МОСКОВСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ИМЕНИ М.В. ЛОМОНОСОВА

ЭКОНОМИЧЕСКИЙ ФАКУЛЬТЕТ

УТВЕРЖДАЮ

Декан экономического факультета МГУ

_____/проф. А.А.Аузан/

«_____» _____2025 г.

РАБОЧАЯ ПРОГРАММА ДИСЦИПЛИНЫ

«Машинное обучение и анализ данных-1 (на англ.языке)»

Уровень высшего образования

магистратура

Направление подготовки (специальность)

38.04.01 Экономика

Направленность (профиль) ОПОП

Анализ данных в экономике

Форма обучения

очная

Москва, 2025

1. **Место и статус дисциплины в структуре основной профессиональной образовательной программы подготовки магистра**

Статус дисциплины: *вариативная (по выбору программы)*
Триместр: *2*

2. **Входные требования для освоения дисциплины**

Для успешного освоения данного курса требуются знания и навыки, полученные в следующих дисциплинах:

— теория вероятностей и математическая статистика;
— эконометрика;
— основы программирования.

3. **Планируемые результаты обучения по дисциплине, соотнесенные с требуемыми компетенциями выпускников**

Формируемые компетенции	**Планируемые результаты обучения по дисциплине, соотнесенные с требуемыми компетенциями**
Способность формулировать научно обоснованные гипотезы, создавать теоретические модели явлений и процессов, применять методологию научного познания в профессиональной деятельности **(М.УК-1)**	**УМЕТЬ** выдвигать научно обоснованные гипотезы, поддающиеся операционализации, моделировать явления и процессы на основе системного видения различных отраслей знаний **М.УК-1.Ум.1**
Способность применять продвинутые инструментальные методы экономического анализа в прикладных и/или фундаментальных исследованиях **(М.ОПК-5)**	**УМЕТЬ** обрабатывать информацию при помощи методов машинного обучения **М.ОПК-5.Ум.1**
Способность проводить самостоятельные исследования в соответствии с разработанной программой **(М.ПК-3)**	**ЗНАТЬ** современные научные методы машинного обучения и анализа данных **М.ПК-3.Зн.1**
	УМЕТЬ применять современные научные методы машинного обучения и анализа данных в экономических исследованиях **М.ПК-3.Ум.1**
Способность представлять результаты проведенного исследования научному сообществу в виде статьи или доклада **(М.ПК-4)**	**УМЕТЬ** представлять результаты научного исследования в систематизированном виде в письменной форме **М.ПК-4.Ум.1**
	УМЕТЬ создавать презентации по итогам исследований и делать устные научные доклады **М.ПК-4.Ум.2**
Способность анализировать и использовать различные источники информации для проведения экономических расчетов	**УМЕТЬ** оценивать качество источников экономической информации **М.ПК-9.Ум.1**

(М.ПК-9)	**УМЕТЬ** применять качественные и количественные методы для проведения прикладных экономических исследований **М.ПК-9.Ум.2**
Способность разрабатывать эконометрические модели и модели машинного обучения исследуемых экономических процессов и явлений, интерпретировать полученные результаты **(М.СПК-1)**	**ЗНАТЬ** современные инструментальные методы, применяемые в экономических исследованиях **М.СПК-1.Зн.1**
	УМЕТЬ применять современные инструментальные методы к релевантным данным для решения заданного или самостоятельно сформулированного исследовательского вопроса **М.СПК-1.Ум.1**
	УМЕТЬ представлять результаты аналитических расчетов, полученные с применением современных инструментальных методов **М.СПК-1.Ум.2**
Способность видеть логические связи в системе собранной, обработанной и проанализированной информации, и на основании этого разрабатывать рекомендации для лиц, принимающих решения на микро- и макроуровне, или бизнес-решения **(М-СПК-4)**	**УМЕТЬ** при помощи методов анализа данных сравнивать альтернативные решения и находить оптимальные по заданным метрикам качества **М.СПК-4.Ум.1**
	УМЕТЬ на основе сделанных выводов об оптимальности решения разрабатывать рекомендации для лиц, принимающих решения **М.СПК-4.Ум.2**

4. **Объем дисциплины по видам занятий**

Объем дисциплины составляет 6 зачетных единицы: 216 академических часов, из которых 56 академических часов составляет аудиторная нагрузка, из них 28 академических часов — лекции и 28 академических часов — семинары, 52 академических часа — групповая контактная работа, 0 академических часов — индивидуальная контактная работа, 108 академических часов составляет самостоятельная работа магистранта.

5. **Формат обучения:** используется электронная информационная среды экономического факультета МГУ имени М.В.Ломоносова «ON.ECON».

