**Правительство Российской Федерации**

**Федеральное государственное автономное образовательное учреждение высшего профессионального образования  
«Национальный исследовательский университет   
«Высшая школа экономики»**

Факультет Компьютерных наук

Департамент больших данных и информационного поиска

Базовая кафедра Яндекс

УТВЕРЖДАЮ

Академический руководитель

образовательной программы

«Программная инженерия»

по направлению 09.03.04

«Программная инженерия»

К.Ю. Дегтярев

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014 г.

**Программа дисциплины** «Анализ данных»

для направления 09.03.04 "Программная инженерия" подготовки бакалавра

для программы бакалавриата "Программная инженерия"

**Автор программы:**

Бабенко А.В. (arbabenko@yandex-team.ru)

Одобрена на заседании базовой кафедры Яндекс

«\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Заведующий кафедрой \_\_\_\_\_\_\_\_\_\_\_\_\_\_ М.А. Бабенко

Рекомендована Академическим советом образовательной программы

«Науки о данных» «\_\_\_»\_\_\_\_\_\_\_\_\_\_\_\_\_ 2014  г.

Менеджер базовой кафедры Яндекс

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ Е.Ф. Баулин

Москва, 2014

*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 направления подготовки **231000.62 «Программная инженерия»**, обучающихся по программе бакалавриата.

Программа разработана в соответствии с:

* Образовательным стандартом государственного образовательного бюджетного учреждения высшего профессионального образования «Государственный университет – Высшая школа экономики», в отношении которого установлена категория «Национальный исследовательский университет»;
* Рабочим учебным планом университета подготовки бакалавра по направлению 231000.62 «Программная инженерия», утвержденным в 2014 г.

# Цели освоения дисциплины

Целью освоения дисциплины «Анализ данных» является обучить студентов применять основные модели и методы анализа данных и машинного обучения для обработки реальных данных из прикладных задач.

# Компетенции обучающегося, формируемые в результате освоения дисциплины

В результате освоения учебной дисциплины «Анализ данных» обучающийся должен

**знать:**

**-** основные алгоритмы машинного обучения;

- преимущества и недостатки методов машинного обучения;

- типы задач, решаемые методами машинного обучения;

- способы оценки качества методов машинного обучения;

- существующие программные инструменты машинного обучения

**уметь:**

– формулировать прикладные задачи на языке машинного обучения;

- выбирать метод, подходящий конкретной прикладной задаче;

- пользоваться готовыми инструментами при решении конкретной задачи;

- правильно ставить эксперименты при решении исследовательской задачи;

- проводить сравнительный анализ различных методов машинного обучения;

**владеть:**

- основными определениями, методами и алгоритмами анализа данных;

- стандартными инструментариями обработки реальных данных.

В результате освоения дисциплины студент осваивает следующие компетенции:

|  |  |  |  |
| --- | --- | --- | --- |
| Общенаучная | ОНК-4 | Готовность использовать основные законы естественнонаучных дисциплин в профессиональной деятельности, применять методы математического анализа и моделирования, теоретического и экспериментального исследования при работе в какой-либо предметной области | Стандартные (лекционно-семинарские) |
| Общенаучная | ОНК-5 | Готовность выявить естественнонаучную сущность проблем, возникающих в ходе профессиональной деятельности, привлечь их для решения соответствующий аппарат дисциплины | Стандартные (лекционно-семинарские) |
| Общенаучная | ОНК-6 | Способность приобретать новые знания с использованием научной методологии и современных образовательных и информационных технологий | Стандартные (лекционно-семинарские) |
| Общенаучная | ОНК-7 | Способность порождать новые идеи (креативность) | Стандартные (лекционно-семинарские) |
| Инструментальные | ИК-2 | Умение работать на компьютере, навыки использования основных классов прикладного программного обеспечения, работы в компьютерных сетях, составления баз данных | Стандартные (лекционно-семинарские) |
| Профессиональные | ПК-1 | Способность демонстрации общенаучных базовых знаний естественных наук, математики и информатики, понимание основных фактов, концепций, принципов теорий, связанных с прикладной математикой и информатикой | Стандартные (лекционно-семинарские) |
| Профессиональные | ПК-2 | Способность понимать и применять в исследовательской и прикладной деятельности современный математический аппарат | Стандартные (лекционно-семинарские) |
| Профессиональные | ПК-4 | способность критически оценивать собственную квалификацию и её востребованность, переосмысливать накопленный практический опыт, изменять при необходимости вид и характер своей профессиональной деятельности | Стандартные (лекционно-семинарские) |
| Профессиональные | ПК-8 | Способность решать задачи производственной и технологической деятельности на профессиональном уровне, включая разработку математических моделей, алгоритмических и программных решений | Стандартные (лекционно-семинарские) |

