1. **Паспорт Образовательной программы**

**«Python для анализа данных»**

|  |  |
| --- | --- |
| **Версия программы** | **1** |
| **Дата Версии** | 15**.**10.2020 |

1. **Сведения о Провайдере**

|  |  |  |
| --- | --- | --- |
| 1.1 | Провайдер | АНО ДПО «СофтЛайн Эдюкейшн» |
| 1.2 | Логотип образовательной организации |  |
| 1.3 | Провайдер ИНН | **7736228783** |
| 1.4 | Ответственный за программу ФИО | **Максименко Денис Владиславович** |
| 1.5 | Ответственный должность | Руководитель проектного офиса |
| 1.6 | Ответственный Телефон | +7-495-232-0023 доб. 0889 |
| 1.7 | Ответственный Е-mail | Edu2035@softline.com |

1. **Основные Данные**

|  |  |  |
| --- | --- | --- |
| **№** | **Название** | **Описание** |
| 2.1 | Название программы | Python для анализа данных |
| 2.2 | Ссылка на страницу программы | <https://www.basispythoncourse.ru/> - страница программы  <https://www.basispythoncourse.ru/#rec237919852> - входное тестирование |
| 2.3 | Формат обучения | Электронное обучение |
|  | Подтверждение от ОО наличия возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа | Подтверждаем наличие возможности реализации образовательной программы с применением электронного обучения и (или) дистанционных образовательных технологий с возможностью передачи данных в форме элементов цифрового следа |
| 2.4 | Уровень сложности | Базовый |
| 2.5 | Количество академических часов | 72 академических часа |
|  | Практикоориентированный характер образовательной программы: не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и (или) выполнению практических заданий в режиме самостоятельной работы (кол-во академических часов) | Не менее 50 % трудоёмкости учебной деятельности отведено практическим занятиям и выполнению практических заданий в режиме самостоятельной работы:  36 ак. часов практических занятий, в том числе 15 практических заданий на программирование (см. п. 7), что составляет не менее 50% трудоемкости учебной деятельности. |
| 2.6 | Стоимость обучения одного обучающегося по образовательной программе, а также предоставление ссылок на 3 (три) аналогичные образовательные программы иных организаций, осуществляющих обучение, для оценки объективности стоимости или обоснование уникальности представленной образовательной программы в случае отсутствия аналогичных образовательных программ на рынке образовательных услуг | 24 900 рублей  [Программирование на Python ВШЭ 40 ак. часов 20 000 рублей](https://busedu.hse.ru/catalog/338121148.html)  [Первичный анализ данных в Python 96 ак.часов 57 000 рублей](https://mipt.ru/cdpo/programs/software/supervised_learning.php?ELEMENT_ID=1911693&clear_cache=Y)  [Программирование на Python МГТУ 40 ак.часов 20 200 рублей](https://www.specialist.ru/course/python1-a) |
| 2.7 | Минимальное количество человек на курсе | 100 |
| 2.8 | Максимальное количество человек на курсе | Не ограничено |
| 2.9 | Данные о количестве слушателей, ранее успешно прошедших обучение по образовательной программе | 250 (на этапе апробации) |
| 2.10 | Формы аттестации | Входное тестирование, итоговое тестирование по темам/модулям, выходное тестирование |
|  | Указание на область реализации компетенций цифровой экономики, к которой в большей степени относится образовательная программа, в соответствии с Перечнем областей | Большие данные |

1. **Аннотация программы**

# Согласно исследованию Академии больших данных MADE и российского портала вакансий HeadHunter (<https://corp.mail.ru/ru/press/releases/10476/> ), начиная с 2015 года потребность в специалистах постоянно растет. В 2018 году количество вакансий под заголовком Data Scientist выросло в 7 раз по сравнению с 2015 годом, а вакансий с ключевыми словами Machine Learning Specialist – в 5 раз. При этом в первом полугодии 2019 года спрос на специалистов по Data Science составил 65% от спроса за весь 2018 год. Анализ данных применяется в самых разных областях - розничная торговля, банковское дело, страхование, телекоммуникации, техника, медицина, молекулярная биология, молекулярная генетика, хемоинформатика и другие. Огромное число современных open source библиотек для анализа и визуализации данных, знание которых требуется практически в каждой вакансии специалиста по анализу данных – в том числе scikit-learn, Pandas, NumPy, Plotly написано на языке Python. Все это делает курс «Python для анализа данных» своевременным и актуальным.

В соответствии с профессиональным стандартом «Специалист по большим данным», программа повышения квалификации «Python для анализа данных» направлена на формирование и качественное изменение профессиональных компетенций в области анализа больших данных с использованием существующей в организации методологической и технологической инфраструктуры. В результате освоения программы слушатель приобретает компетенции по созданию и совершенствованию программного кода с использованием специализированных программных средств и библиотек экосистемы Python. Слушатели программы приобретут знания в области основных библиотек для анализа и визуализации данных, использования специализированных сред разработки, алгоритмов машинного обучения, основных архитектур нейросетей, использования библиотек Python для решения задач. Слушатели получат навыки по созданию программного кода на Python, решения задач регрессии и классификации, создания нейросетей с использованием актуальных фреймворков.

Для успешного освоения программы обучающимся необходимо владеть базовыми компетенциями цифровой грамотности, в том числе навыками владения ПК на уровне среднего или продвинутого пользователя; компетенции в области алгоритмизации, в том числе знаниями и умениями в области алгоритмизации (знание основных алгоритмических структур, умения выражать их на псевдокоде и в виде блок-схем) и программирования, базовыми знаниями в области языка Python, а также знаниями основ линейной алгебры, математического анализа и теории вероятностей.