6. **Содержание дисциплины, структурированное по темам (разделам) с указанием отведенного на них количества академических часов и виды учебных занятий**

Название темы	Трудоёмкость (в академических часах) по видам работ				
	Всего часов	Контактная работа студента с преподавателем, часы			Самостоятельная работа студента, часы
		Занятия лекционного типа	Занятия семинарского типа	Контактные часы (консультации групповые)	
1. Основные задачи машинного обучения.		4	4		
2. Регрессия (линейные модели).		2	2	8	
3. Оптимизационные методы.		2	2	4	
4. Классификация (линейные модели).		2	2	4	
5. Оценка качества модели.		2	2	4	
6. Ядерные методы (регрессия).		2	2	4	
7 Ядерные методы (классификация).		2	2	4	
8. Кластеризация		2	2	4	
9. Снижение размерности		2	2	4	
10. Деревья решений		2	2	4	
11. Бутстрап и анализ моделей		4	4	8	
12. Бустинг		2	2	4	
Аттестация (экзамен)	4	4			
Всего часов	216	28	28	52	108

Краткое содержание тем дисциплины

1. **Основные задачи машинного обучения.** Онтологические допущения (главное — законосообразность и познаваемость мира). Классификация задач машинного обучения: с учителем (классификация, регрессия) и без учителя (кластеризация, снижение размерности). Примеры наборов данных и решаемых на них задач. Типы данных (непрерывные, категориальные). Метод K ближайших соседей (KNN) для регрессии и классификации. Библиотеки Python для хранения данных, визуализации и машинного обучения. Среда Jupyter Notebook. Обзор курса.
   Семинар: настройка Питона, загрузка датасетов, работа с Polars, Matplotlib, NumPy, реализация KNN.
   Литература: (probML2022, гл. 1, 16.1), (ISL2024/ISL2021, гл. 2)

2. **Регрессия (линейные модели).** Формулировка модели, матрица признаков, понятие эмбеддинга (вектора признаков). Оценка методом максимального правдоподобия с разными распределениями шума (нормальное и Лапласа). Оптимизация правдоподобия: аналитические решения, градиентный спуск. Метрики качества модели: Mean Squared Error, Mean Absolute Error.
   Семинар: датасет для регрессии, sklearn.
   Литература: (probML2023, sec. 11), (ISL2024/2021, гл. 3)

3. **Оптимизационные методы.** Оптимизационные задачи, максимизация лог-правдоподобия как оптимизационная задача. Решения в явном виде, итеративные алгоритмы для поиска решения. Вывод градиентного спуска и метода Ньютона. Свойства градиентных методов, намёк на существование не-градиентных оптимизационных алгоритмов.
   Семинар: датасет и простая модель, требующая градиентных методов; реализация градиентного спуска с нуля.
   Литература: (probML2022, гл. 8), (EDO2021, sec. 4), (NumOpt2006, sec. 2)

4. **Классификация (линейные модели).** Формулировка логистической «регрессии» (для 2-х классов, для 3+ классов), оценка методом максимального правдоподобия (распределения: Бернулли и категориальное). Логистическая сигмоида, softmax. Оптимизация правдоподобия: градиентный спуск. Метрики качества модели: accuracy, precision, recall, F-мера, ROC-AUC, …
   Семинар: датасет для классификации (ирисы Фишера?), sklearn, confusion matrix.
   Литература: (probML2022, sec. 10)

5. **Оценка качества модели.** Переобучение (на примере полиномиальных признаков), обучающая и тестовая выборки, кросс-валидация. Регуляризация (Ridge, LASSO), интерпретация регуляризованных оценок как оценок методом апостериорного максимума. Кросс-валидация.
   Семинар: визуальная демонстрация переобучения, подбор гиперпараметров на кросс-валидации.
   Литература: (ISL2024/2021, раздел 6.2), (probML2022, раздел 4.5)

6. **Ядерные методы (регрессия).** Двойственная задача в методе наименьших модулей (у шума распределение Лапласа) с нелинейными признаками, ядерная функция в результате её решения. Квадратичное программирование. Примеры ядерных функций, их свойства, работа с бесконечным количеством признаков.
   Зачем: прелюдия к SVM.
   Проблемы: это «нестандартные» модели, но они есть в sklearn (SVR с какими-то конкретными гиперпараметрами); немного более простая задача — с методом наименьших квадратов, но там надо считать обратную матрицу, это запарно.
   Литература: (PRML2006/2020, sec. 6)

