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 «Математика»
направления 01.03.01 «Математика»,
и образовательной программы «Совместный бакалавриат ВШЭ и ЦПМ»
направления 01.03.01 «Математика»
уровень бакалавр

Разработчики программы
Щуров Илья Валерьевич ischurov@hse.ru

Одобрена на заседании департамента больших данных и информационного поиска
«___»_____ 2018 г.

Руководитель департамента
В.В.Подольский _____ [подпись]

Утверждена Академическим советом образовательной программы
«___»_____ 2018 г., № протокола_____

Академический руководитель образовательной программы
К.А. Букин _____ [подпись]
К.А. Паниди _____ [подпись]

Москва, 2018

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 01.03.01 «Математика», обучающихся по образовательным программам «Математика» и «Совместный бакалавриат ВШЭ и ЦПМ».

Программа учебной дисциплины разработана в соответствии с:

- ФГОС ВО/Образовательным стандартом НИУ ВШЭ по направлению подготовки 01.03.01 «Математика»;
- Образовательной программой 01.03.01 «Математика».
- Объединенным учебным планом университета по образовательной программе «Математика», утвержденным в 2018 г.

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

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:

Код по ОС ВШЭ	Уровень формирования компетенции	Описание компетенции
УК-1	РБ СД	Способен учиться, приобретать новые знания, умения, в том числе в области, отличной от профессиональной
УК-6	РБ	Способен вести исследовательскую деятельность, включая анализ проблем, постановку целей и задач, выделение объекта и предмета исследования, выбор способа и методов исследования, а также оценку его качества
ПК-11	РБ СД	Способен осуществлять сбор, анализ и обработку статистических данных, информации, научно-аналитических материалов, необходимых для решения поставленных экономических задач
ПК-12	СД МЦ	Способен выбрать инструментальные средства для обработки экономических данных в соответствии с поставленной задачей, проанализировать результаты расчетов и обосновать полученные выводы

ПК-17	РБ СД МЦ	Способен использовать для решения аналитических и исследовательских задач современные технические средства и информационные технологии

4 Место дисциплины в структуре образовательной программы

Для специализации «Математика» 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	16	4	4	8
2	Линейные модели	32	4	4	24
3	Признаковые представления	32	4	4	24
3	Решающие деревья и композиции	32	4	4	24
5	Нейронные сети	42	8	8	26
4	Кластеризация и методы снижения размерности	16	4	4	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 неделей ранее.

Промежуточный	Контрольная работа	*		Письменный
Итоговый	Экзамен		*	Письменный

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
- Контрольная работа в конце 1-го модуля
- Письменный 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)

Постановки основных классов задач в машинном обучении. Обучение с учителем (supervised learning): регрессия и классификация; обучение без учителя (unsupervised learning): кластеризация, снижение размерности; semi-supervised learning, рекомендательные системы, обработка текстов: тематическое моделирование, построение аннотаций, извлечение ответов на вопросы, машинный перевод; обработка изображений: порождение, преобразование; обучение представлений; обучение с подкреплением. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук, логи. Признаки.

2. **Статистические оценки и проверка гипотез** (1 лекция, 1 семинар)

Основные понятия математической статистики: статистические оценки (точечные и интервальные), их свойства, проверка гипотез.

3. **Машинное обучение как математическое моделирование** (1 лекция, 1 семинар)

Статистические модели. Теоретико-вероятностная постановка задачи обучения с учителем. Минимизация ожидаемой ошибки. No free lunch theorem. Пример: задача регрессии, минимизация квадрата отклонения. Регрессионная функция: условное матожидание. Линейная регрессия и метод k ближайших соседей. Переобучение и недообучение. Разложение ошибки на шум, смещение и разброс. Проклятие размерности. Методы оценивания обобщающей способности, кросс-валидация.

4. **Введение в линейные модели и задача регрессии** (1 лекция, 1 семинар)

Градиентный спуск, методы оценивания градиента. Функции потерь. Метрики качества регрессии. Линейная регрессия, метод наименьших квадратов и максимизация правдоподобия.

Теорема Гаусса—Маркова. Явный вид решения в методе наименьших квадратов. Ковариационная матрица для коэффициентов. Практические соображения: что делать с категориальными данными? Вычислительные соображения: точное решение vs градиентный спуск. Регуляризация.

5. **Линейные модели и задача классификации** (1 лекция, 1 семинар)
Задачи классификации. Общая постановка. 0-1 ошибка. Байесовский классификатор. Линейные методы для классификации. Логистическая регрессия, максимизация правдоподобия, кросс-энтропия.

6. **Выбор и оценка моделей, работа с признаками** (1 лекция, 1 семинар)
Кросс-валидация: тонкости (отбор переменных, переобучение на валидационное множество). Оценки ожидаемой ошибки для линейной регрессии: AIC и другие. L1 и L2 регуляризация. Методы отбора признаков. Метод главных компонент и singular spectrum analysis. Ядровые методы. Ядра и спрямляющие пространства, методы их построения. Операции в спрямляющих пространствах.

7. **Деревья и ансамбли** (1 лекция, 1 семинар)
Ограничения линейных методов (пример: XOR). Решающие деревья. CART. Ансамбли. Бутстреп. Бэггинг. Случайный лес.

8. **Бустинг** (1 лекция, 1 семинар)
AdaBoost, градиентный бустинг. XGBoost.

9. **Признаковые представления для дискретных входных данных** (1 лекция, 1 семинар)
Практические соображения. Кодирование категориальных данных. Пропущенные значения. Обработка текстов: bag of words, tf-idf, векторные эмбеддинги.

10. **Введение в нейросети** (1 лекция, 1 семинар)
Нейронные сети: общая архитектура. Реализация XOR с помощью трёх персептронов. Теорема об универсальной аппроксимации. Многослойные сети. Обратное распространение ошибки. Стохастический градиентный спуск. Проблемы: затухающие и взрывающиеся градиенты, невыпуклость функции потерь.

11. **Современные нейросетевые архитектуры** (2 лекции, 2 семинара)
Нейронные сети в обработке изображений. Фильтры. Сверточные слои. Нейронные сети и обучение представлений. Обработка последовательностей. Рекуррентные нейронные сети.

12. **Кластеризация** (1 лекция, 1 семинар)
K-means. EM-алгоритм. Другие методы кластеризации: иерархическая кластеризация, DBSCAN, Affinity Propagation.

13. **Снижение размерности** (1 лекция, 1 семинар)
SVD-разложение. Метод главных компонент. t-SNE, UMAP.

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

Результирующая оценка по дисциплине рассчитывается по формуле:

$$O_итог = 0.7 * O_накопл + 0.3 * O_экз$$

Накопленная и итоговая оценки округляются арифметически, оценки вида «целое + 1/2» округляются вверх.

Накопленная оценка рассчитывается по формуле:

$$O_накопл = 0.2 * O_самост + 0.4 * O_дз + 0.4 * O_кр$$

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.

11 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).

12 Учебно-методическое и информационное обеспечение дисциплины

12.1 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.

12.2 Дополнительная литература

1. Ian Goodfellow and Yoshua Bengio and Aaron Courville, Deep Learning, MIT Press, 2016.
2. Mohri M., Rostamizadeh A., Talwalkar A. Foundations of Machine Learning. MIT Press, 2012.
3. Murphy K. Machine Learning: A Probabilistic Perspective. MIT Press, 2012.
4. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014.