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
Методы машинного обучения

для образовательной программы «Прикладная политология»
направления подготовки 41.04.04
уровень магистр

Разработчик программы
Шестаков А.В., старший преподаватель, avshestakov@hse.ru

Одобрена на заседании департамента больших данных и информационного поиска
«___» _____ 201_ г.

Зав. Департамента
В.В. Подольский _____ [подпись]

Утверждена Академическим советом образовательной программы
«___» _____ 201_ г., № протокола_____

Академический руководитель образовательной программы
А.С. Ахременко _____ [подпись]

Москва, 2018

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 «Методы машинного обучения», учебных ассистентов и студентов направления подготовки/специальности 41.04.04 «Политология», обучающихся по образовательной программе «Прикладная политология».

Программа учебной дисциплины разработана в соответствии с:

- ФГОС ВО/ Образовательным стандартом НИУ ВШЭ по направления подготовки 41.04.04 «Политология»;
- Образовательной программой 41.04.04 «Политология»;
- Объединенным учебным планом университета по образовательной программе 41.04.04 «Политология» утвержденным в 2018 г.

2 Цели освоения дисциплины

Целями освоения дисциплины «Методы машинного обучения» являются:

- Ознакомление студентов с основными принципами машинного обучения – а именно, видами задач машинного обучения, классами моделей, способами обработки различных типов данных и измерение качества результатов
- Формирование у студентов практических навыков работы с данными и решения прикладных задач анализа данных с помощью современного программного обеспечения.

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:

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
Общепрофессиональные компетенции	ОПК-2	РБ СД	Способен осуществлять поиск информации и решать задачи профессиональной деятельности с помощью современных информационно-коммуникационных технологий и программных средств		
Общепрофессиональные	ОПК-7	РБ СД	Способен разрабатывать предложения и рекоменда-		

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
компетенции			ции для проведения прикладных исследований и консалтинга		
Профессиональные компетенции	ПК-3	РБ СД	Способен использовать в научной и проектной деятельности современные базы эмпирических данных (в том числе зарубежные), самостоятельно создавать базы данных для реализации исследовательских и практических задач		
Профессиональные компетенции	ПК-4	РБ СД	Способен осуществлять анализ эмпирических данных (политических, экономических и социологических исследований) с помощью современных качественных и количественных методов и с использованием соответствующего программного обеспечения		
Профессиональные компетенции	ПК-5	РБ СД	Способен строить и анализировать математические модели социально-политических систем и процессов		
Профессиональные компетенции	ПК-11	РБ СД	Способен осуществлять самостоятельную подготовку обобщающих аналитических материалов (обзоров, записок, докладов, отчетов, рекомендаций и др.) и предложений для лиц, принимающих решения в политической сфере		

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к дисциплинам по выбору.

Для освоения учебной дисциплины студенты должны владеть следующими знаниями и компетенциями:

- «Математика и статистика»
- «Основы программирования на Python»

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы				Самостоятельная работа
			Лекции	Семинары	Практические занятия	Другие виды работы[1]	
1	Введение в машинное обучение и анализ данных	27	5	5			17
2	Базовые модели классификации и регрессии	27	5	5			17
3	Методы обучения без учителя: задача кластеризации и понижения размерности	33	6	6			21
4	Модели и методы работы с текстовыми данными.	33	6	6			21
5	Методы сбора и анализа сетевых данных	21	4	4			13
6	Введение в рекомендательные системы	11	2	2			7
	Итого	152	28	28			96

[1] Указать другие виды аудиторной работы студентов, если они применяются при изучении данной дисциплины.

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год				Параметры **
		1	2	3	4	
Текущий	Домашнее задание		*	*		Практическая работа на языке программирования.
	Самостоятельные работы		*	*		Тест по материалам предыдущих лекций.
	Индивидуальный или групповой проект			*		Самостоятельная работа студентов для решения практической задачи с использованием изученных методов. На одном из занятий студенты презентуют проекты.
Промежуточный	Коллоквиум		*			Письменная работа (120 минут)
Итоговый	Экзамен			*		Письменная работа (120 минут)

7 Критерии оценки знаний, навыков

В курсе предусмотрено несколько форм контроля знаний:

- Практические домашние работы с элементами программирования, направленные на формирование навыков работы с данными и их анализом, а также помогающие освоить изученный материал
- Самостоятельные работы в виде тестов для проверки усваивания материалов прошедших лекций
- Групповой или индивидуальный проект с поэтапным исполнением в течение всего курса. Студенты должны сформулировать исследовательскую задачу, при необходимости собрать необходимые данные, определиться с методами решения, провести анализ результатов.
- Письменный коллоквиум в конце 2-го модуля
- Письменный экзамен

Оценки по всем формам текущего контроля осуществляются по 10-ти бальной шкале

8 Содержание дисциплины

<u>Раздел 1. Введение в машинное обучение и анализ данных</u>

Введение. Что такое анализ данных, машинное обучение, Data Science, KDD. Мотивация к изучению данного направления. Основных классов задач с машинном обучении: классификация, регрессия, ранжирование, понижение размерности. Основные типы методов: обучение с

учителем, обучение без учителя, обучение с подкреплением, частичное обучение. Типы данных и методы их обработки. Способы оценки качества результатов моделей.

Раздел 2. Основные модели и методы классификации и регрессии
Метрические методы, деревья решений и линейные модели. Функции потерь и методы оптимизации (градиентный спуск). Оценка обобщающей способности алгоритмов. Регуляризация.