В процессе обучения по программе обучающимся рекомендуется прослушать и конспектировать все видеолекции, выполнять по каждой теме практические задания с автоматизированной проверкой и проработать самостоятельно дополнительные материалы, после этого необходимо выполнить тесты по каждой теме и, для успешного завершения обучения, сдать итоговый тест (не менее 60% верных ответов).

Обучающиеся в результате освоения программы будут знать синтаксис основных библиотек для анализа данных – NumPy, SciPy, Pandas, Sklearn и построения нейронных сетей – Keras, TensorFlow; синтаксис основных библиотек для визуализации данных – Matplotlib, Plotly; особенности работы в среде Jupyter notebook и Colab; основные алгоритмы обучения на размеченных данных; основные алгоритмы обучения на неразмеченных данных; основные архитектуры нейросетей; будут владеть навыками создания программного кода на Python, который визуализирует данные; навыками создания программного кода на Python, который решает задачи регрессии и классификации; навыками создания кода, который решает задачу построения нейросети с использованием фреймворка Keras.

Результаты обучения по программе будут востребованы при решении таких профессиональных задач, как написание программного кода на языке Python для анализа и визуализации данных.

1. ШАБЛОН ОБРАЗОВАТЕЛЬНОЙ ПРОГРАММЫ (ДПО)

**Автономная некоммерческая организация**

**дополнительного профессионального образования**

**«СофтЛайн Эдюкейшн»**

**\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

|  |  |
| --- | --- |
|  | УТВЕРЖДАЮ:  Генеральный директор  АНО ДПО «СофтЛайн Эдюкейшн»  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.Э. Разуваев  «\_\_\_\_» \_\_\_\_\_\_\_\_\_\_\_\_\_\_ 20\_\_ г. |

**Дополнительная профессиональная программа повышения квалификации**

**«Python для анализа данных»**

72 часа

Москва, 2020

**ОБЩАЯ ХАРАКТЕРИСТИКА ПРОГРАММЫ**

**1.Цель программы**

Программа повышения квалификации «Python для анализа данных» направлена на освоение (совершенствование) профессиональных компетенций в области анализа больших данных с использованием существующей в организации методологической и технологической инфраструктуры по направлению развития цифровой экономики «Большие данные».

Программа повышения квалификации «Python для анализа данных» также формирует профессиональные компетенции для решения прикладных задач с помощью языка Python.

**2.Планируемые результаты обучения:**

2.1. Знание (осведомленность в областях)

2.1.1. синтаксис основных библиотек для анализа данных – NumPy, SciPy, Pandas, Sklearn и построения нейронных сетей – Keras, TensorFlow;

2.1.2. синтаксис основных библиотек для визуализации данных – Matplotlib, Plotly;

2.1.3. особенности работы в среде Jupyter notebook и Colab;

2.1.4. основные алгоритмы обучения на размеченных данных;

2.1.5 основные алгоритмы обучения на неразмеченных данных;

2.1.6. основные архитектуры нейросетей.

2.2. Умение (способность к деятельности)

2.2.1. использовать библиотеки Python для решения задач регрессии, классификации, понижения размерности,

2.2.2. использовать библиотеки Keras и TensorFlow для построения нейросети;

2.2.3. использовать библиотеки Matplotlib и Plotly для визуализации данных.

2.3. Навыки (использование конкретных инструментов)

2.3.1. создание программного кода на Python, который визуализирует данные;

2.3.2. создание программного кода на Python, который решает задачи регрессии и классификации;

2.3.3. создание кода, который решает задачу построения нейросети с использованием фреймворка Keras.

**3.Категория слушателей** (возможно заполнение не всех полей)

* 1. Образование: среднее профессиональное и/или высшее образование
  2. Квалификация: не ниже 4
  3. Наличие опыта профессиональной деятельности: разработчики программного обеспечения, руководители служб и подразделений в сфере информационно-коммуникационных технологий, программисты приложений, специалисты аналитических служб.
  4. Предварительное освоение иных дисциплин/курсов /модулей: необходимы знания в области теории вероятностей, математического анализа и линейной алгебры в объёме первого курса вуза.

**4.Учебный план программы «Python для анализа данных»**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль** | **Всего, час** | **Виды учебных занятий** | | |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| 1 | Математика для анализа данных и библиотеки для работы с данными | 12 | 3 | 6 | 3 |
| 2 | Визуализация и статистический анализ данных | 12 | 2 | 6 | 4 |
| 3 | Обучение с учителем | 16 | 3 | 8 | 5 |
| 4 | Методы обучения без учителя | 14 | 2 | 8 | 4 |
| 5 | Нейронные сети | 16 | 2 | 8 | 6 |
| **Итоговая аттестация** | | **2** | **Указывается вид (экзамен, зачёт, реферат и т.д.)** | | |
| **ИТОГО:** | | 72 | Итоговое тестирование | | |

**5.Календарный план-график реализации образовательной программы**

(дата начала обучения – дата завершения обучения) в текущем календарном году, указания на периодичность набора групп (не менее 1 группы в месяц)

|  |  |  |  |
| --- | --- | --- | --- |
| **№ п/п** | **Наименование учебных модулей** | **Трудоёмкость (час)** | **Сроки обучения** |
| **1** | Математика для анализа данных и библиотеки для работы с данными | 12 | 01.11.20-03.11.20 |
| **2** | Визуализация и статистический анализ данных | 12 | 04.11.20-05.11.20 |
| 3 | Обучение с учителем | 16 | 06.11.20-08.11.20 |
| 4 | Методы обучения без учителя | 14 | 09.11.20-11.11.20 |
| 5 | Нейронные сети | 16 | 12.11.20-14.11.20 |
| 6 | Итоговое тестирование | 2 | 15.11.20 |
| **Всего:** | | 72 | 15 календарных дней |