# Место дисциплины в структуре образовательной программы

Программа учебной дисциплины «Анализ данных» предназначена для подготовки студентов, обучающихся по направлению 231000.62 «Программная инженерия». Программа составлена в соответствии с Образовательными стандартами и учебными планами государственного образовательного бюджетного учреждения высшего профессионального образования [Национального исследовательского университета «Высшая школа экономики»](http://www.hse.ru/).

«Анализ данных» является самостоятельной учебной дисциплиной, относится к математическому и естественнонаучному циклу дисциплин.

Для специализаций 231000.62 «Программная инженерия» настоящая дисциплина является базовой.

Изучение данной дисциплины базируется на следующих дисциплинах:

* Математический анализ;
* Геометрия и алгебра;
* Теория вероятностей и математическая статистика.

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

* Знаниями основных определений и теорем перечисленных выше дисциплин;
* Навыками решения типовых задач этих дисциплин.

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

* Эконометрика;
* Компьютерное моделирование;
* При подготовке ВКР.

# Тематический план учебной дисциплины

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| № | Название раздела | Всего часов | Аудиторные часы | | | Самостоятельная работа |
| ***Лекции*** | ***Семинары*** | ***Практические занятия*** |
|  | Основные понятия машинного обучения и примеры прикладных задач. | 10 | 4 |  | 4 | 10 |
|  | Метрические методы классификации | 20 | 4 |  | 4 | 10 |
|  | Логические методы классификации | 20 | 4 |  | 4 | 10 |
|  | Линейные методы классификации | 30 | 6 |  | 6 | 16 |
|  | Регрессионный анализ | 20 | 4 |  | 4 | 10 |
|  | Байесовская классификация | 40 | 8 |  | 8 | 22 |
| 7. | Нейронные сети | 20 | 4 |  | 4 | 10 |
| 8. | Кластерный анализ | 20 | 6 |  | 6 | 12 |
|  | Итого | 180 | 40 |  | 40 | 100 |

# Формы контроля знаний студентов

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Тип контроля | Форма контроля | 1 год | | | | Параметры |
| 1 | 2 | 3 | 4 |
| Текущий  (неделя) | Лабораторная работа |  |  | 9 |  | письменная работа 80 минут |
|  |  |  |  |  |
| Домашнее задание |  |  |  | 9 |  |
| Итоговый | Экзамен |  |  |  | э | Письменная работа на 100 минут |

## 6.1 Критерии оценки знаний, навыков

Для прохождения контроля студент должен, как минимум, продемонстрировать знания основных определений и описание алгоритмов; умение решать типовые задачи на ЭВМ, разобранные на семинарских занятиях и умение выбирать оптимальный метод обработки данных конкретной прикладной задачи.

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

**6.2 Порядок формирования оценок по дисциплине**Накопленная оценка за текущий контроль учитывает результаты студента по текущему контролю следующим образом:

О*накопленная*= *Отекущий*

где*Отекущий* рассчитывается как взвешенная сумма всех форм текущего контроля, предусмотренных в РУП

*Отекущий* = *0.4\*Одз* + *0.6\*Окр*.

