультурные компетенций (ОК) и общепрофессиональные компетенций	«Прикладная математика и информатика (уровень бакалавриата)» код: 01.03.02
1.	Способность к критическому анализу и оценке современных научных достижений, генерированию новых идей при решении исследовательских и практических задач, в том числе в междисциплинарных областях	УК-1
2.	Способность к абстрактному мышлению, анализу, синтезу, способность совершенствовать и развивать свой интеллектуальный и общекультурный уровень	ОК-2
3.	Готовность действовать в нестандартных ситуациях, нести социальную и этическую ответственность за принятые решения	ОК-3
4.	Способность использовать базовые знания естественных наук, математики и информатики, основные факты, концепции, принципы теорий, связанных с прикладной математикой и информатикой в контексте анализа данных и смежных задач	ОПК-1
5	Способность приобретать новые научные и профессиональные знания, используя современные образовательные и информационные технологии	ОПК-2

1.2. Планируемые результаты обучения

Планируемые результаты обучения представлены в таблице 3.

Таблица 3

№	Знать	Коды компетенций
1.	Принцип работы алгоритмов классификации	ПК-1 , ОПК-2
2.	Принцип работы алгоритмов регрессии	ПК-1 , ОПК-2
3.	Принцип работы методов отбора признаков (одномерные, жадные, на основе моделей)	ПК-1, ОПК-2
4.	Постановка задачи обучения по прецедентам	ПК-1, ОПК-2
5.	Критерии выбора модели для построения прогноза	ПК-1, ОПК-2
6.	Язык Python и библиотеки для анализа данных (NumPy, SciPy, Pandas, Matplotlib, Sklearn)	ПК-1, ОПК-2
8.	Процедуру вычисления значения метрик качества моделей	ПК-1, ОПК-2
10.	Методы работы с естественным языком	ПК-1, ОПК-2
11.	Анализ временных рядов методами машинного обучения	ПК-1, ОПК-2

№	Уметь	Коды компетенций
1.	Решать задачу классификации данных	ПК-2, ОК-2
	Решать задачу регрессии на данных	ПК-2, ОК-2
2.	Выбрать алгоритм классификации или регрессии, оптимальный для конкретной задачи	ПК-2, ПК-3, УК-1
3.	Перечислить и охарактеризовать метрики качества	ПК-2, ОК-2, ОПК-1
4.	Применить средства sklearn для предобработки и процессинга объектов	ПК-2, ПК-3, ОПК-1
5.	Описать решение задачи классификации	ПК-2, ОК-2
6.	Описать решение задачи регрессии	ПК-2, ОК-2
7.	Применить на практике классификацию	ПК-2, ПК-3, УК-1, ОПК-1
8.	Применить на практике регрессию	ПК-2, ПК-3, УК-1, ОПК-1
9.	Проанализировать проблемы, возникающие при решении задачи классификации, и предложить пути их решения	ПК-3, УК-1
10.	Проанализировать проблемы, возникающие при решении задачи регрессии, и предложить пути их решения	ПК-3, УК-1
11.	Использовать STL-разложение временного ряда	ПК-1, ПК-2, ОК-2

1.3 Категории обучающихся: уровень образования – ВО, область профессиональной деятельности –для начинающих разработчиков и аналитиков данных.

1.4 Форма обучения – заочному принципу (с применением электронного обучения и дистанционных образовательных технологий) в формате вебинаров.

1.5 Объем программы – 128 академических часов:

Режим обучения – два раза в неделю по 4 академических часа, контрольное мероприятие – защита проекта.

2. Содержание программы

2.1. Учебный (тематический) план

Таблица 4

№ п/п	Наименование модулей	Всего, час.	в том числе:				
			лекции	практические занятия (семинары)	самостоятельная работа	контрольные задания	Форма контроля
1	**Базовые алгоритмы**	61,5	18	18	24	1,5	тестирование
2	**Глубокое обучение**	36,5	12	12	12	0,5	тестирование
5	**Итоговая аттестация**	30	4	0	26	0	проект
	Итого:	**128**	**34**	**30**	**36**	**2**	

2.2. Учебная программа

Таблица 5

№ п/п	Содержание обучения, наименование и тематика практических занятий (вебинаров), самостоятельных работы	Объем, ак.час.
	Модуль 1. Базовые алгоритмы.	61,5
1	Введение в ML	6
	Чем заниматься в ML?	1
	Постановка задачи ML, типы разных постановок задач	1
	Типы данных	1
	Конвейер обработки данных, Crispr-dm	1
	Метрики качества и функция потерь	1
	Метрики качества задачи классификации и регрессии (их получение и интерпретация)	1
2	Предобработка	6,25
	Проблемы и варианты загрузки	1
	Пример EDA	1
	Подготовка: стандартизация, полиномиальные признаки (что такое kernel trick)	1
	Метрики качества задачи классификации алгоритма KNN, перебор параметров алгоритмов с помощью sklearn	1,25

