Федеральное государственное автономное образовательное учреждение высшего образования
"Национальный исследовательский университет "Высшая школа экономики"

Факультет компьютерных наук

Департамент анализа данных и искусственного интеллекта

_____

Майнор Интеллектуальный анализ данных
Рабочая программа дисциплины Введение в анализ данных

для уровня подготовки - бакалавриат

Разработчики программы
Игнатов Д.И., к.т.н., доцент, dignatov@hse.ru
Соколов Е.А., старший преподаватель, esokolov@hse.ru

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

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

Академический руководитель образовательной программы **предлагающей(его) майнор**)
А.С. Конушин _____

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 минимальные требования к знаниям и умениям студента и определяет содержание и виды учебных занятий и отчетности.

Программа предназначена для преподавателей, ведущих данную дисциплину, учебных ассистентов и студентов, выбравших майнор «Интеллектуальный анализ данных».

Программа разработана в соответствии с образовательным стандартом Федерального государственного автономного образовательного учреждения высшего профессионального образования «Национальный исследовательский университет «Высшая школа экономики» по направлению подготовки 01.03.02 Прикладная математика и информатика для квалификации «бакалавр».

2 Цели освоения дисциплины

Целями освоения дисциплины «**Введение в 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.

3 Компетенции обучающегося, формируемые в результате освоения дисциплины

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

В результате освоения дисциплины студент осваивает следующие компетенции:

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
Способен рефлексировать (оценивать и перерабатывать) освоенные научные методы и способы деятельности.	СК-М1	Понимание места и ценности методов машинного обучения и разработки данных в современной науке и практической деятельности.	Лекции и практикумы.
Способен использовать в профессиональной деятельности знания в области естественных наук, математики и информатики,	ИК-М7.1пми	Корректно применяет полученные ранее знания в таких дисциплинах как математический анализ, дифференциальные уравнение, дискретная математика и	Практикумы. Работа над проектом по анализу реальных данных.

Компетенция	Код по ФГОС/ НИУ	Дескрипторы – основные признаки освоения (показатели достижения результата)	Формы и методы обучения, способствующие формированию и развитию компетенции
понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой.		стохастическое моделирование при изучении материалов курса.	
Способен строить и решать математические модели в соответствии с направлением подготовки и специализацией.	ИК-М7.2пми	Адекватно определяет тип задачи, строит модель и подбирает параметры методов.	Практикумы. Работа над проектом по анализу реальных данных.
Способен применять в исследовательской и прикладной деятельности современные языки программирования и языки манипулирования данными, операционные системы, электронные библиотеки и пакеты программ, сетевые технологии и т.п.	ИК-М7.5пми	Способен адекватно разрабатывать программы на одном из доступных языков программирования, использовать программные средства (в том числе самостоятельно разработанные) при сборе, предобработке и анализе данных.	Практикумы.

4 Место дисциплины в структуре образовательной программы

Для специализаций подготовки бакалавров настоящая дисциплина является неотъемлемой дисциплиной в рамках майнора «Интеллектуальный анализ данных», который предоставляет дополнительную специализацию.

Изучение данной дисциплины базируется на следующих дисциплинах:

- Введение в программирование

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

- Необходимо владеть основами программирования на языке Python (изучается в рамках первой дисциплины курса «Введение в программирование») и знаниями математики в объеме программы средней школы.

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

- Введение в машинное обучение
- Интеллектуальные системы

5 Тематический план учебной дисциплины

№	Название раздела	Всего часов	Аудиторные часы			Самостоя-тельная работа
			Лекции	Семинары	Практические занятия	
1	Введение, основные понятия анализа данных		4	2		
2	Математические объекты и методы в анализе данных		8	8		
3	Метрические методы		2	2		
4	Линейная регрессия и классификация		4	4		
5	Оценивание качества алгоритмов		4	4		
6	Логические методы		4	4		
7	Композиции алгоритмов		2	2		
8	Особенности реальных данных		2	2		
9	Кластеризация данных		2	2		

6 Формы контроля знаний студентов

Тип контроля	Форма контроля	1 год		Параметры **
		3	4	
Текущий (неделя)	Домашнее задание	*		срок выполнения и требования указываются в тексте задания
	Домашнее задание	*		срок выполнения и требования указываются в тексте задания
	Домашнее задание		*	срок выполнения и требования указываются в тексте задания
	Домашнее задание		*	срок выполнения и требования указываются в тексте задания
Промежу-точный	Коллоквиум			В письменной форме
Итоговый	Экзамен			В устной форме

6.1 Критерии оценки знаний, навыков

Оценки по всем формам текущего контроля выставляются по 10-ти балльной шкале.

