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 «Прикладная математика и информатика»
направления 01.04.02
уровень магистр

Разработчик программы
Соколов Е.А., старший преподаватель, esokolov@hse.ru

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

Руководитель департамента
В.В.Подольский _____ [подпись]

Утверждена Академическим советом образовательной программы
«___»_____ 2017 г., № протокола_____

Академический руководитель образовательной программы
А.А. Масютин _____ [подпись]

Москва, 2017

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 1, учебных ассистентов и студентов направления подготовки 01.04.02 «Прикладная математика и информатика», обучающихся по образовательной программе «Финансовые технологии и анализ данных».

Программа учебной дисциплины разработана в соответствии с:

- Образовательным стандартом федерального государственного автономного образовательного учреждения высшего профессионального образования «Национального исследовательского университета «Высшая школа экономики»;
- Образовательной программой подготовки магистра по направлению 01.04.02 «Прикладная математика и информатика»;
- Объединенным учебным планом университета по образовательной программе «Финансовые технологии и анализ данных», утвержденным в 2017г.

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

В результате освоения дисциплины студент осваивает компетенции:

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
ПК18	СД	Способен понимать и применять в исследовательской и прикладной деятельности современный математический аппарат	Аудиторная работа на семинарах, выполнение теоретических домашних заданий	Домашние задания
ПК19	СД	Способен в составе научно-исследовательского и производственного коллектива решать задачи профессиональной деятельности в соответствии с профилем подготовки, общаться с экспертами в других предметных областях	Аудиторная работа на семинарах, выполнение теоретических домашних заданий	Домашние задания
ПК-3	СД	Способен понимать, совершенствовать и применять современный математический аппарат	Аудиторная работа на семинарах, выполнение теоретических домашних заданий	Домашние задания
ПК-4	СД	Способен формализовать и алгоритмизировать поставленную задачу	Выполнение практических домашних заданий	Домашние задания
ПК-5	СД	Способен писать, оформлять, отлаживать и оптимизировать программный код	Выполнение практических домашних заданий	Домашние задания
ПК-7	МЦ РБ СД	Способен провести сбор, обработку и анализ данных с использованием существующих методов машинного обучения	Выполнение практических домашних заданий	Домашние задания
ПК-9	СД	Способен разработать и реализовать в виде программного модуля алгоритм решения поставленной теоретической или прикладной задачи на основе математической модели	Выполнение практических домашних заданий	Домашние задания

4 Место дисциплины в структуре образовательной программы

Настоящая дисциплина относится к циклу дисциплин по машинному обучению и анализу данных.

Для освоения учебной дисциплины студенты должны владеть знаниями и компетенциями по математическому анализу, линейной алгебре и теории вероятностей, которые проверяются в рамках вступительных экзаменов на магистерскую программу.

Основные положения дисциплины должны быть использованы в дальнейшем при изучении дисциплин:

- Современные методы анализа данных: Глубинное обучение.
- Современные методы принятия решений: алгоритмы обработки больших данных
- Управление данными и исполнение моделей
- Анализ текстов. Генеративные модели.

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы		Самостоятельная работа
			Лекции	Семинары	
1	Введение в машинное обучение	34	2	4	28
2	Линейные модели	44	6	10	28
3	Решающие деревья и композиции	44	6	10	28
4	Нейронные сети и обучение без учителя	58	10	20	28
5	Рекомендательные системы	48	4	12	32
	ИТОГО	228	28	56	144

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры
		1	2	
Текущий	Домашнее задание	*		Практическая работа: линейные модели и инструменты для работы с данными
	Домашнее задание	*		Теоретическая работа: линейные модели, градиентные методы обучения
	Домашнее задание		*	Практическая работа: композиции деревьев, обучение представлений и обучение без учителя
	Домашнее задание		*	Теоретическая работа: разложение ошибки на смещение и разброс, решающие деревья и их композиции, графы вычислений
Промежуточный	Коллоквиум	*		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
- Письменный коллоквиум 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)

Введение. История анализа данных. Постановки задач в машинном обучении: классификация, регрессия, ранжирование, кластеризация, латентные модели. Примеры задач. Виды данных: структурированные таблицы, тексты, изображения, звук. Признаки.