	Интерпретация метрик качества классификации	1
	Метрики задачи регрессии, перебор параметров алгоритма для максимизации метрик регрессии	1
3	Линейная регрессия	8,25
	Постановка задачи линейной регрессии	2,25
	- Постановка задачи линейной регрессии - Аналитическое решение -- переопределенная система линейных уравнений -- множественная регрессия -- линейные трансформации пространства -- визуализаиции матричных перемножений -- нормальные уравнения и визуализация аналитического решения -- реализация на numpy линейной и полиномиальной регрессии (проблема некорректной постановки задачи) и через sklearn	3
	- Функция правдоподобия -- Максимизация функции правдоподобия и условия теоремы Маркова-Гаусса -- Компромисс дисперсии-смещения -- Вывод аналитического решения через максимизацию функции правдоподобия (получение MSE) - Градиентный спуск на примере линейной регрессии -- Предел, секущая и касательная, метод конечных разностей, правила дифференцирования на простом примере, исследование функции на экстремумы. -- Вывод шагов градиентного спуска из MSE -- Реализация на numpy (стахостического) градиентного спуска	3
4.	Регуляризация и логистическая регрессия	6,25
	- Компромисс смещения-разброса - Регуляризация -- Некорректно поставленная задача: регуляризация логистической регрессии - l1 и l2 регуляризации. -- Вывод с шагами градиентного спуска. -- Реалзиация на numpy и через sklearn. -- Отбор признаков с l1 регуляризацией	3
	- Логистическая регрессия -- Кросс-энтропия. Вывод шагов спуска логистический регрессии (связь с линейной регрессией). -- Визуализация отступа и logistic loss, связь с кросс-энтропией -- Принцип максимального правдоподобия и логистическая регрессия, сравнение с zero-one-loss -- Реализация на numpy логистической регрессии и перебор параметров через sklearn -- Регуляризация l1, l2 логистической регрессии, перебор параметров с помощью sklearn -- Полиномиальные признаки и логистическая регрессия	3,25
5.	SVM, работа с представлением, наивный Байес	6
	- SVM -- Постановка задачи SVM, визуализация -- Вывод шагов hard-margin svm -- Hinge loss: регуляризация SVM. Связь с logloss и ReLU -- Ядра в SVM (теорема Мерсера) -- Реализация на numpy SVM, перебор параметрво через sklearn	2
	- Что такое обучение представлению в ML, отбор признаков -- Варианты подготовки текста, TF-IDF -- Алгоритмы аугментации - Smote, adasyn, tomek links -- Кривые обучения - как обнаружить нерепрезентативные данные -- Методы поиска выбросов	2

	- Теория вероятностей и наивный Байес -- Случайная величина, понятия вероятности и условной вероятноси, (не)зависимые события, теорема Байеса -- Алгоритм найивного байеса, постановка задачи, проблемы и применения -- Варианты наивного Байеса и реализация на numpy. Модификации наивного байеса (tf-idf + фильтрация предлогов + bow с парами слов), перебор параметров на sklearn. Поддержка программы чтения с экрана включена.	2
6.	Методы понижения размерности	8,25
	-Зачем нужны методы понижения размерности - Разложения -- Спектральное разложение -- Сингулярное разложение -- Оценка ранга и сжатие без потерь -- Прроксимация матрицей меньшего ранга, норма Фробениуса -- Преобразование признаков -- Сингулярное разложение и низкоранговое приближение	2
	PCA -- Постановка задачи -- Реализация через матрицу ковариации -- Реализация чреез сингулярное разлежение -- Выбор числа компонент -- Реализация через sklearn на примере понижения размерности tf-idf -- Недостатки PCA	2
	T-SNE -- Постановка задачи и отличие от PCA -- Функция потерь и перплексия. -- Недостатки T-SNE	2
	UMAP -- Сравнение с T-SNE -- Анализ топологических данных и упрощенные комплексы -- Низкоранговое представление -- Efficient K-Nearest Neighbor Graph Construction for Generic Similarity Measures	1,25
	- Работа с tensorboard projector	1
7.	Кластеризация	6,25
	- Постановка задачи кластеризаций и сферы применения - Нормы - Типы алгоритмов - Метрики задачи кластеризации - Алгоритмы -- k-means / k-means++ / k-medians -- Mean shift -- Агломеративная кластеризация -- DBSCAN -- Affinity Propagation -- Спектральная кластеризация	3,25
	- Приложения алгоритвом кластеризации на примере FAISS - Реализация кластеризации через sklearn, сравнение алгоритмов	3
8.	Деревья и градиентный бустинг	14,25