**6.Учебно-тематический план программы «Python для анализа данных»**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **№ п/п** | **Модуль / Тема** | **Всего, час** | **Виды учебных занятий** | | | **Формы контроля** |
| **лекции** | **практические занятия** | **самостоятельная работа** |
| **1** | **Математика для анализа данных и библиотеки для работы с данными** | **12** | **3** | **6** | **3** |  |
| 1.1 | Математика для анализа данных | 4 | 1 | 2 | 1 | Итоговое тестирование по теме |
| 1.2 | Библиотеки для работы с данными | 4 | 1 | 2 | 1 |  |
| 1.3 | Подготовка данных | 4 | 1 | 2 | 1 |  |
| **2** | **Визуализация и статистический анализ данных** | **12** | **2** | **6** | **4** |  |
| 2.1 | Визуализация данных | 6 | 1 | 3 | 2 |  |
| 2.2 | Статистический анализ | 6 | 1 | 3 | 2 | Итоговое тестирование по теме |
| **3** | **Обучение с учителем** | **16** | **3** | **8** | **5** |  |
| 3.1 | Введение. Линейные модели | 4 | 1 | 2 | 1 | Промежуточное тестирование, итоговое тестирование по теме |
| 3.2 | Измерение качества моделей | 6 | 1 | 3 | 2 | Итоговое тестирование по теме |
| 3.3 | Ансамблевые модели | 6 | 1 | 3 | 2 | Промежуточное тестирование, итоговое тестирование по теме |
| **4** | **Методы обучения без учителя** | **14** | **2** | **8** | **4** |  |
| 4.1 | Обучение без учителя. Методы кластеризации | 7 | 1 | 4 | 2 | Промежуточное тестирование, итоговое тестирование по теме |
| 4.2 | Методы понижения размерности | 7 | 1 | 4 | 2 | Промежуточное тестирование, итоговое тестирование по теме |
| **5** | **Нейронные сети** | **16** | **2** | **8** | **6** |  |
| 5.1 | Основы нейронных сетей | 8 | 1 | 4 | 3 | Промежуточное тестирование, итоговое тестирование по теме |
| 5.2 | Архитектуры нейронных сетей | 8 | 1 | 4 | 3 | Промежуточное тестирование, итоговое тестирование по теме |
|  | Итоговая аттестация | 2 |  |  | 2 | Итоговое тестирование по теме |

**7. Учебная (рабочая) программа повышения квалификации «Python для анализа данных»**

Модуль 1. Математика для анализа данных и библиотеки для работы с данными (12 часов)

Тема 1.1 Математика для анализа данных (4 часа)

Содержание темы: Повторение основных понятий линейной алгебры (кратко)

Тема 1.2 Библиотеки для работы с данными (4 часа)

Содержание темы: Знакомство с библиотеками NumPy и SciPy

Тема 1.3 Подготовка данных (4 часа)

Содержание темы:

- Знакомство с библиотекой Pandas

- Объект pandas.Series

- Объект pandas.DataFrame

- Группировка данных

- Работа с несколькими таблицами

- Преобразование признаков

Модуль 2. Визуализация и статистический анализ данных (12 часов)

Тема 2.1. Визуализация данных (6 часов)

Содержание темы:

- Расширенная визуализация с matplotlib

- Визуализация с pandas

- Интерактивная визуализация с plotly

Тема 2.2 Статистический анализ (6 часов)

Содержание темы:

- Определение вероятности

- Случайная величина

- Показатели центра распределения

- Нормальное распределение

- Центральная предельная теорема

- Зависимость между случайными величинами

- Распределение Стьюдента

- Статистика в SciPy

- Доверительный интервал

- Проверка гипотез и распределение Стьюдента

Модуль 3. Обучение с учителем (16 часов)

Тема 3.1 Введение. Линейные модели (4 часа)

Содержание темы:

- Виды машинного обучения

- Линейная регрессия

- Функционал качества и градиентный спуск

- Логистическая регрессия

- Применение линейных моделей

Тема 3.2 Измерение качества моделей (6 часов)

Содержание темы:

- Данные и переобучение

- Метрики качества

- Применение метрик качества

Тема 3.3 Ансамблевые модели (6 часов)

Содержание темы:

- Решающие деревья

- Случайный лес

- Градиентный бустинг

- Применение ансамблевых моделей

Модуль 4. Методы обучения без учителя (14 часов)

Тема 4.1 Обучение без учителя. Методы кластеризации (7 часов)

Содержание темы:

- Задача кластеризации, группы методов

- Метод K-средних

- Иерархическая кластеризация. Агломеративный алгоритм

- DBSCAN

- Оценки качества кластеризации

Тема 4.2 Методы понижения размерности (7 часов)

Содержание темы:

- Метод главных компонент (Principal Component Analysis)

- Сингулярное разложение матрицы и связь с PCA

- Многомерное шкалирование

- T-SNE

Модуль 5. Нейронные сети (16 часов)

Тема 5.1 Основы нейронных сетей (8 часов)

Содержание темы:

- Определение нейросети

- Обучение нейросети

Тема 5.2 Архитектуры нейронных сетей (8 часов)

Содержание темы:

- Свёрточные сети

- Рекуррентные сети

- Современные архитектуры

- Введение в Tensorflow

- Классификация изображений на Tensorflow

**Описание практико-ориентированных заданий и кейсов**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Номер темы/модуля** | **Наименование практического занятия** | **Описание** |
| 1 | Тема 1.3 Подготовка данных | Python для анализа данных | Практическое задание на отработку умения анализировать данные в библиотеке Pandas.  Для выполнения данного теста нужно скачать датасет [201809-citibike-tripdata.csv](https://s3.amazonaws.com/tripdata/201809-citibike-tripdata.csv.zip) расположенный по адресу <https://s3.amazonaws.com/tripdata/201809-citibike-tripdata.csv.zip> и найти:  - общее количество строк и столбцов в датасете - указать первым число строк, вторым - число столбцов  - среднюю длину поездок в минутах(столбец tripduration) c точностью до 2 знака  - поездки, которые начались и закончились в той же самой станции  - идентификатор велосипеда (bikeid), у которого в среднем продолжительность поездок выше, чем у всех остальных  - количество строк, в которых отсутствуют данные о start station id?  - средняя продолжительность поездки в минутах в зависимости от типа подписки c точностью до 2 знака?  - среднюю длину поездок в километрах с точностью до 2 знака, предварительно выкинув замкнутые траектории  - станцию (start station id) с максимальным количеством отправлений с 18 до 20 вечера  - идентификаторы станций(end station id), в которые приезжают с 6 до 10 утра |
| 2 | Тема 2.2 Статистический анализ | Доверительный интервал | Практическое задание на отработку умения нахождения доверительного интервала.  Нужно загрузить данные из предоставленного файла a\_b\_testing.csv при помощи функции read\_csv из библиотеки pandas и найти доверительный интервал для средней конверсии пользователей из контрольной выборки с уровнем значимости 95% |
| 3 | Тема 2.2 Статистический анализ | Задача на вычисление статистики Стьюдента по формуле | Практическое задание на отработку умения нахождения значения статистики Стьюдента в предположении независимости выборок по предоставленной формуле. |
| 4 | Тема 2.2 Статистический анализ | Задача на вычисление статистики Стьюдента используя библиотеку SciPy | Практическое задание на отработку умения нахождения значения статистики Стьюдента в предположении независимости выборок используя функцию stats.ttest\_ind из библиотека SciPy. |
| 5 | Тема 3.1 Введение. Линейные модели | Применение линейной регрессии | Практическое задание на отработку умения работать c линейной регрессией и нормализацией признаков. Загрузите данные из предоставленного файла data.csv, который содержит 100 признаков f1, f2, ..., f100 и целевую переменную target. Для загрузки данных используете функцию read\_csv из библиотеки pandas. Выделите матрицу признаков и целевую переменную из загруженных данных.  Далее, разделите загруженный датасет на тренировочную и тестовую выборку. В качестве ответа answer1 приведите это значение с точностью до двух знаков после запятой.  Далее вам предлагается нормализовать данные используя StandardScaler из sklearn.preprocessing. Приведите значение среднеквадратической ошибки в качестве ответа answer2 с точностью до двух знаков после запятой. |
| 6 | Тема 3.2 Измерение качества моделей | Настройка параметра регуляризации в Гребневой регрессии | Практическое задание на отработку умения работать гребневой регрессии. Модель гребневой регрессии представлена классом Ridge из sklearn.linear\_model. Конструктор этого класса содержит аргумент alpha, обратное значение которого соответствует параметру регуляризации 𝜆. При использовании данной модели установите значение аргумента random\_state=42. Найдите оптимальное значение коэффициента alpha в диапазоне {10, 20, 30, ..., 90}. |
| 7 | Тема 3.2 Измерение качества моделей | Практическое задание на отработку умения работать с методом регрессии LASSO. | Загрузите набор данных Diabetes, используя функцию load\_diabetes из sklearn.datasets. Diabetes содержит в себе информацию о различных признаках, таких как возраст, пол, среднее кровяное давление человека и прочие. В качестве целевой переменной выступает количественный показатель прогрессирования заболевания диабет через год после определенного периода (baseline).  Метод LASSO представлен классом Lasso из sklearn.linear\_model. Обучите эту модель на всем наборе данных Diabetes. В качестве ответа answer3 на это задание приведите отношение числа выбранных моделью признаков к их общему числу. Ответ округлите до одного знака после запятой.  Далее обучите модель Lasso с параметром конструктора random\_state=42 на тренировочной выборке датасета Boston House Prices и посчитайте значение среднеквадратической ошибки на тестовой выборке. Сравните это значение с результатом, полученным Гребневой регрессией. |
