Программа учебной дисциплины «Машинное обучение»

Утверждена

Академическим советом ООП

Протокол № 2.03-09/2706-01 от «27» июня 2018г.

Автор	Соколов Евгений Андреевич
Число кредитов	6
Контактная работа (час.)	60
Самостоятельная работа (час.)	168
Курс	Машинное обучение
Формат изучения дисциплины	Без использования онлайн курса

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

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.

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

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

- Математический анализ
- Линейная алгебра и геометрия
- Теория вероятностей
- Математическая статистика
- Алгоритмы и структуры данных

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

- Современные методы принятия решений: Алгоритмы обработки больших данных
- Байесовские методы машинного обучения
- Современные методы анализа данных: Глубинное обучение

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

1. Введение в машинное обучение

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

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

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

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

Методы отбора признаков. Метод главных компонент.

6. Решающие деревья

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

7. Композиции алгоритмов

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

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

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

11. Обучение без учителя

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

12. Рекомендательные системы

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

III. ОЦЕНИВАНИЕ

В рамках курса предусмотрены самостоятельные работы на занятиях, теоретические домашние задания, практические домашние задания, письменная контрольная работа и письменный экзамен.

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.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры практических заданий можно найти по ссылке http://wiki.cs.hse.ru/Машинное_обучение_1#.D0.9F.D1.80.D0.B0.D0.BA.D1.82.D0.B8.D1.87.D0.B5.D1.81.D0.BA.D0.B8.D0.B5_.D0.B7.D0.B0.D0.B4.D0.B0.D0.BD.D0.B8.D1.8F

Примеры вопросов к экзамену:

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

1. Основная литература

1. James, Witten, Hastie, Tibshirani. An Introduction to Statistical Learning, 2013. (http://www-bcf.usc.edu/~gareth/ISL/ISLR Sixth Printing.pdf)
2. Bishop C.M. Pattern Recognition and Machine Learning, 2006. (https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf)

2. Дополнительная литература

1. Boyd, Vandenberghe. Convex Optimization (http://stanford.edu/~boyd/cvxbook/bv_cvxbook.pdf)

3. Программное обеспечение

№ п/п	Наименование	Условия доступа
1.	Anaconda	*Свободно распространяемое ПО*

4. Профессиональные базы данных, информационные справочные системы, интернет-ресурсы (электронные образовательные ресурсы)

№ п/п	Наименование	Условия доступа
	Интернет-ресурсы (электронные образовательные ресурсы)	
1.	Открытое образование	URL: https://openedu.ru/
2.	Coursera	URL: https://www.coursera.org

5. 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 (операционная система, офисные программы, антивирусные программы);

– мультимедийный проектор с дистанционным управлением.

Учебные аудитории для лабораторных и самостоятельных занятий по дисциплине оснащены ПЭВМ (операционная система, офисные программы), с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.