Раздел 3. Методы обучение без учителя: задача кластеризации и понижения размерности
Постановка задачи кластеризации. Группы алгоритмов кластеризации. Метод К-средних, иерархическая кластеризация, DBSCAN, EM-алгоритм для смеси распределений. Оценка качества кластеризации. Задача понижения размерности признаков и отбор признаков. Метод главных компонент, многомерное шкалирование и T-SNE.

Раздел 4. Модели и методы работы с текстовыми данными
Представление текстовых данных. Методы предобработки текстовых данных – стэмминг, лемматизация, нормализация. Выявление ключевых слов и словосочетаний. Модель мешка слов, TF-IDF. Тематические модели текстов – PLSA, LDA, аддитивные тематические модели. Дистрибутивная семантика и эмбеддинги – word2vec, fasttext.

Раздел 5. Методы сбора и анализа сетевых данных
Представление информации в виде графа (сети). Основные характеристики сетевых структур, дескриптивные статистики элементов сети. Меры центральности и выявление сообществ в сети. Выгрузка информации из социальных сетей.

Раздел 6. Введение в рекомендательные системы
Постановка задач рекомендательных систем. Коллаборативная фильтрация. Методы основанные на скрытых переменных. Оценка качества рекомендательных систем.

9 Оценочные средства для текущего контроля и аттестации студента

9.1 Примеры экзаменационных вопросов

1) В таблице даны попарные расстояние между объектами из обучающей выборки. Выполните иерархическую кластеризацию с использованием complete-linkage расстоянием между кластерами
2) Выпишите формулу критерия silhouette. Дайте определение каждой составляющей. Приведите пример разбиения объектов на кластеры, при котором у некоторых объектов будет отрицательные силуэт
3) Что такое мультиколлинеарность, как она влияет на линейные модели. По данному набору признаков запишите вид уравнения регрессии избежав при этом строгой мультиколлинеарности
4) Дайте определение F-меры для оценки качества классификации. Почему эта мера лучше чем min(Precision, Recall)?
5) Для данной таблицы с результатами модели нарисуйте ROC кривую и посчитайте ROC-AUC
6) Что такое коллаборативная фильтрация? В чем отличие Item-based и User-based подходов?

7) Опишите метод T-SNE. Объясните, чем он отличается от методов многомерного шкалирования. Какие ошибки можно допустить при его использовании.
8) По данному набору данных, известным собственным векторам и собственным числам корреляционной матрицы выполните переход в сжатое признаковое пространство с помощью метода главных компонент.
9) По данной сети посчитайте степенную центральность вершин. Почему для неориентированных графов степенная центральность будет очень сильно коррелировать с PageRank?
10) Что такое модулярность? Для данной сети выделите 3 сообщества с помощью метода EdgeBetwenneess. Посчитайте модулярность.
11) Какая гипотеза заложена в метрических методах машинного обучения? Почему необходимо нормировать признаки перед применением метода ближайшего соседа. Что делать, если исходные признаки имеют разный тип?
12) Что такое нормализация, лемматизация и стэмминг. Их плюсы и минусы?
13) Опишите модель мешка слов для представления текстов. Как TF-IDF может помочь в задаче классификации текстов?
14) Что такое суррогатный сплит в деревьях решений? Каким образом можно вводить регуляризацию в деревьях?
15) Что такое L1 и L2 регуляризация в линейных моделях? При какой из них происходит более агрессивное размежевание коэффициентов модели?

10 Порядок формирования оценок по дисциплине

Результирующая оценка по дисциплине рассчитывается по формуле

О_рез = 0.7 О_накопл + 0.3 О_экз

Накопленная оценка по дисциплине рассчитывается по формуле:

О_накоп = 0.1 О_самост + 0.6 О_дз + 0.3 О_коллоквиум

Оценка за самостоятельную работу считается как сумма баллов по всем самостоятельным, переведенная в 10-ти бальную шкалу. Оценка за домашнюю работу – как сумма баллов по всем практическим домашним заданиям и групповому проекту, переведенную в 10-ти бальную шкалу. Все промежуточные оценки могут быть нецелыми. Накопленная и результирующая оценки округляются математически.

Сроки выполнения домашних заданий являются жесткими если не указано обратного. При обнаружении плагиата оценки обнуляются всем вовлеченным сторонам.

При наличии уважительной причины, пропущенную проверочную или домашнюю работу можно сдать позже в установленный преподавателем срок.

По курсу возможно получение оценки автоматом без сдачи экзамена, при условии накопленной оценки более 8 баллов после округления (если не указано иного)

11 Учебно-методическое и информационное обеспечение дисциплины

11.1 Базовый учебник

1) Elements of Statistical Learning, Hastie T., Timshirani R., and Friedman J.
2) Data Mining and Analysis: Fundamental Concepts and Algorithms, Mohammed J. Zaki, Wagner Meira Jr.

3) Mining of Massive Datasets, Leskovec J. Rajaraman A., Ullman J.D.

11.2 Дополнительная литература

1) Core Concepts in Data Analysis: Summarization, Correlation, Visualization, Mirkin B.
2) Recommender Systems Handbook, Ricci F., Rokach L., Bracha S., Kantor P.
3) Speech and Language Processing, Jurafsky D., Martin J.
4) Neural Network Methods in Natural Language Processing, Goldberg Y.
5) Networks: An Introduction, Newman M.
6) Introduction to Machine Learning with Python, Muller A., Guido S.
7) Tutorial on Probabilistic Topic Modelling: Additive Regularization for Stochastic Matrix Factorization, Vorontsov K., Potapenko A.

11.3 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства:

- Python
- Jupyther Notebook