| 8 | Тема 3.3 Ансамблевые модели Тема 3.3 Ансамблевые модели | Решение задачи классификации при помощи случайного леса (Random Forest) | Практическое задание на отработку умения работать с алгоритмом классификации Random Forest. Загрузите датасет Wine Data Database с помощью функции load\_wine из модуля sklearn.datasets.  Модель случайного леса для классификации представлена классом RandomForestClassifier из модуля sklearn.ensemble. Конструктор этого класса содержит аргумент n\_estimators, который соответствует количеству базовых алгоритмов в случайном лесе. Целью этого задания будет настройка этого параметра. Далее сравните модель градиентного бустинга GradientBoostingClassifier из sklearn.ensemble с логистической регрессией LogisticRegression из sklearn.linear\_model на этой выборке. Используете параметр random\_state=42 при создании экземпляров классов. Какая из моделей работает лучше? Приведите лучшую оценку, округленную до трех знаков после запятой, в качестве ответа answer2 на это задание. Какие выводы из этого можно сделать? |
| 9 | Тема 3.3 Ансамблевые модели | Решение задачи регрессии при помощи градиентного бустинга | Практическое задание на отработку решения задачи регрессии при помощи градиентного бустинга.  Загрузите уже известную вам выборку Boston House Prices и разделите ее случайным образом на тренировочную и тестовую выборку. Какая из моделей показывает наименьшее значение среднеквадратической ошибки на тестовых данных? В качестве ответа answer3 приведите это значение, округленное до двух цифр после запятой. |
| 10 | Тема 4.1 Обучение без учителя. Методы кластеризации | Применение метода Kmeans для кластеризации данных | Практическое задание на отработку умения применять алгоритм Kmeans для кластеризации данных. Загрузите данные из предоставленного файла weather-check.csv  Обучите метод к-средних с n\_clusters от 2 до 20 и random\_state=123  Создайте массив scores  Для каждого запуска вычислите silhouette\_score полученной кластеризации  В переменную answer1 сохраните значение наибольшего значения из scores c точностью до сотых |
| 11 | Тема 4.1 Обучение без учителя. Методы кластеризации | Применение метода агломеративной кластеризации для кластеризации данных | Практическое задание на отработку умения применять алгоритм агломеративной кластеризации для кластеризации данных.  Прогоните метод с параметрами method='average', metric='cosine'  Визуализируйте полученную дендрограмму  Выделите 5 кластеров из полученной иерархии  В переменную answer2 запишите долю мужчин во втором по величине кластере с точностью до сотых |
| 12 | Тема 4.1 Обучение без учителя. Методы кластеризации | Применение метода DBSCAN для кластеризации | Практическое задание на отработку умения применять алгоритм DBscan для кластеризации данных. Возможно вы уже успели заметить, что часть респондентов практически не отвечали на заданные вопросы - то есть в соответствующих графах стоит символ "-". "Молчунами" будем называть участников опроса, которые не ответили ни на один вопрос. Таких респондентов можно сразу исключить из дальнейших исследований.  Переберите значение параметра eps в интервале [0.1, 1] с шагом 0.1  Остальные параметрамы установите следующим образом min\_samples=20, metric='cosine'  В переменную answer3 запишите значение eps (c точностью до **сотых**) при котором алгоритм выделит всех "молчунов" в отдельный кластер с выбросами |
| 131 | Тема 4.2 Методы понижения размерности | Метод главных компонент | Практическое задание на отработку умения применять метод главных компонент.  В этом задании мы применим метод главных компонент на многомерных данных и постараемся найти оптимальную размерность признаков для решения задачи классификации  Исходными [данными](http://archive.ics.uci.edu/ml/machine-learning-databases/auslan2-mld/auslan.data.html) являются показания различных сенсоров, установленных на руках человека, который умеет общаться на языке жестов.  В данном случае задача ставится следующим образом: по показаниям датчиков (по 11 сенсоров на каждую руку) определить слово, которое было показано человеком. |
| 14 | Тема 5.1 Основы нейронных сетей | Реализация перцептрона | Практическое задание на отработку умения обучить перцептрон.  Перцептрон - это модель, предложенная Френком Розенблаттом в 1957 году и являющаяся прообразом современных нейронных сетей. По своей сути она представляет из себя значительно упрощенную схему восприятия информации мозгом. Целью этого практического задания будет реализация собственной модели перцептрона. |
| 15 | Тема 5.2 Архитектуры нейронных сетей | Задача классификации цифр из датасета mnist с помощью нейронной сети | Практическое задание на отработку умения создать нейросеть для классификации цифр.  В этом практическом заданий мы будем решать задачу классификации цифр на датасете mnist с помощью полносвязной и сверточной нейронной сети. Для этого мы будем использовать надстройку над tensorflow, которая называется keras. Для начала обсудим данные. mnist датасет состоит из черно-белых изображений цифр размером 28×28 пикселей. В данном случае, мы работаем с одним каналом, хотя в случае цветных изображений, общее число каналов равно трем. Загрузим наши данные используя функцию load\_data объекта mnist из модуля keras.dataset. Перед выполнением этого задания убедитесь, что ваша версия tensorflow >= 1.4. |