7. **Ядерные методы (классификация).** Метод опорных векторов. Формулировка для классификации на 2 класса (через линал, максимизацию расстояния между разделяющей плоскостью и облаками данных). Hard SVM (для линейно разделимых классов), soft SVM (для линейно неразделимых классов, допускает ошибки). Интерпретация функции потерь как регуляризованной hinge loss. Оптимизационная задача, методы решения: градиентный спуск, квадратичное программирование.
   Семинар: опять fit-predict, меняется пара строк кода по ср. с темой 3; скучновато? мб подбор гиперпараметров на кросс-валидации?
   Литература: (MLPP2012, sec. 14.5), (ISL2024/ISL2021, гл. 9), (PRML2006/2020, sec. 7)

8. **Кластеризация.** Постановка задачи, алгоритмы K-means, DBSCAN. Конечные смеси нормальных распределений, EM-алгоритм для поиска оценок смесей методом максимального правдоподобия.

Семинар: датасеты для кластеризации и для смесей (Old Faithful?), реализация k-means вручную, k-means и смеси в sklearn.
Литература: (ISL2024/ISL2021, раздел 12.4), (PRML2006/2020, sec. 9), (probML2022, sec. 21), (MLPP2012, sec. 11)

9. **Снижение размерности.** Постановка задачи, алгоритмы: Principal Component Analysis, факторизация матриц. Использование преобразованных эмбеддингов для дальнейшего обучения моделей.
Семинар: датасет с большим количеством признаков.
Литература: (ISL2024/ISL2021, раздел 12.2), (probML2022, sec. 20)

10. **Деревья решений.** Бинарные деревья как последовательность инструкций if/else. Невозможность построения дерева градиентными методами. Алгоритмы построения (локально) оптимальных деревьев, CART. Интерпретация «деревянных» моделей.
Семинар: построение деревьев, визуализация, интерпретация, подбор гиперпараметров кросс-валидацией.
Литература: (ISL2024/ISL2021, гл. 8), (probML2022, sec. 18).

11. **Бутстрэп и ансамбли моделей.** Бутстрэп как метод аппроксимации sampling distribution оценок, альтернатива асимптотическим подходам. Бутстрэп доверительные интервалы, бутстрэп прогнозы, снижение дисперсии оценок с помощью бутстрэпа (Bootstrap AGGregatING). Stacking. Случайный лес.
Семинар: построение бутстрэп доверительных интервалов для параметров логистической регрессии, построение случайных лесов.
Литература: (PRML2006/2020, sec. 14), (probML2022, раздел 4.7.3, гл. 18).

12. **Бустинг.** Бустинг для произвольных моделей, для деревьев решений. Бустинг со среднеквадратичной функцией потерь. Связь бустинга и градиентного спуска.
Литература: (probML2022, sec. 18.5).

2. **Фонды оценочных средств результатов обучения**

Результаты обучения по дисциплине	Оценочные средства
ОПК-5.И-1.У-1.	Домашние задания, проект, экзамен
ПК-8.И-2.У-1.	Домашние задания, проект, экзамен
ПК-8.И-2.У-2.	Домашние задания, проект, экзамен
МПК-1.И-1.У-1.	Домашние задания, проект, экзамен
МПК-1.И-1.У-2.	Домашние задания, проект, экзамен
МПК-1.И-1.У-3.	Домашние задания, проект, экзамен
МПК-4.И-1.У-1.	Домашние задания, проект, экзамен
МПК-4.И-1.У-2.	Домашние задания, проект, экзамен

8. **Балльная система оценки**

Максимальные значения баллов, которые студент может получить за выполнение формы проверки знаний (текущая и промежуточная аттестация):

Формы текущей и промежуточной аттестации (оценочные средства)	Баллы
Курсовой проект	100
Домашние задания	140
Экзамен	60
Итого	**300**

Оценка по курсу выставляется, исходя из следующих критериев:

Оценка	Минимальное количество баллов	Максимальное количество баллов
Отлично	255	300
Хорошо	195	254.9
Удовлетворительно	120	194.9
Неудовлетворительно	60	119.9

Пример домашнего задания:

Используемые данные находятся в файле price.csv. Выберите итоговую метрику. Что касается эмпирического распределения показателей, проверьте с помощью соответствующих тестов, что оно статистически отличается от нормального. На основе результата проведите сравнение средних значений в двух группах для выбранной метрики. Напишите краткий отчет с окончательным решением по ценовой политике компании N. Отправьте эту работу в Jupiter Notebook.

Пример задания из экзамена.

Для бинарного классификатора с порогом 0.5 и заданной выборки (**y**, **p**), где **y** – истинные ответы, а **p** – соответствующие оценки вероятностей принадлежности позитивному классу, выданные классификатором.