Результирующая оценка за дисциплину рассчитывается следующим образом:

*Орезульт = 0.8\* Онакопл + 0.2\*·Оэкз/зач*

Способ округления результирующей оценки: если дробная часть результирующей оценки составляет меньше 0.7, то результирующая оценка равна целой части полученного значения; если дробная часть результирующей оценки не менее 0.7, то результирующая оценка равна целой части полученного значения плюс 1.

Оценка за итоговый контроль – не блокирующая.

# Содержание дисциплины

***Раздел 1.***

**Основные понятия машинного обучения и примеры прикладных задач.(Л.-6ч.,ПЗ.-4ч.,СРС-16ч.: проработка лекций-8ч., подготовка к практическим занятиям -4 ч., выполнение домашнего задания-4ч.)**

Основные понятия и примеры прикладных задач. Постановка задач обучения по прецедентам. Объекты и признаки. Типы шкал: бинарные, номинальные, порядковые, количественные. Типы задач: классификация, регрессия, кластеризация. Основные понятия: модель алгоритмов, метод обучения, функция потерь и функционал качества, принцип минимизации эмпирического риска, обобщающая способность, скользящий контроль, переобучение. Примеры прикладных задач.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел 2.***

**Метрические методы классификации** **(Л.-6ч., С.-8ч.,СРС -18ч. .: проработка лекций-6ч., подготовка к практическим занятиям -8 ч., выполнение домашнего задания-4ч.)**

Метрические алгоритмы классификации. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Обобщенный метрический классификатор, понятие отступа.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел 3*.**

**Логические методы классификации (Л.-6 ч., ПЗ.-6ч., СРС -16ч. .: проработка лекций-6ч., подготовка к семинарским занятиям -6 ч., выполнение домашнего задания-4ч.)**

Логические закономерности и решающие деревья. Понятие логической закономерности. Определение информативности. Разновидности закономерностей: шары, гиперплоскости, гиперпараллелепипеды (конъюнкции). Бинаризация признаков, алгоритм выделения информативных зон. «Градиентный» алгоритм синтеза конъюнкций, частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция. Решающее дерево. Псевдокод: жадный алгоритм ID3. Недостатки алгоритма и способы их устранения. Проблема переобучения. Редукция решающих деревьев: предредукция и постредукция.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел 4*.**

**Линейные методы классификации** **(Л.-10ч., С.-12ч., СРС -26ч. .: проработка лекций-10ч., подготовка к практическим занятиям -10ч., выполнение домашнего задания-6ч.)**

Линейные алгоритмы классификации. Квадратичная функция потерь, метод наименьших квадратов. Метод стохастического градиента и частные случаи: перcептрон Розенблатта, правило Хэбба. Недостатки метода стохастического градиента и способы их устранения. Ускорение сходимости, «выбивание» из локальных минимумов. Проблема переобучения, редукция весов (weight decay). Логистическая регрессия. Метод опорных векторов (SVM). Оптимальная разделяющая гиперплоскость. Понятие зазора между классами (margin). Случаи линейной разделимости и отсутствия линейной разделимости.Связь с минимизацией регуляризованного эмпирического риска. Кусочно-линейная функция потерь. Задача квадратичного программирования и двойственная задача. Понятие опорных векторов. Рекомендации по выбору константы C. Функция ядра (kernel functions), спрямляющее пространство, теорема Мерсера. Способы конструктивного построения ядер. Примеры ядер. Шкалирование Платта для обучения вероятностного вывода в SVM.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел* 5.**

**Регрессионный анализ (Л.-4ч., ПЗ.-6ч., СРС-12ч. .: проработка лекций-4ч., подготовка к семинарским занятиям -6 ч., выполнение домашнего задания-2ч.)**