	- Дерево решений, отличие от остальных алгоритмов -- Принципы построения в задачах регрессии и классификации -- Работа с количественными признаками -- Комбинаторная энтропия, вывод энтропии по Шеннону, прирост информации. Реализация с нуля на питоне на игрушечном примере. -- Реализация через sklearn -- Недостатки алгоритма дерева	4
	- Ансамблевые методы -- Бэггинг -- Случайный лес, сверхслучайные деревья -- Оценка важности признаков -- Недостатки деревьев	3,25
	Градиентные методы -- Что было до: адабуст -- Алгоритм Фридмана. Градиентный бустинг для задачи регрессии, вывод шагов обновления для различных функций потерь -- Градиентный бустинг для задачи классификации, вывод шагов обновления для различных функций потерь	4
	- Работа с catbost и перебор моделей с tensorboard	3
	Модуль 2. Глубокое обучение	36,5
1	Введение в Pytorch	6
	— тензоры и операции с ними, вычислительный граф ИНС и обратное распространение ошибки — полносвязная нейронная сеть, обратное распространение ошибки, загрузка данных, автодифференцирование на примере логистической регрессии Универсальные способы оптимизации моделей — оптимизаторы, батчинг, нормализации, функции активации, работа с моделями, pytorch lightning Оптимизация, часть 2 — Дистилляция, прунинг, квантизация, onnx	6
2	Машинное зрение, введение	6,25
	— Основы CNN и типичные операции в сверточных сетях Машинное зрение, архитектуры — Виды архитектур и их реализация, дообучение Машинное зрение - детекция 2 — faster r-cnn, реализация Машинное зрение - детекция 1 — Yolo подобные системы, реализация Машинное зрение - сегменатция — U-net Машинное зрение - GAN — stylegan 2, normalizing flow Машинное зрение в 3d — 3d ml, point cloud	6,25
3	Обработка текста, введение	6

	— дистрибутивная семантика, w2v, lstm Обработка текста, lstm+seq2seq — задачи Seq2seq на примере задач перевода Обработка текста, attention, attention block — Механизм внимания, реализация на pytorch, визуализация Обработка текста, transformer — Реализация трансформер модели на pytorch Обработка текста, bert — Трансформеры как универсальная модель машинного обучнения, трансформеры и Bert-ология Обработка текста, SOTA — huggin face, deepPavlov на примере разных задач NLP	6
4	Обработка временных рядов, классический подход	6,25
	— эконометрические подходы Обработка временных рядов, глубокое обучение — lstm, cnn, transformer Обработка звука, sota — tacotron2, huggin face, asteroid	6,25
5	Обучение с подкреплением, введение	6
	— online/offline learning, q-learning, policy-gradients Обучение с подкреплением, policy gradients — gym, road to PPO и RND Обучение с подкреплением, model-based — Alpha zero Обучение с подкреплением, transformer — upside-down rl, dicision transformer	6
	Итого	**128**

3. Формы аттестации и оценочные материалы

Оценка качества освоения программы проводится по двухбалльной системе: «зачтено», «не зачтено» по результатам промежуточного контроля (тестирование, проверочные задания на взаимную оценку), контроля посещаемости практических занятий (вебинаров) и результатам итоговой аттестации.

Слушатель считается аттестованным в случае положительных результатов работы (не менее 70% баллов от итоговой оценки) в процессе обучения и успешной сдачи экзамена. При этом баллы за экзамены начисляются только при достижении 50% порога при прохождении каждого экзаменационного испытания. После аттестации слушатель получает оценку “отлично”, если набрано не меньше 80% баллов от возможного максимума, “хорошо”, если набрано не меньше 65%, “удовлетворительно”, если набрано не меньше 40%, в противном случае слушатель курс не сдает.

Результат тестирования, решения проверочных заданий и написания кода проверяется автоматически системой на образовательной платформе. Экзаменационная работа проверяется преподавателем.

Составляющие процесса обучения, которые оцениваются в ходе обучения, и их вклад в итоговую оценку представлены в таблице 6.

Таблица 6 – Составляющие процесса обучения

	Основные показатели оценки	**Вклад в итоговую оценку**
1	Основной курс обучения на образовательной платформе	50%
2	Практические занятия	10%
3	Итоговая аттестация	40%

Оценочные материалы

Пример тестового задания

Из чего можно вывести основные метрики задачи бинарной классификации?

F1
Recall (Полнота)
Precision (Точность)
Confusion matrix (Матрица ошибок)

Ответ: Confusion matrix

4. Организационно-педагогические условия реализации программы

4.1 Учебно-методическое обеспечение и информационное обеспечение программы

Основная литература

1. Прикладная и компьютерная лингвистика / Под ред. И. С. Николаева, О. В. Митрениной, Т. М. Ландо. — М.: URSS, 2016. — 320 с.