**8.Оценочные материалы по образовательной программе**

**8.1. Вопросы тестирования по модулям**

|  |  |  |  |
| --- | --- | --- | --- |
| **№ модуля** | **Вопросы входного тестирования** | **Вопросы промежуточного тестирования** | **Вопросы итогового тестирования** |
| Входное тестирование | 1. Какая функция выводит что-либо в консоль?  1. print()  2. out()  3. log()  4. write()  2. Как не может называться переменная в Python?  1. \_name  2. \_\_name  3. name  4. 1name  3. Что будет напечатано в результате выполнения кода  x = 20  if x > 10:  num = 0  else:  num = 10  y=30  print(num)  1. 10  2. 0  3. 20  4. 30  4. Что будет напечатано в результате выполнения кода:  s = 0  for i in range(10)  s+=i  print(s)  1. 1 2 3 4 5 6 7 8 9 10  2. 45  3. 55  4. 25  5. К чему приведет обращение к непустому списку по индексу «-1»  1. Вернется первый элемент  2. Ошибка KeyError  3. Ошибка IndexError  4. Вернется последний элемент  6. Какое число будет напечатано после выполнения следующего кода  def sum\_square(a, b):  return (a\*\*2 + b\*\*2)  print(sum\_square(3, 4))  1. 7  2. 5  3. 25  4. 19  7. Чему равен определитель матрицы( мы записываем ее как список списков) [[3,2],[7,5]]?  1. 1  2. -29  3. -7  4. 11  8. Векторы a = (1,2,3) и b = (2,4,6) являются:  1. Линейно зависимыми  2. Линейно независимыми  3. Неколлинеарными  4. Ортогональными  9. Решением системы уравнений  x + y = 10  x - y = 0  будет пара чисел:  1. 5, 5  2. 10, 0  3. 0, 10  4. 10, 10  10. Какая из функций не является непрерывной?  1. f(x) = x^2  2. f(x) = sin(x)  3. f(x) = 1/x  4. f(x) = cos(x)  11. Чему равно значение функции f(x) = sin(x) + cos(x) в точке 0  1. 2  2. 0  3. 1  4. 2  12. Найдите значение производной функции f(x) = x^2 + x^3 в точке 1.  1. 2  2. 5  3. 0  4. 6  13. Найдите значение производной функции f(x) = sin(cos(x)) в точке 0.  1.1  2.2  3.0  4. -1  14.  Если производная функции положительна в каждой точке некоторого интервала, то функция на этом промежутке:  1. убывает  2. возрастает  3. положительна  4. постоянна  15. Наименьшее значение функции f(x) = x^2 + 2 равно:  1.3  2.0  3. -1  4. 2  16. В турнире по шахматам участвуют 3 человека. Сколькими способами может выглядеть финальная таблица, если считать, что делёжки мест не было?  1. 2  2. 4  3. 6  4. 3  17. На шахматном турнире каждый сыграл с каждым по одной партии. Сколько было сыграно партий, если в турнире участвовало 15 человек?  1. 14  2.225  3.210  4.29  18. Брошен игральный кубик. Какова вероятность, что выпадет нечётное число?  1. 1/2  2. 1/5  3. 1/3  4. 1/4  19. Монетка брошена два раза. Какова вероятность, что оба раза выпадет герб?  1. 1/2  2. 1/4  3. 1/3  4. 1  20. Вероятность того, что новая лампочка будет бракованной равна 0,05. Покупатель, не глядя, берёт одну лампочку из коробки. Найдите вероятность того, что эта лампочка будет исправной.  1. 0,05  2. 0,95  3. 0,9  4. 0,85 |  |  |
| 1.1 | нет | нет | 1. Является ли данная система векторов линейно независимой?  $$\vec{a} = \{\ 2, -1 , 0 \} , \vec{b} = \{\ 0, 1 , 3 \} , \vec{c} = \{\ 4, -3 , -3 \}$$  -Вектора линейно зависимы  -Вектора линейно независимы  2. Найти детерминант матрицы  $$\begin{pmatrix}1 & \quad 13 \\2 & \quad 3 \end{pmatrix}$$  -\(0\)  -\(-23\)  -\(23\)  -\(-1\)  3. Найдите результат перемножения матриц \(A\) и \(B\)  -$$A = \begin{pmatrix}\quad 42\quad&1\quad &6\\\\1&0\quad&-3\\\\3&5\quad&1\ \end{pmatrix}$$, $$B = \begin{pmatrix}-1 \quad & \quad 1\\0 \quad & \quad 7 \\9 \quad & -2 \end{pmatrix}$$  -$$B = \begin{pmatrix}12 \quad & \quad 37\\-28 \quad & \quad 7 \\6 \quad & 36 \end{pmatrix}$$  $$B = \begin{pmatrix}12 \quad & \quad -28 & \quad 6\\37 \quad & \quad 7\quad&36 \end{pmatrix}$$  -Несоответствие количества строк и столбцов, нельзя перемножить матрицы  -$$B = \begin{pmatrix}12 \quad & \quad -37\\28 \quad & \quad 7 \\6 \quad&36 \end{pmatrix}$$  4. Имеет ли данная матрица обратную?  \begin{pmatrix}\quad 2 \quad & 1 \quad & -3 \ \\\ 1 & 0 \quad & 0 \ \\\ 1 & 0 \quad & 1 \ \end{pmatrix}  -Нет  -Да  5. Найти собственные числа матрицы \(A\)  $$A = \begin{pmatrix}0 & \quad 3 \\3 & \quad 8 \end{pmatrix}$$  -\(1, -3\)  -\(-1\)  -\(-1, 9\)  -\(0, 8\)  6. К какому виду относится матрица \(A\)?  $$A = \begin{pmatrix}0.2 & \quad 0.98 \\0.98 & \quad -0.2 \end{pmatrix}$$  -Симметричная  -Диагональная  -Ортогональная  -Вырожденная \((detA = 0)\)  7. Какое линейное преобразование задает данная матрица?  \(\begin{bmatrix}k & \quad 0 \\0& \quad 5 \end{bmatrix}, 0 < k < 1\)  -Растяжение по оси \(x\) в \(k\) раз и в \(5\) раз по оси \(y\)  -Растяжение по оси \(y\) в \(5\) раз  -Сжатие по оси \(x\) в \(k\) раз  8. Какое линейное преобразование задает данная матрица?  \begin{bmatrix}1 & \quad 0 \\k & \quad 1 \end{bmatrix}  -Растяжение по оси \(x\) в \(k\) раз  Сдвиг по оси \(y\) на \(kx\)  -Не задает никакое преобразование  -Поворот  9. Если поменять местами две строки в матрице, то детерминант матрицы:  -останется неизменным  -будет равен бесконечности  -значение детерминанта поменяет знак  -будет равен нулю  10. Матрица размера \(m\) x \(n, m <n\), может быть симметричной?  1)Да  2) Нет |
| 1.2 | нет | нет | нет |
| 1.3 | нет | нет | нет |
| 2.1 | нет | нет | нет |
| 2.2 | нет | нет | Какие из перечисленных случайных величин являются непрерывными?  -Количество покупок в интернет-магазине  -Время ожидания автобуса на остановке  -Время зарядки ноутбука в часах  -Нет правильного варианта  Пусть скорость автомобиля описывается нормально распределенной случайной величиной с параметрами \(N(40, 5)\). Какова вероятность того, что скорость автомобиля будет равна 60 км/ч?  -\(0.34\)  -\(0\)  -\(0.89\)  -\(0.001\)  Корреляция позволяет оценить следующую зависимость между случайными величинами:  -Нелинейная зависимость  -Линейная зависимость  -Корреляция не оценивает зависимость между случайными величинами  -Корреляция оценивает только положительную зависимость между величинами  Правило 3 сигм показывает, что:  -Нормальное распределение имеет вид \(N(3\sigma, 1)\)  -\(95%\) данных лежит в интервале \(+/- 3\sigma\)  -Нормальное распределение имеет вид \(N(0.3\sigma, 1)\)  -\(99.7%\) данных лежит в интервале \(+/- 3\sigma\) |