Выполнение домашнего задания оценивается в баллах по 10-ти балльной шкале. В тексте задания могут быть указаны баллы для каждого из подзаданий. Основные критерии: корректность и полнота представленного решения.

7 Содержание 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. Метрические методы

Гипотеза компактности. Функция расстояния между объектами. Метрические алгоритмы классификации, их модификация с весами. Метрические алгоритмы регрессии.

Раздел 4. 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 5. 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 6. 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 7. 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 8. 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 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.

8 Образовательные технологии

Необходимое для выполнения работ программное обеспечение, как правило, находится в свободном доступе и его можно загрузить в сети Интернет или скопировать из репозитория, предоставляемого курсу.

В число программных пакетов входят инструменты PyData (http://pydata.org/downloads/):

- Пакеты scipy и numpy.
- Сборка Anaconda
- Pandas
- Scikit-learn и др.

Дополнительно к каждой из тем доступны слайды лекций, изложение которых адаптировано с учетом используемых программных пакетов.
В рамках курса используется также проектная форма работы.

8.1 Методические рекомендации преподавателю

Преподавателю рекомендуется использовать демонстрацию работы изучаемых методов анализа данных с помощью предустановленных программных продуктов (PyData) во время лекционных занятий.

В рамках семинаров возможно решение задач, а домашние задания предлагается составлять практико-ориентированными.

Для проектов по анализу данных рекомендуется предлагать доступные источники данных, например, UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/).

9 Оценочные средства для текущего контроля и аттестации студента

9.1 Тематика заданий текущего контроля

Примерные темы домашних заданий:

Домашнее задание 1. Пакеты Numpy, Scipy, математические операции в них. Пакет Pandas, работа с данными в нем.

Домашнее задание 2. Линейные методы классификации и регрессии.

Домашнее задание 3. Деревья решений, их построение.

Домашнее задание 4. Композиции алгоритмов. Случайные леса.

9.2 Вопросы для оценки качества освоения дисциплины

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.

10 Порядок формирования оценок по дисциплине

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 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.

11 Учебно-методическое и информационное обеспечение дисциплины

11.1 Базовый учебник

Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)

11.2 Основная литература

1. Mohammed J. Zaki, Wagner Meira Jr. Data Mining and Analysis. Fundamental Concepts and Algorithms. Cambridge University Press, 2014 (http://www.dataminingbook.info/pmwiki.php/Main/BookDownload)
2. Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010 (http://www.hse.ru/data/2010/10/14/1223126254/Mirkin_All.pdf)

11.3 Дополнительная литература

1. Boyd, Vandenberghe. Convex Optimization (http://stanford.edu/~boyd/cvxbook/)
2. Dekking, F.M., Kraaikamp, C., Lopuhaä, H.P., Meester, L.E., A Modern Introduction to Probability and Statistics (http://www.ewi.tudelft.nl/index.php?id=50508 и http://www.springer.com/gp/book/9781852338961)

11.4 Справочники, словари, энциклопедии

Портал http://www.machinelearning.ru

11.5 Программные средства

Для успешного освоения дисциплины, студент использует следующие программные средства: язык программирования Python, его библиотеки NumPy, SciPy, Pandas, Scikit-Learn

11.6 Дистанционная поддержка дисциплины

Дистанционная поддержка может осуществляться с помощью LMS или хранилища слайдов и данных в DropBox папке автора курса (возможно использование иных облачных сервисов).

12 Материально-техническое обеспечение дисциплины

Используется проектор (для лекций или семинаров), слайды мультимедийных презентаций и компьютеры с предустановленным программным обеспечением и доступ в Интернет.