Программа учебной дисциплины «Интеллектуальный анализ данных»

Утверждена
Академическим советом ООП
Протокол № 8.1.2.1-14/01 от «28» июня 2018 г.

Автор	Адаптировано из программы НИУ ВШЭ “Введение в анализ данных” Разработчики программы: Игнатов Д.И., к.т.н., доцент, dignatov@hse.ru Соколов Е.А., старший преподаватель, esokolov@hse.ru Авторы адаптации: Петровичева А. Л., старший преподаватель, anna@xperience.ai Серебряков Г. А., grigory@xperience.ai Ханова Т. А., tatiana@xperience.ai
Число кредитов	5
Контактная работа (час.)	56
Самостоятельная работа (час.)	134
Курс	2 курс магистратуры
Формат изучения дисциплины	Без использования онлайн-курса.

I. ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ

Целями освоения дисциплины «Интеллектуальный анализ данных» являются овладение студентами моделями и методами интеллектуального анализа данных и машинного обучения в задачах поиска информации, обработки и анализа данных, а также приобретение навыков исследователя данных (data scientist) и разработчика математических моделей, методов и алгоритмов анализа данных.

В результате освоения дисциплины студент должен:

- Знать основные модели и методы машинного обучения и разработки данных;
- Уметь адекватно применять указанные модели и методы, а также программные средства, в которых они реализованы;
- Иметь навыки (приобрести опыт) анализа реальных данных с помощью изученных методов.

Изучение данной дисциплины базируется на следующих дисциплинах:

- математический анализ, линейная алгебра, дискретная математика, теория вероятностей и математическая статистика, исследование операций;
- технологии программирования.

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

- КР, ВКР.

II. СОДЕРЖАНИЕ УЧЕБНОЙ ДИСЦИПЛИНЫ

Раздел 1. Введение, основные понятия анализа данных

Введение в машинное обучение и анализ данных. Анализ данных в различных прикладных областях. Основные определения. Этапы анализа данных. Постановки задач машинного обучения. Примеры прикладных задач и их типы: классификация, регрессия, ранжирование, кластеризация, поиск структуры в данных.

Раздел 2. Математические объекты и методы в анализе данных

Линейная алгебра и анализ данных. Линейные пространства, их примеры из машинного обучения (признаки в кредитом скоринге, векторные представления текстов). Коллинеарность и линейная независимость. Скалярное произведение, косинус угла, примеры их применения. Векторы и матрицы, операции над ними. Матричное умножение. Системы линейных уравнений. Обратная матрица.

Математический анализ и анализ данных (на примере парной линейной регрессии и МНК). Производная и градиент, их свойства и интерпретации. Типы функций: непрерывные, разрывные, гладкие. Градиентный спуск. Выпуклые функции и их особое место в оптимизации.

Теория вероятностей и анализ данных. Случайные величины. Дискретные и непрерывные распределения, их свойства. Примеры распределений и их важность в анализе данных: биномиальное, пуассоновское, нормальное, экспоненциальное. Характеристики распределений: среднее, медиана, дисперсия, квантили. Пример их использования при генерации признаков. Центральная предельная теорема.

Математическая статистика и анализ данных. Оценивание параметров распределений. Метод максимального правдоподобия. Пример использования: анализ текстов и наивный байесовский классификатор. Доверительные интервалы и бутстрэппинг.

Раздел 3. Линейная регрессия и классификация

Линейная регрессия. Квадратичная функция потерь и предположение о нормальном распределении шума. Метод наименьших квадратов: аналитическое решение и оптимизационный подход. Стохастический градиентный спуск. Тонкости градиентного спуска: размер шага, начальное приближение, нормировка признаков. Проблема переобучения. Регуляризация.

Линейная классификация. Аппроксимация дискретной функции потерь. Отступ. Примеры аппроксимаций, их особенности. Градиентный спуск, регуляризация. Классификация и оценки принадлежности классам. Кредитный скоринг. Логистическая регрессия: откуда берется такая функция потерь и почему она позволяет предсказывать вероятности. Максимизация зазора как пример регуляризации и устранения неоднозначности решения.

Раздел 4. Оценивание качества алгоритмов

Регрессия: квадратичные и абсолютные потери, абсолютные логарифмические отклонения. Примеры использования.

Классификация: доля верных ответов, ее недостатки. Точность и полнота, их объединение: арифметическое среднее, минимум, гармоническое среднее (F-мера).