1. Вычислить:
   - accuracy
   - precision
   - recall
   - f1-меру
   - auc roc
2. Построить roc-кривую.

Y	p
1	0.6
0	0.25
1	0.25
0	0.9
0	0.8
1	0.9
0	0.6
0	0.4

0	0.3
1	0.7

Пример задания на проект.
Проект представляет собой краткий текст (до 15 страниц), содержащий постановку задачи, краткий обзор литературы, анализ теоретической модели (или моделей) с предпосылками и выводами, а также подборку примеров из эмпирических статей, иллюстрирующих выводы модели и собственные расчёты. Требуется подбор данных и их описание. Обучающие данные должны быть взяты из открытых источников.
Задание выполняется в группах (2–3 человека) или индивидуально.
Работа должна быть представлена в формате Jupiter Notebook. Рекомендуется использовать сервис Яндекс.DataSphere.

Требования к выполнению заданий:

Домашние задания представляют собой практические задания, направленные на закрепление навыков машинного обучения и анализа данных на Python по различным темам. Задание предоставляется в формате Jupyter-notebook .ipynb с кодом, комментариями и ответами на вопросы задания. В комментариях необходимо описать методы и данные, использованные для решения задачи, подробно обосновать выбор алгоритма и представить результаты расчётов (при необходимости с использованием таблиц и рисунков).

Проект представляет собой небольшой текст (до 15 страниц), содержащий постановку задачи, краткий обзор литературы, анализ теоретической модели (или моделей) с предпосылками и выводами, а также подборку примеров из эмпирических статей, иллюстрирующих выводы модели и собственные расчёты. Выбор данных и их описание обязательны. Обучающие данные должны быть взяты из открытых источников. Задание выполняется в группах (2–3 человека) или индивидуально. Работа должна быть представлена в Jupiter Notebook. Рекомендуется использовать сервис Yandex DataSphere.

Информационное обеспечение дисциплины

a. Основная литература

i. (ISL2024) Джеймс Г., Уиттен Д., Хасти Т., Тибширани Р., Тейлор Дж. Введение в статистическое обучение с примерами на языке Python / пер. с англ. А.Ю. Гинько. -М.: ДМК Пресс, 2024. - 846 с.: ил.

ii. (PRML2020) Бишоп, Кристофер М. Распознавание образов и машинное обучение. : Пер. с англ. — СПб. : ООО «Диалектика», 2020. — 960 с. : ил. — Парал. тит. англ.

iii. (probML2022) Мэрфи К.П. Вероятностное машинное обучение: введение / пер. с англ. А.А. Слинкина. - М.: ДМК Пресс, 2022. - 940 с.: ил.

iv. (ISL2021) James G., Witten D., Hastie T., Tibshirani R., «An Introduction to Statistical Learning with Applications in R», Springer, 2021. https://doi.org/10.1007/978-1-0716-1418-1

v. (PRML2006) Bishop C.M., «Pattern Recognition and Machine Learning», Springer, 2006. ISBN-10: 0-387-31073-8.

vi. (MLPP2012) Murphy K.P., «Machine Learning: a Probabilistic Perspective», The MIT Press, 2012. ISBN 978-0-262-01802-9.

vii. (probML2022) Murphy K.P., «Probabilistic Machine Learning: an Introduction», The MIT Press, 2022.

viii. (EDO2021) Joaquim R.R.A. Martins and Andrew Ning. Engineering Design Optimization. Cabridge University Press, 2021. ISBN: 9781108833417. https://mdobook.github.io

ix. (NumOpt2006) Nocedal J., Wright S.J. «Numerical Optimization», Springer, 2006. ISBN-10: 0-387-30303-0.

b. Документация:

i. Polars: https://docs.pola.rs/api/python/stable/reference/index.html

ii. NumPy: https://numpy.org/doc/

iii. Matplotlib: https://matplotlib.org/

iv. scikit-learn: https://scikit-learn.org/stable/index.html

9. **Материально-техническое обеспечение дисциплины**

Описание материально-технической базы

Для организации занятий по дисциплине необходимы следующие технические средства обучения: компьютерный класс с установленным проектором, доской, маркерами. Для подключения к сервисам Yandex DataSphere студентам необходимо завести электронный ящик на @yandex.ru

Язык преподавания:

Английский

Преподаватель (преподаватели):

Иванов Михаил Алоизович (ассистент кафедры ММАЭ), Машин Иван Сергеевич, руководитель группы разработки, ООО Салютдевайсы (внешний преподаватель)

Разработчики программы:

Иванов Михаил Алоизович (ассистент кафедры ММАЭ), Машин Иван Сергеевич, руководитель группы разработки, ООО Салютдевайсы (внешний преподаватель)