Методы восстановления регрессии. Задача восстановления регрессии, метод наименьших квадратов. Одномерная непараметрическая регрессия (сглаживание): оценка Надарая-Ватсона, выбор ядра и ширины окна сглаживания. Многомерная линейная регрессия. Сингулярное разложение. Регуляризация: гребневая регрессия и лассо Тибширани. Метод главных компонент и декоррелирующее преобразование.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел* 6.**

**Байесовская классификация** **(Л.-8 ч., С.-4ч., СРС -12ч.: проработка лекций-8ч., подготовка к семинарским занятиям -4 ч)**

Байесовские алгоритмы классификации, непараметрические методы. Вероятностная постановка задачи классификации. Основные понятия: априорная вероятность, апостериорная вероятность, функция правдоподобия класса. Функционал среднего риска. Ошибки I и II рода. Оптимальный байесовский классификатор. Оценивание плотности распределения: три основных подхода. Наивный байесовский классификатор. Непараметрическое оценивание плотности распределения по Парзену-Розенблатту. Квадратичный дискриминант. Вид разделяющей поверхности. Подстановочный алгоритм, его недостатки и способы их устранения. Линейный дискриминант Фишера.EM-алгоритм и его обобщения. Логистическая регрессия.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел* 7.**

**Нейронные** **сети (Л.-8 ч., С.-4ч., СРС -12ч.: проработка лекций-8ч., подготовка к семинарским занятиям -4 ч)**

Нейронные сети. Структура многослойной нейронной сети. Функции активации. Проблема полноты. Задача исключающего или. Полнота двухслойных сетей в пространстве булевских функций. Алгоритм обратного распространения ошибок. Формирование начального приближения. Проблема паралича сети. Методы оптимизации структуры сети. Выбор числа слоев и числа нейронов в скрытом слое. Постепенное усложнение сети. Оптимальное прореживание сети (optimal brain damage).

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

***Раздел* 8.**

**Кластерный анализ (Л.-8 ч., С.-4ч., СРС -12ч.: проработка лекций-8ч., подготовка к семинарским занятиям -4 ч)**

Постановка задачи кластеризации. Примеры прикладных задач. Типы кластерных структур. Графовые алгоритмы кластеризации. Выделение связных компонент. Кратчайший незамкнутый путь. Функционалы качества кластеризации. Статистические алгоритмы: EM-алгоритм и Алгоритм k средних (k-means). Эвристики для ускорения сходимости k-means.

**Литература по разделу.**

Базовая: [1]

Основная: [2],[3]

# Оценочные средства для текущего контроля и аттестации студента

## 8.1 Тематика заданий текущего контроля

1.Переобучение и способы его выявления

2.Выбор метрики в метрических классификаторах

3.Определение способа редукции решающих деревьев

4.Выбор регуляризатора для линейной классификации

5.Выбор структуры нейронной сети

***Вариант билета контрольной работы .***

1. Какова асимптотическая сложность этапов обучения и классификации для классификатора Parzen Window с фиксированной шириной окна, если обучающая выборка содержит n объектов, каждый из которых описан d признаками?

2. Какова асимптотическая сложность этапов обучения и классификации для QDA и LDA, если обучающая выборка содержит n объектов, каждый из которых описан d признаками?

3. Как можно применить метод ближайшего соседа к данным, у которых есть номинальные признаки?

4. Назовите основную причину переобучения решающего дерева (алгоритм ID3)

# Учебно-методическое и информационное обеспечение дисциплины

## 9.1 Базовый учебник

1. Воронцов К.В. Курс лекций по машинному обучению.

## 9.2 Основная литература

2. Christopher Bishop. Pattern Recognition and Machine Learning.

3. Trevor Hastie, Robert Tibshirani, Jerome Friedman. Elements of Statistical Learning.

# Материально-техническое обеспечение дисциплины

Для проведения семинарских занятий требуется компьютерный класс с установленным ППП Python с набором пакетов Anaconda.