Дополнительная литература

1. Murphy K.P. Machine Learning: A Probabilistic Perspective. The MIT Press, 2012.
2. Sheldon A. Linear Algebra Done Right. Springer, 2015.
3. DasGupta A. Probability for statistics and machine learning. Springer, 2011.
4. Умнов. Аналитическая геометрия и линейная алгебра (2011) — МФТИ.
5. Нестеров. Методы выпуклой оптимизации (2010)
6. Diez, Barr, Çetinkaya-Rundel, Dorazio. Advanced High School Statistics (2015)
7. DasGupta. Probability for Statistics and Machine Learning: Fundamentals and Advanced Topics (2011)
8. Vance Martin, Stan Hurn, David Harris. Econometric Modelling with Time Series.
9. Specification, Estimation and Testing. Cambridge University Press, 2013.
10. Lütkepohl, H., 2005, New introduction to multiple time series analysis. Springer Science & Business Media.
11. Box, G.E.P., Jenkins G.M. and G.C. Reinsel, 2008, Time Series Analysis: Forecasting and Control (Wiley Series in Probability and Statistics), Wiley Publ., 4th ed.

12. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.
13. Введение в информационный поиск / Кристофер Д. Маннинг, Прабхакар Рагхаван, Хайнрих Шютце. — М.: Вильямс, 2011. — 528 с.
14. Natural Language Processing for the Working Programmer / Daniël de Kok, 2011.

4.2 Материально-технические условия реализации программы

Таблица 7

Наименование специализированных аудиторий, кабинетов, лабораторий	**Вид занятий**	**Наименование оборудования, программного обеспечения**
Система дистанционного обучения провайдера массовых открытых онлайн курсов	Лекции	Слушателю необходимо наличие доступа в сеть интернет, компьютер. Преподавателю курса необходимо наличие доступа администратора курса на LMS-платформе к материалам курса.
Информационно-коммуникационная платформа дистанционных семинаров	Практические занятия (дистанционные семинары)	Слушателю необходимо наличие доступа в сеть интернет, компьютер. Преподавателю курса необходимо оборудование для проведения дистанционных семинаров (вебинаров), качественный отказоустойчивый доступ в сеть интернет.
Система дистанционного обучения провайдера массовых открытых онлайн курсов	Самостоятельная работа	Наличие компьютера и доступа в сеть интернет.
Система дистанционного обучения провайдера массовых открытых онлайн курсов	Рубежный контроль, Итоговая аттестация	Наличие компьютера и доступа в сеть интернет.

5. Организация образовательного процесса

Слушатели получают доступ к электронным учебным материалам посредством ресурсов поддержки электронного обучения ЦИОТ МФТИ и партнерских образовательных площадок. Форматы представления электронных учебных материалов: в виде массовых онлайн курсов (МООС) в системе дистанционного обучения провайдера массовых открытых онлайн курсов.

Преподаватель проводит практические занятия дистанционно в форме вебинаров с использованием платформы ZOOM (или аналогичной).

Самостоятельная работа выполняется слушателем в удобном для него режиме.

В таблице 8 описаны образовательные технологии.

Таблица 8

№ п/п	Вид занятия	Форма проведения занятий	Цель
1	Лекция	Самостоятельный просмотр видеолекций	Ознакомление слушателей с базовым материалом по тематике курса
2	Практические занятия	Выполнение практических заданий, получение обратной связи от преподавателя. Обсуждение вопросов, возникших в результате просмотра видеолекций и изучения литературы.	Практическое освоение теоретических знаний, а также углубление знаний по курсу
3	Самостоятельная работа	Самостоятельное изучение дополнительных материалов и литературы. Выполнение тренировочных тестов и заданий.	Углубление знаний по курсу.
4	Выполнение контрольных заданий	Выполнение тестов, проверочных заданий. Написание кода на языке Python.	Практическое освоение теоретических знаний, контроль освоения материалов.
5	Итоговая аттестация	Защита выпускной квалификационной работы.	Практическое освоение теоретических знаний, контроль освоения материалов.

6. Составители программы:

Райгородский Андрей Михайлович
Доктор физико-математических наук, директор ФПМИ МФТИ
Благодарный Евгений Владимирович
заведующий учебно-методической лабораторией инноватики
Илья Тихонов
преподаватель и составитель методических материалов
Иванова Анастасия Сергеевна
руководитель проектов учебно-методической лаборатории инноватики ФПМИ

Согласовано

Зам. директора ЦДПО

У.Б. Вещезерова _____

Согласовано

Директор ФПМИ, д.ф.-м.н.

_____ А.М. Райгородский