Оценки принадлежности классам: площади под кривыми. AUC-ROC, AUC-PRC, их свойства.

Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.

Практические особенности кросс-валидации. Стратификация. Потенциальные проблемы с разбиением зависимой или динамической выборки.

Раздел 5. Логические методы

Логические методы и их интерпретируемость. Простейший пример: список решений. Пример решающего списка для задачи фильтрации нежелательных сообщений. Деревья решений. Проблема построения оптимального дерева решений. Жадный алгоритм, основные его параметры.

Построение деревьев решений. Критерий ветвления. Выбор оптимального разбиения в задачах регрессии. Сложности выбора разбиения в задаче классификации. Примеры критериев: энтропийный (прирост информации), Джини и их модификации. Критерии завершения построения. Регуляризация и стрижка деревьев.

Раздел 6. Композиции алгоритмов

Простейший пример: уменьшение дисперсии при усреднении алгоритмов методом бутстреп. Блендинг алгоритмов.Понятие смещения и разброса (иллюстрация на примере линейных методов и решающих деревьев). Уменьшение разброса с помощью усреднения. Случайный лес. Оценка out-of-bag.

Раздел 7. Особенности реальных данных

Неполнота и противоречивость. Шумы и выбросы в данных. Методы поиска выбросов. Пропуски в данных, методы их восстановления. Несбалансированные выборки: проблемы и методы борьбы. Задача отбора признаков, примеры подходов.

Раздел 8. Анализ частых множеств признаков и ассоциативных правил

Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритм Априори. Меры “интересности правил”.

Раздел 9. Кластеризация данных

Простые эвристические подходы. Алгоритм K-Means. Проблема устойчивости результатов и важность грамотной инициализации, алгоритм K-Means++. Выбор числа кластеров. Оценка качества кластеризации.

Раздел 10. Нейронные сети

Типичные задачи. Алгоритм обратного распространения ошибки. Блоки нейронной сети. Архитектуры современных нейронных сетей. Типы нейронных сетей для различных видов данных. Нейронные сети для анализа изображений и видео.

III. ОЦЕНИВАНИЕ

Преподаватель оценивает работу студентов на практических занятиях: учитывается активность студентов и правильность решения задач. Оценки за работу на практических занятиях преподаватель выставляет в рабочую ведомость. Накопленная оценка по 10-ти балльной шкале за работу на практических занятиях определяется перед промежуточным или итоговым контролем – $O_{аудиторная}$.

Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

$$O_{накопленная} = 0{,}8 \cdot O_{текущий} + 0{,}2 \cdot O_{ауд}$$

где $O_{текущий}$ рассчитывается как взвешенная сумма всех форм текущего контроля, предусмотренных в РУП. Если в группе не оценивалась аудиторная активность, то $O_{накопленная} = O_{текущий}$.

$$O_{текущая\ 1\ модуля} = 0{,}2 \cdot O_{проект} + 0{,}8 \cdot O_{дз};$$

$$O_{текущая\ 2\ модуля} = 0{,}6 \cdot O_{дз} + 0{,}4 \cdot O_{проект}$$

Способ округления накопленной оценки текущего контроля: арифметический.

Результирующая оценка за дисциплину рассчитывается следующим образом:

$$O_{результ} = 0{,}7 \cdot O_{накопл\ Итоговая} + 0{,}3 \cdot O_{экз},$$

$$O_{накопленная\ Итоговая} = (O_{промежуточная\ 1} + O_{накопленная\ 2})/2,$$

$$O_{промежуточная\ 1} = 0{,}5 \cdot O_{накопленная\ 1\ модуля} + 0{,}5 \cdot O_{коллоквиум},$$ где

$O_{промежуточная\ 1}$ – промежуточная оценка модуля 3, а $O_{накопленная\ 2}$ – накопленная оценка 4 модуля перед итоговым экзаменом

Способ округления накопленной оценки промежуточного (итогового) контроля в форме экзамена: арифметический.

В случае особых обстоятельств студент может получить возможность пересдать низкие результаты за текущий контроль или работу на занятиях, а также самостоятельную работу.

На пересдаче студенту не предоставляется возможность получить дополнительный балл для компенсации оценки за текущий контроль.

На экзамене студент может получить дополнительный вопрос (дополнительную практическую задачу, решить к пересдаче домашнее задание), ответ на который оценивается в 1 балл.

В диплом выставляет результирующая оценка по учебной дисциплине.