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

Аналитическое и численное решение задачи МНК. Градиентный спуск, методы оценивания градиента. Функции потерь. Регуляризация. Квантильная регрессия (постановка задачи и примеры использования). Методы оценивания обобщающей способности, кросс-валидация. Метрики качества регрессии.

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

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

Аппроксимация эмпирического риска. Персептрон. Метод опорных векторов, его двойственная задача (без ядер). Задача оценивания вероятностей, логистическая регрессия. Идея калибровки вероятностей. Оптимизация второго порядка (идея и предпосылки для использования). Обобщённые линейные модели. Метрики качества в задачах классификации.

Multiclass- и multilabel-классификация. Особенности многоклассовых задач. Метрики качества. Методы решения multilabel-задач, основанные на матричных разложениях.

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 (1 лекция, 1 семинар)

Методы отбора признаков. Метод главных компонент.

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 (3 лекции, 3 семинара)

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

Бустинг. Градиентный бустинг над решающими деревьями. Модель xgboost.

8. Нейронные сети (1 лекция, 1 семинар)

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

10. Подходы к извлечению признаков для сложных данных (1 лекция, 1 семинар)

Работа с изображениями (фильтры, извлечение признаков с помощью нейросетей), текстами (word embeddings).

11. Обучение без учителя (1 лекция, 1 семинар)

Задача кластеризации. K-Means, DBSCAN, MeanShift. Spectral clustering. Иерархическая кластеризация. Consensus clustering. Автокодировщики. Визуализация и t-SNE.

12. Рекомендательные системы (1 лекция, 1 семинар)

Постановки задачи. Метрики качества. Методы, основанные на коллаборативной фильтрации. Методы, основанные на матричных разложениях.

9 Образовательные технологии

Необходимое для выполнения домашних заданий программное обеспечение находится в свободном доступе и его можно загрузить в сети Интернет. К каждой теме студентам выдаются конспекты или слайды лекций и семинаров, а также программный код, иллюстрирующий практическое использование изучаемых методов.

10 Оценочные средства для текущего контроля и аттестации студента

10.1 Оценочные средства для оценки качества освоения дисциплины в ходе текущего контроля

Примеры домашних заданий (теоретических):

- https://github.com/esokolov/ml-course-hse/blob/master/2016-fall/homeworks-theory/homework-theory-01-linregr.pdf
- https://github.com/esokolov/ml-course-hse/blob/master/2016-fall/homeworks-theory/homework-theory-05-trees.pdf
- https://github.com/esokolov/ml-course-hse/blob/master/2016-fall/homeworks-theory/homework-theory-08-ensembles.pdf

Примеры домашних заданий (практических):

- https://github.com/esokolov/ml-course-hse/blob/master/2016-fall/homeworks-practice/homework-practice-01-linregr.ipynb
- https://github.com/esokolov/ml-course-hse/blob/master/2016-fall/homeworks-practice/homework-practice-03-ensembles.ipynb

10.2 Примеры заданий промежуточной аттестации

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.

11 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.

Накопленная оценка рассчитывается по формуле

$$O_{накопл}=0.1\ O_{самост} + 0.4\ O_{практ} + 0.3\ O_{теор} + 0.2\ O_{контрольные}$$

Оценка за домашние задания рассчитывается как среднее значение оценок за все выданные домашние задания. Оценка за самостоятельную работу рассчитывается как среднее значение оценок за все проверочные работы, проведённые на семинарских занятиях. В конце семестра разрешается переписать все самостоятельные работы, пропущенные по уважительной причине.

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. 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.