**Программа учебной дисциплины «Машинное обучение»**

Утверждена
Академическим советом ООП
Протокол № 2.9.1-12/11 от «16» мая 2017 г.

| | |
|---|---|
| Автор | Артемов А.В., кандидат физ.-мат. наук, доцент<br>Лобачева Е.М., старший преподаватель<br>Филатов А.А., преподаватель |
| Число кредитов | 5 |
| Контактная работа (час.) | 64 |
| Самостоятельная работа (час.) | 126 |
| Курс | 3 |
| Формат изучения дисциплины | без использования онлайн курса |

## I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целью освоения дисциплины «Машинное обучение» является ознакомление студентов с теоретическими основами и основными принципами машинного обучения

В результате освоения дисциплины студент должен:

- Знать основные модели и методы машинного обучения и разработки данных;
- Уметь применять указанные модели и методы, а также программные средства, в которых они реализованы;
- Владеть навыками анализа реальных данных с помощью изученных методов.

Изучение дисциплины «Машинное обучение» базируется на следующих дисциплинах:

- Математический анализ
- Линейная алгебра и геометрия
- Теория вероятностей и математическая статистика

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

- знать основы высшей математики;
- знать основы теории вероятностей и математической статистики;
- обладать базовыми навыками программирования.

## II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

### 1. Введение в машинное обучение

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

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**

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

**4. Работа с признаками**

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.

5. **Решающие деревья**

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

6. **Введение в методы кластеризации и снижения размерности данных**

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

**7. Композиции алгоритмов**

Общая идея bias-variance decomposition. Бэггинг и метод случайных подпространств. Случайные леса. Бустинг. Градиентный бустинг над решающими деревьями.

**8. Введение в нейронные сети и глубинное обучение**

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

**9. Работа с текстами**

Методы кодирования текстовых данных: векторизация, хэширование, TF-IDF. Косинусная метрика. Нейронные сети в задачах анализа текстов: рекуррентные нейронные сети.

10. **Введение в методы моделирования и прогнозирования временных рядов**

Прогнозирование временных рядов как задача регрессии: авторегрессия, тренды и сезонности. Адаптивные методы работы с временными рядами: экспоненциальное скользящее среднее, фильтр Калмана.

## III. ОЦЕНИВАНИЕ

Результирующая оценка по дисциплине рассчитывается по формуле:

О_итог = 0.7*О_накопл + 0.3*О_экз

Накопленная и итоговая оценки округляются арифметически.

Накопленная оценка рассчитывается по формуле:

О_накопл = 0.2*О_самост + 0.6*О_дз + 0.2*О_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. 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

## IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

**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.

## V. РЕСУРСЫ

### 5.1 Основная литература

1. Hastie T, Tibshirani R, Friedman JH. The Elements of Statistical Learning : Data Mining, Inference, and Prediction [Internet]. New York: Springer; 2009. (Springer Series in Statistics; vol. Second edition, corrected 7th printing). Available from: http://proxylibrary.hse.ru:2048/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=edsebk&AN=277008&site=eds-live
2. Bishop C. M. Pattern Recognition and Machine Learning [Internet]. Singapore: Springer; 2006 Available from: https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

### 5.2 Дополнительная литература

1. Mohri M, Talwalkar A, Rostamizadeh A. Foundations of Machine Learning [Internet]. Cambridge, MA: The MIT Press; 2012 [cited 2019 Jan 24]. (Adaptive Computation and Machine

Learning Series). Available from: http://proxylibrary.hse.ru:2048/login?url=http://search.ebsco-host.com/login.aspx?direct=true&db=edsebk&AN=478737&site=eds-live

### 5.3 Программное обеспечение

| № п/п | Наименование | Условия доступа |
|---|---|---|
| 1. | Microsoft Windows 7 Professional RUS<br>Microsoft Windows 10<br>Microsoft Windows 8.1 Professional RUS | *Из внутренней сети университета (договор)* |
| 2. | Microsoft Office Professional Plus 2010 | *Из внутренней сети университета (договор)* |
| 3. | Anaconda Community | *Свободно распространяемое лицензионное соглашение* |
| 4. | Python Software Foundation Python | *Свободно распространяемое лицензионное соглашение* |

**5.4**

### 5.5 Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

| № п/п | Наименование | Условия доступа |
|---|---|---|
| | ***Профессиональные базы данных, информационно-справочные системы*** | |
| 1. | Консультант Плюс | *Из внутренней сети университета (договор)* |
| 2. | Электронно-библиотечная система Юрайт | URL: https://biblio-online.ru/ |
| | ***Интернет-ресурсы (электронные образовательные ресурсы)*** | |
| 1. | Открытое образование | URL: https://openedu.ru/ |

### 5.6 Материально-техническое обеспечение дисциплины

Учебные аудитории для лекционных занятий по дисциплине обеспечивают использование и демонстрацию тематических иллюстраций, соответствующих программе дисциплины в составе:

- ПЭВМ с доступом в Интернет;
- мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ноутбуками, с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.