IV. ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примерные темы домашних заданий:

Домашнее задание 1. Пакеты Numpy, Scipy, математические операции в них.

Домашнее задание 2. Пакет Pandas, работа с данными в нем.

Домашнее задание 3. Линейные методы классификации и регрессии.

Домашнее задание 4. Метрики качества алгоритмов машинного обучения, кросс-валидация.

Домашнее задание 5. Деревья решений, их построение.

Домашнее задание 6. Композиции алгоритмов. Случайные леса.

Домашнее задание 7. Работа с реальными данными. Предобработка признаков.

Домашнее задание 8. Кластеризация реальных данных.

Домашнее задание 9. Поиск частых множеств и ассоциативных правил.

Примерный перечень вопросов к экзамену:

1. Основные понятия машинного обучения. Основные постановки задач. Примеры прикладных задач.
2. Линейные пространства. Векторы и матрицы. Линейная независимость. Обратная матрица.
3. Производная и градиент функции. Градиентный спуск. Выпуклые функции.
4. Случайные величины. Дискретные и непрерывные распределения. Примеры.
5. Оценивание параметров распределений, метод максимального правдоподобия. Бутстрэппинг.
6. Линейные методы классификации и регрессии: функционалы качества, методы настройки, особенности применения.
7. Метрики качества алгоритм регрессии и классификации.
8. Оценивание качества алгоритмов. Отложенная выборка, ее недостатки. Оценка полного скользящего контроля. Кросс-валидация. Leave-one-out.
9. Деревья решений. Методы построения деревьев. Их регуляризация.
10. Композиции алгоритмов. Разложение ошибки на смещение и разброс.
11. Случайный лес, его особенности.
12. Методы поиска выбросов в данных. Методы восстановления пропусков в данных. Работа с несбалансированными выборками.
13. Задача анализа потребительской корзины. Поддержка и достоверность. Частые, замкнутые и максимальные частые множества. Алгоритм Априори.
14. Задача кластеризации. Алгоритм K-Means. Оценки качества кластеризации.
15. Нейронные сети для анализа изображений.

5. РЕСУРСЫ

5.1 Основная литература

1 Ben-Tal, A. Lectures on modern convex optimization: analysis, algorithms, and engineering applications [Электронный ресурс] /A.Ben-Tal, A.Nemirovski; DB Books24x7. – Philadelphia: Society for Industrial and Applied Mathematics (SIAM), 2001. – 504 p. - (MPS-SIAM series on optimization). - Режим доступа: https://library.books24x7.com/toc.aspx?bookid=9410. – Загл. с экрана.

2. Mirkin, B. Core Concepts in Data Analysis: Summarization, Correlation, Visualization [Электронный ресурс] / Boris Mirkin; DB Springer Books. – London: Springer-Verlag Limited, 2011. - Режим доступа: https://link.springer.com/book/10.1007/978-0-85729-287-2. - Загл. с экрана.

5.2 Дополнительная литература

1. Кацко, И.А. Практикум по анализу данных на компьютере: учебно-практическое пособие / И.А.Кацко, Н.Б.Паклин; под ред. проф. Г.В.Гореловой. - М.: КолосС, 2009. - 278 с.

2. A Modern Introduction to Probability and Statistics: Understanding Why and How [Электронный ресурс] / F.M.Dekking, C.Kraaikamp, H.P.Lopuhaä, L.E.Meester; DB Springer

Books. – London: Springer-Verlag Limited, 2005. – Режим доступа: https://link.springer.com/book/10.1007/1-84628-168-7. - Загл. с экрана.
3. Nisbet, R. Handbook of statistical analysis and data mining applications [Электронный ресурс] / R.Nisbet, J.Elder, G.Miner; DB ebrary. – Amsterdam: Elsevier, 2009. – 824 с. + CD-ROM. – Режим доступа: https://ebookcentral.proquest.com/lib/hselibrary-ebooks/reader.action?docID=452830&query=Nisbet%2C+Robert. – Загл. с экрана.

5.3. Программное обеспечение

№ п/п	Наименование	Условия доступа/скачивания
1	Язык программирования Python	*Свободный бесплатный доступ в сети Интернет.*
2	Библиотеки NumPy, SciPy, Pandas, Scikit-Learn	*Свободный бесплатный доступ в сети Интернет.*

5.4. 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 проектором (для лекций или семинаров), с возможностью подключения к сети Интернет и доступом к электронной информационно-образовательной среде НИУ ВШЭ.