



Национальный исследовательский университет «Высшая школа экономики»
Программа дисциплины «Машинное обучение» для направления 38.03.01 «Экономика» образовательных программ «Экономика» и «Совместная программа по экономике НИУ ВШЭ и РЭШ» подготовки бакалавра

**Федеральное государственное автономное образовательное учреждение
высшего образования**

**"Национальный исследовательский университет
"Высшая школа экономики"**

Факультет Компьютерных наук
Департамент больших данных и информационного поиска

**Рабочая программа дисциплины
Машинное обучение**

для образовательной программы «Экономика»
направления 38.03.01 «Экономика»,
и образовательной программы «Совместная программа
по экономике НИУ ВШЭ и РЭШ»
направления 38.03.01 «Экономика»
уровень бакалавр

Разработчики программы

Артемов А.В., кандидат физ.-мат. наук, доцент, aartemov@hse.ru

Лобачева Е.М., старший преподаватель, elobacheva@hse.ru

Филатов А.А., преподаватель, filatovartm@gmail.com

Одобрена на заседании департамента больших данных и информационного поиска
«__» _____ 2017 г.

Руководитель департамента

В.В.Подольский _____ [подпись]

Утверждена Академическим советом образовательной программы

«__» _____ 2017 г., № протокола _____

Академический руководитель образовательной программы

К.А. Букин _____ [подпись]

К.А. Паниди _____ [подпись]

Москва, 2017



Национальный исследовательский университет «Высшая школа экономики»
Программа дисциплины «Машинное обучение» для направления 38.03.01 «Экономика» образовательных программ «Экономика» и «Совместная программа по экономике НИУ ВШЭ и РЭШ»
подготовки бакалавра

Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.



1 Область применения и нормативные ссылки

Настоящая программа учебной дисциплины устанавливает требования к образовательным результатам и результатам обучения студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих дисциплину Машинное обучение, учебных ассистентов и студентов направления подготовки/специальности 38.03.01 «Экономика», обучающихся по образовательным программам «Экономика» и «Совместная программа по экономике НИУ ВШЭ и РЭШ».

Программа учебной дисциплины разработана в соответствии с:

- ФГОС ВО/Образовательным стандартом НИУ ВШЭ по направлению подготовки 38.03.01 «Экономика»;
- Образовательной программой 38.03.01 «Экономика».
- Объединенным учебным планом университета по образовательной программе «Экономика», утвержденным в 2016г.

2 Цели освоения дисциплины

Целями освоения дисциплины «Машинное обучение» являются:

- Ознакомление студентов с теоретическими основами и основными принципами машинного обучения — а именно, с классами моделей (линейные, логические, нейросетевые), метриками качествами и подходами к подготовке данных.
- Формирование у студентов практических навыков работы с данными и решения прикладных задач анализа данных.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

[Компетенции для программы учебной дисциплины берутся из стандартов: ФГОС/ ОС НИУ ВШЭ для соответствующего уровня и направления подготовки и из числа закрепленных за дисциплиной в матрице компетенций образовательной программы]

Уровни формирования компетенций:

РБ — ресурсная база, в основном теоретические и предметные основы (знания, умения);

СД – способы деятельности, составляющие практическое ядро данной компетенции;

МЦ – мотивационно-ценностная составляющая, отражает степень осознания ценности компетенции человеком и готовность ее использовать

В результате освоения дисциплины студент осваивает компетенции:

Компетенция	Код по ОС ВШЭ	Уровень формирования компетенции	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции	Форма контроля уровня сформированности компетенции
	УК-1	РБ СД			
	УК-6	РБ			
	ПК-11	РБ СД			
	ПК-12	СД МЦ			
	ПК-17	РБ СД МЦ			



4 Место дисциплины в структуре образовательной программы

Для специализации «Экономика» настоящая дисциплина является дисциплиной по выбору.

Для освоения учебной дисциплины студенты должны владеть знаниями и компетенциями следующих дисциплин:

- Математический анализ
- Линейная алгебра и геометрия
- Теория вероятностей и математическая статистика
- Основы программирования и информатики

Основные положения дисциплины должны быть использованы в дальнейшем при написании КР и ВКР, предусмотренных учебным планом.

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Введение в машинное обучение	12	2	2	8
2	Линейные модели	84	14	14	56
3	Решающие деревья и композиции	36	6	6	24
4	Нейронные сети и обучение без учителя	24	4	4	16
5	Модели временных рядов	12	2	2	8
	ИТОГО	170	28	28	114

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры
		1	2	
Текущий	Домашнее задание	3	2	Практическое задание: нужно запрограммировать метод, провести эксперименты и сделать выводы
	Соревнование по анализу данных		*	Соревнование включает в себя решение прикладной задачи. Оценка зависит от качества решения, в том числе и от места в рейтинге.
	Самостоятельные работы	*	*	Самостоятельная работа в начале каждого семинара: включает в себя несколько вопросов на понимание последней лекции + решение задач, выданных на дом недель ранее.
Промежуточный	Коллоквиум	*		Устный
Итоговый	Экзамен		*	Устный



7 Критерии оценки знаний, навыков

В курсе предусмотрено несколько форм контроля знания:

- Самостоятельные работы на семинарах, проверяющие знание основных фактов с лекций и выполнение теоретических домашних заданий
- Практические домашние работы на Python, формирующие навыки работы с основными инструментами анализа данных, а также помогающие освоить основные концепции машинного обучения
- Соревнование по анализу данных, направленные на приобретения опыта решения практических задач анализа данных
- Устный коллоквиум в конце 1-го модуля
- Устный экзамен

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

8 Содержание дисциплины

1. Введение в машинное обучение (1 лекция, 1 семинар)

Введение. Постановки основных классов задач в машинном обучении: классификация, регрессия, ранжирование, кластеризация, оценка скрытого состояния модели. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук, логи. Признаки.

