**Федеральное государственное автономное образовательное учреждение высшего профессионального 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 Академическим советом ОП бакалаврской программы «Прикладная математика и информатика»

 Протокол № \_\_\_\_\_\_\_\_\_\_\_\_\_\_ «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2015 г.

Утверждена «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_ 2015 г.

Академический руководитель образовательной программы/руководитель подразделения (название ОП/подразделения, **предлагающей(его) майнор**)

Конушин Антон Сергеевич \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ [подпись]

Москва, 2015

*Настоящая программа не может быть использована другими подразделениями университета и другими вузами без разрешения подразделения-разработчика программы.*

# 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».

# 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.

# 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 анализировать, верифицировать, оценивать полноту информации в ходе профессиональной деятельности, при необходимости восполнять и синтезировать недостающую информацию. | СК-М6 | Умеет самостоятельно находить научно-техническую литеру по тематике и критически подходить к оценке моделей и методов машинного обучения и разработки в ней рассмотренных. | Групповая работа над анализом реальных данных. |
| Способен задавать, транслировать правовые и этические нормы в профессиональной и социальной деятельности. | СЛК-М2 | Относится с уважением к международным этическим стандартам сохранения анонимности персональных данных (Data Privacy Preserving). | Работа над проектом по анализу реальных данных. |
| Способен публично представлять результаты профессиональной деятельности (в том числе с использованием информационных технологий). | ИК-М2.5 | Умеет представлять результаты проведенного исследования в виде отчета или устного рассказа с показом мультимедийной презентации. | Подготовка и представление результатов отчета по проекту. |
| Способен использовать в профессиональной деятельности знания в области естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой. | ИК-М7.1пми | Корректно применяет полученные ранее знания в таких дисциплинах как математический анализ, дифференциальные уравнение, дискретная математика и стохастическое моделирование при изучении материалов курса. | Практикумы. Работа над проектом по анализу реальных данных. |
| 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 понимать и применять в исследовательской и прикладной деятельности современный математический аппарат. | ИК-М7.3пми | Корректно применяет математические модели и методы прикладной математики в анализе данных. | 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. Работа над проектом по анализу реальных данных. |

# 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](http://electives.hse.ru/minor_intel/machine/)
* [Интеллектуальные системы](http://electives.hse.ru/minor_intel/systems/)

# Тематический план учебной дисциплины

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

# Формы контроля знаний студентов

|  |  |  |  |
| --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | Параметры \*\* |
| 3 | 4 |  |
| Текущий(неделя) | Домашнее задание | 3 |  | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание | 5 |  | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание | 7 |  | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание | 9 |  | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание |  | 2 | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание |  | 5 | срок выполнения и требования указываются в тексте задания |
|  | Домашнее задание |  | 7 | срок выполнения и требования указываются в тексте задания |
|  | Проект по анализу реальных данных |  | 1 | Предполагается промежуточный контроль в конце первого модуля и итоговая защита в рамках экзамена во втором модуле. |
| Промежу­точный | Коллоквиум |  |  | В устной форме |
| Итоговый | Экзамен  |  |  | В устной форме, а также собеседование по результатам домашнего задания и проекта |

## 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.

Проект по анализу реальных данных подразумевает коллективное или индивидуальное выполнение задания, связанного с применением программных средств и методов разработки данных и машинного обучения на реальных данных. Защита проекта происходит в два этапа, на первом модуле происходит выбор данных, обоснование методов для решения задачи и первичное исследование данных. К экзамену студент предоставляет (высылает в электронной форме) отчет по проекту, а рамках устного собеседования на экзамене происходит его защита.

Осуществляется следующая дистанционная поддержка при проведении контроля: задания и материалы лекций, рассылаются по электронной почте, учебная литература и материалы лекций доступны в DropBox (или LMS, GitHub, SageMath и прочие облачные сервисы).

# 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. 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.

# 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.

## 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/>).

# Оценочные средства для текущего контроля и аттестации студента

## Тематика заданий текущего контроля

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.

## 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.

# Порядок формирования оценок по дисциплине

Преподаватель оценивает работу студентов на практических занятиях: . 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 – *Оаудиторная*.

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

*Онакопленная*= 0,8∙ *Отекущий* + 0,2∙ Оауд

где *Отекущий* рассчитывается как взвешенная сумма всех форм текущего контроля, предусмотренных в РУП. Если в группе не оценивалась аудиторная активность, то *Онакопленная*= *Отекущий*.

*Отекущая 1 модуля* = *0,2 ·Опроект +0,8 ·Одз* ;

*Отекущая 2 модуля* = *0,6 ·Одз +0,4 ·Опроект*

Способ округления накопленной оценки текущего контроля: арифметический.

Результирующая оценка за дисциплину рассчитывается следующим образом:

*Орезульт = 0,6∙ Онакопл Итоговая + 0,4 ∙Оэкз,*

О*накопленная Итоговая=* (О*промежуточная 1+* О*накопленная 2)/2,*

*Опромежуточная 1* = *0,5·Онакопленная 1 модуля +0,5·Околлоквиум,* где

О*промежуточная 1 –* промежуточная оценка модуля 3, а О*накопленная 2 –* накопленная оценка 4 модуля перед итоговым экзаменом

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

В случае особых обстоятельств .

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.

# Учебно-методическое и информационное обеспечение дисциплины

## 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>)

## 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>)

## 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>)

## Справочники, словари, энциклопедии

Портал [http://www.machinelearning.ru](http://www.machinelearning.ru/)

## Программные средства

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

## 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).

# Материально-техническое обеспечение дисциплины

м и доступ в Интернет.