**Программа учебной дисциплины \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Утверждена

Академическим советом ООП

Протокол № от «\_\_»\_\_\_\_\_20\_\_ г.

|  |  |
| --- | --- |
| Автор | Соколов Евгений Андреевич |
| Число кредитов | 5 |
| Контактная работа (час.) | 72 |
| Самостоятельная работа (час.) | 180 |
| Курс | Введение в анализ данных |
| Формат изучения дисциплины | Без использования онлайн курса |

1. **ЦЕЛЬ, РЕЗУЛЬТАТЫ ОСВОЕНИЯ ДИСЦИПЛИНЫ И ПРЕРЕКВИЗИТЫ**

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.

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

Для специализаций подготовки бакалавров настоящая дисциплина является неотъемлемой дисциплиной в рамках майнора «Интеллектуальный анализ данных».

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:

* [Современные методы машинного обучения](http://electives.hse.ru/minor_intel/machine/)
* [Прикладные задачи анализа данных](http://electives.hse.ru/minor_intel/systems/)

# Содержание УЧЕБНОЙ ДИСЦИПЛИНЫ

Тема 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.

Тема 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.

Тема 3. Метрические методы

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

Тема 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.

Тема 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.

Тема 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.

Тема 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.

Тема 8. Рекомендательные системы

Постановка задачи. Коллаборативная фильтрация. Методы на основе матрицы оценок: item-based, user-based. Матричные разложения.

Тема 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.

# ОЦЕНИВАНИЕ

В рамках курса предусмотрены самостоятельные работы на занятиях, 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, экзаменационная и итоговая оценки округляются арифметически.

Накопленная оценка рассчитывается по формуле

O\_накопл=0.2 O\_самост + 0.6 O\_дз + 0.2 O\_контрольная

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.

# ПРИМЕРЫ ОЦЕНОЧНЫХ СРЕДСТВ

Примеры практических заданий можно найти по ссылке <http://wiki.cs.hse.ru/Майнор_Интеллектуальный_анализ_данных/Введение_в_анализ_данных/2017-2018#.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 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 кластеризации. Алгоритм K-Means. Оценки качества кластеризации.
14. **РЕСУРСЫ**
    1. **Основная литература**
15. 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>)
16. Boris Mirkin. Core Concepts in Data Analysis: Summarization, Correlation, Visualization. 2010 (<http://www.hse.ru/data/2010/10/14/1223126254/Mirkin_All.pdf>)
    1. **Дополнительная литература**
17. Boyd, Vandenberghe. Convex Optimization (<http://stanford.edu/~boyd/cvxbook/>)
18. 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>)
    1. **Программное обеспечение**

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

* 1. **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 |

* 1. **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.