2. Линейные методы регрессии (2 лекции, 2 семинара)

Аналитическое и численное решение задачи МНК. Градиентный спуск, методы оценивания градиента. Функции потерь. Метрики качества регрессии. Регуляризация. Методы оценивания обобщающей способности, кросс-валидация.

3. Линейные методы классификации (2 лекции, 2 семинара)

Аппроксимация эмпирического риска. Задача оценивания вероятностей, логистическая регрессия. Идея калибровки вероятностей. Перцептрон. Метод опорных векторов, его двойственная задача (без ядер). Обобщённые линейные модели. Метрики качества в задачах классификации. Постановки задач multiclass- и multilabel-классификации.

4. Особенности работы с реальными данными (1 лекция, 1 семинар)

Пропуски в данных. Предобработка признаков. Чистка данных. Категориальные признаки: кодирование, хэширование, счётчики. Работа с текстами. Разреженные признаки: векторизация, хэширование, TF-IDF. Косинусная метрика.

5. Работа с признаками (2 лекция, 2 семинара)

Методы отбора признаков. Метод главных компонент и singular spectrum analysis. Ядровые методы. Ядра и спрямляющие пространства, методы их построения. Операции в спрямляющих пространствах. Ядра в SVM и PCA.

6. Решающие деревья (1 лекция, 1 семинар)

Общий алгоритм построения, критерии информативности. Конкретные критерии для классификации и регрессии. Тонкости решающих деревьев: обработка пропущенных значений, стрижка, регуляризация.

7. Композиции алгоритмов (2 лекции, 2 семинара)

Общая идея bias-variance decomposition. Бэггинг и метод случайных подпространств. Случайные леса и extra random trees.

Бустинг. Градиентный бустинг над решающими деревьями.

8. Введение в нейронные сети и глубинное обучение (1 лекция, 1 семинар)



Структура нейронной сети. Обратное распространение ошибки. Применение нейросетей для анализа изображений: свёрточные слои, примеры архитектур как наборов кубиков.

9. Введение в методы кластеризации и снижения размерности данных (1 лекция, 1 семинар)

Задача кластеризации. K-Means. Визуализация и t-SNE.

10. Введение в методы моделирования и прогнозирования временных рядов (1 лекция, 1 семинар)

Прогнозирование временных рядов как задача регрессии: авторегрессия, тренды и сезонности. Оценивание качества скользящим окном.

9 Оценочные средства для текущего контроля и аттестации студента

Примеры экзаменационных вопросов

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.

2. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.

3. Метрики качества алгоритм регрессии и классификации.

4. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.

5. Деревья решений. Методы построения деревьев. Их регуляризация.

6. Композиции алгоритмов. Разложение ошибки на смещение и разброс.

7. Случайный лес, его особенности.

8. Градиентный бустинг, его особенности при использовании деревьев в качестве базовых алгоритмов.

9. Нейронные сети. Метод обратного распространения ошибок. Свёрточные сети.

10. Кластеризация. Алгоритм K-Means.

10 Порядок формирования оценок по дисциплине

Результующая оценка по дисциплине рассчитывается по формуле

$$O_{\text{итог}} = 0.7 O_{\text{накопл}} + 0.3 O_{\text{экз}}$$

Накопленная и итоговая оценки округляются арифметически.

Накопленная оценка рассчитывается по формуле

$$O_{\text{накопл}} = 0.2 O_{\text{самост}} + 0.6 O_{\text{дз}} + 0.2 O_{\text{коллоквиум}}$$

Оценка за самостоятельную работу вычисляется как сумма баллов по всем самостоятельным, переведенная в 10 бальную шкалу. Оценка за домашнюю работу — как сумма баллов по всем практическим заданиям и соревнованию, переведенная в 10 бальную шкалу. Количество баллов за разные задания может различаться в зависимости от их сложности. Все промежуточные оценки (за домашние, самостоятельные и коллоквиум) могут быть не целыми. Накопленная и итоговая оценки округляются математически.

Дедлайны по всем домашним заданиям являются жёсткими, то есть после срока работы не принимаются. При обнаружении плагиата оценки за домашнее задание обнуляются всем задействованным в списывании студентам, а также подаётся докладная записка в деканат.

При наличии уважительной причины пропущенную проверочную можно написать позднее, а дедлайн по домашнему заданию может быть перенесён (при этом получить дополнительные баллы за призовые места на конкурсе можно только при участии в общий срок).



По курсу возможно получение оценки автоматом без сдачи экзамена. Необходимым условием для получения автомата является накопленная оценка, равная 8 или выше.

11 Учебно-методическое и информационное обеспечение дисциплины

11.1 Базовые учебники

1. Hastie T., Tibshirani R, Friedman J. [The Elements of Statistical Learning \(2nd edition\)](#). Springer, 2009.
2. Bishop C. M. [Pattern Recognition and Machine Learning](#). Springer, 2006.

11.2 Дополнительная литература

1. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. MIT Press, 2012.
2. Murphy K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
3. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.