| 3.1 | нет | В чем отличие задачи регрессии от задачи классификации?  -Одна это задача обучения с учителем, а вторая классификации (неверно: обе этих задачи обучения с учителем)  -Особого отличия нет, это просто устоявшаяся терминология. В обоих случаях мы прогнозируем вещественное число. Например, стоимость квартиры и вероятность отдать кредит. (неверно, классификаторы не обязаны возвращать вероятность, но дискретное число, скажем, класс, 0 или 1)  -классификаторы не обязаны возвращать вероятность, но дискретное число, скажем, класс, 0 или 1 (правильно)  О чем свидетельствует нулевой вес отдельного коэффициента в линейной регрессии?  -Нет линейной зависимости между целевой переменной и этим признаком (правильно. Вообще говоря, получить именно нулевой коэффициент довольно сложно, если не использовать регуляризацию)  -У нас есть высококореллированные признаки (неверно)  -Мы неправильно подготовили признаки для модели (неверно)  При использовании метода градиентного спуска, мы подбираем веса таким образом, чтобы двигаться в сторону:  -Градиента (неправильно, градиент указывает направление наискорейшего роста функции)  -Антиградиента (правильно)  Какая функция используется для получения вероятности принадлежности классу в двуклассовой логистической регрессии? (наверно надо будет привести формулы)  -sign (неправильно, эта функция имеет лишь 3 числа в области значений -- -1, 0, 1)  -sigmoid (правильно)  -softmax (неправильно, эта функция используется в случае многоклассовой классификации) | 1. Выберите пункты обучения с учителем:  -Поиск подозрительных транзакций по банковской карте (отмеченное неверно -- это поиск аномалий)  -Прогнозирование курса доллара к рублю (правильно)  -Обучение бота для игры в Dota 2 для International 2018 (отмеченное неверно -- это обучение с подкреплением)  -Сегментация абонентов сотовой связи по их активности (отмеченное неверно -- это кластеризация, обучение без учителя)  -Оценка вероятности клиента интернет-магазина сделать покупку в следующем месяце (правильно)  2. Может ли линейная регрессия аппроксимировать нелинейные зависимости?  -Не может, линейная регрессия, как следует из названия аппроксимирует исключительной линейную функцию(-)  -Не может, только линейные или кусочно-линейные зависимости(-)  -Может. Если ввести, например, квадрат исходного признака, то получится квадратичная аппроксимация целевой переменной.  3. Какая функция ошибки используется для оптимизации коэффициентов линейной регрессии градиентным методом?  -Средняя абсолютная ошибка, MAE (неверно, она не является дифференцируемой на всей области определения)  -Средняя квадратичная ошибка,  MSE (верно)  -Для линейной регрессии имеется аналитическое решение, которое всегда предпочтительнее использовать (неверно, хотя оно действительно позволяет получить точный ответ, вычисления становятся громоздкими и затратными в случае высокой размерности)  4. Какая функция ошибки используется в логистической регрессии?  -Средняя абсолютная ошибка, MAE (неверно)  -Средняя квадратичная ошибка,  MSE (неверно)  -Logloss, она же кросс-энтропия (верно)  5. Вспомните формулу логистическую функцию ошибки (тут надо вставить функцию логлосс из лекции). В каких пределах она изменяется?  -От нуля до бесконечности (верно, поэтому одно совершенно неправильное значение может свести на нет все попытки оптимизировать функцию ошибки. На практике вместо бесконечности используют замену на большое число)  -От минус бесконечности до нуля (неверно)  -От минус бесконечности, до плюс бесконечности (неверно)  6. Для предсказания вероятности принадлежности объекта к позитивному классу в библиотеке sklearn для логистической регрессии используется метод:  -fit() (неверно, этот метод используется для обучения модели)  -predict() (неверно, этот метод используется для получения класса объекта)  -predict\_proba() (верно) |
| 3.2 | нет | нет | 1) Для модели с набором параметров wi выберите верные утверждения про L1 регуляризацию:  • В функцию ошибки добавляется член, пропорциональный ∑wi2  • В функцию ошибки добавляется член, пропорциональный ∑|wi|  • Веса при некоторых признаках могут полностью занулиться. Происходит отбор признаков.  • Используется для борьбы с переобучением.  • Возможна только в линейных моделях  2) Для модели с набором параметров wi выберите верные утверждения про L2 регуляризацию:  • В функицонал ошибки добавляется член, пропорциональный ∑wi2  • В функионал ошибки добавляется член, пропорциональный ∑|wi|  • Веса при некоторых признаках могут полностью занулиться. Происходит отбор признаков.  • Используется для борьбы с переобучением.  • Возможна только в линейных моделях  3) Какие из указанных ниже метрик может использовать в качестве функции потерь градиентный спуск?  Подсказка: для корректной работы градиентного спуска необходимо, чтобы функция потерь всегда имела однозначно определённую частную производную, не тождественную 0 по каждому предсказанному значению ypred на объекте.  • MSE  • accuracy  • f-measure  • precision  • logloss  • MAE  4) Рассмотрим следующую confusion matrix, полученную после того как некая модель на 1000 объектов решила задачу бинарной классификации:  Confusion matrix 0 1  0 920 5  1 60 15  Выберите для данной confusion matrix все верные утверждения:  • Precision более чем в три раза превосходит Recall  • Accuracy > 0.93  • F-measure > 1/3  • Классы в выборке хорошо сбалансированы  • Среди Precision, Recall, Accuracy, F-measure самой большой вышла Accuracy  5) Предположим, что вы решаете задачу бинарной классификации об отказе оборудования не предприятии. Вы должны предсказывать 1, если оборудование откажет в течение ближайших шести месяцев и 0, если этого не случится. Вы знаете, что отказы оборудования – редкое событие. Какую метрику НЕ стоит использовать для оценки качества модели?  • Recall  • F-measure  • Precision  • Accuracy  6) Предположим, вы решаете задачу оттока (churn) игроков из онлайн-игры. Вы должны предсказать число как можно ближе к 1, если игрок уйдёт из игры и как можно ближе к 0, если он не уйдёт из неё. Уходом считается отсутствие игрока в течение недели. Для 5-ти игроков ваша модель дала следующие предсказания: [0.6 , 0.4 , 0.8 , 0.4 , 0.8] при том, что реальные метки выглядели как [ 0 , 0 , 1 , 1 , 1]. Для какой метрики её значение не изменится, если предсказания модели [0.6 , 0.4 , 0.8 , 0.4 , 0.8] вы округлите до  [1 , 0 , 1 , 0 , 1 ] ?  • MSE  • logloss  • MAE  • Площадь под кривой ROC-AUC  • RMSE  7) Сколько раз обучается модель при кросс-валидации на 5 фолдах?  • 1  • 4  • 5  • ни разу  8) Выберите наиболее верное из следующих увтерждений:  • Кросс-валидация борется с переобучением  • Кросс-валидация позволяет оценить обощающую способность модели  • Кросс-валидация оценивает качество, используя ансамбль, построенный с помощью исходной модели.  • Вместо обычной валидации на отложенной выборке, всегда лучше применять кросс-валидацию. |
| 3.3 | нет | 3.3.1) Выберите верные утверждения.  • Решающее дерево содержит в себе набор критериев, которые позволяют соотнести новый объект с листом этого дерева.  • В случае задачи классификации листями решающего дерева являются классы.  • Увеличение глубины дерева – это способ избежать его переобучения на тренировочной выборке.  3.3.2) Выберите верные утрверждения.  • Алгоритм Bootstrap позволяет создавать обучающие выборки без повторяющихся объектов.  • Случайный лес строится из неглубоких решаюших деревьев для того, чтобы избежать переобучения.  • Построение решающих деревьев в случайном лесе происходит независимо.  • Число решающих деревьев в случайном лесе влияет на качество модели.  3.3.3) Выберите верные утверждения.  • В градиентном бустинге базовые алгоритмы строятся зависимо.  • Увеличение числа базовых алгоритмов в градиентном бустинге всегда приводит к улучшению качества модели на новых данных. | нет |
| 3.3 | нет | нет | 1. Выберете верные утверждения.  • Решающие деревья склонны к переобучению.  • Результатом классификации нового объекта в решающем дереве является класс, содержащийся в листе этого дерева.  • Решающие деревья не используются в качестве регрессионных моделей.  2. Каким образом может достигается различие между решающими деревьями в алгоритме случайного леса?  • Использованием случайного подмножества признаков при разбиении объектов в вершинах решающего дерева.  • Выбором случайного подмножества объектов при обучении решающего дерева.  • Всем из вышеперечисленного.  3. Какие из пунктов описывают достоинства случайного леса?  • Существуют методы оценки предсказательной способности признаков в модели.  • Случайный лес не склонен к переобучению при увеличении числа базовых алгоритмов.  • В случайном лесе обычно используются глубокие решающие деревья.  • В этой модели решающие деревья строятся независимо, что делает возможным использование методов паралелльных вычислений.  • Все из вышеперечисленных.  4. В отличие от случайного леса в градиентном бустинге над случайными деревьями  • базовые алгоритмы строятся независимо.  • каждый следующий алгоритм пытается исправить ошибки предыдущего.  • необходимо использовать глубокие деревья.  • модель может переобучаться при увеличении числа решающих деревьев.  5. Как сохранить соотношения классов при случайном разбиении выборки на тренировочную и тестовую при помощи функции train\_test\_split из библиотеки scikit-learn?  • Эта опция включена по-умолчанию.  • Установить параметр shuffle = True.  • Нужно задать собственное числовое значение аргумента функции random\_state.  • Используя параметр stratify. |
| 4.1 | нет | 4.1.2) Выберите верные утверждения  ● Кластеризация = Классификация  ● Кластер - группа из “похожих” друг на друга объектов  ● Плотностные методы могут найти кластеры произвольной формы  ● В результате иерархической кластеризации можно получить не просто разбиение на кластеры а иерархию целиком  4.1.3) Выберите верные утверждения  ● В стандартном методе К-средних количество кластеров задается исследователем  ● Пусть дано 5 объектов. Разбиения на 3 кластера C1 = [1,2,2,3,1] и C2 = [2,1,1,3,2] эквиваленты  ● К-means++ - это отдельный алгоритм кластеризации  ● Метод локтя - догма!  4.1.5) Выберите верные утверждения  ● Иерархическая кластеризация бывает дивизивной и агломеративной  ● Для того, чтобы выполнить иерархическую кластеризацию с Single Linkage достаточно знать только попарные расстояния между объектами  ● Кофенетическое расстояние считается для всех пар кластеров  4.1.7) Выберите верные утверждения  ● Количество кластеров в DBSCAN задается исследователем  ● Метод DBSCAN может находить объекты-выбросы  ● Если поставить значение параметра min\_pts=1 то все объекты окажутся в одном кластере  ● Для начала конструирования кластера из текущей точки необходимо, чтобы в ее окрестности было достаточно других точек | 1)  Какие из данных проблем могут решаться с помощью кластеризации?  ● Сегментация посетителей интернет магазина  ● Прогнозирование спроса на товары на следующей неделе  ● Выявление аномальных наблюдений  ● Выделение групп единомышленников по результатам опроса  2)  Пусть дано 100 объектов случайно распределенных на плоскости. Вы запускаете алгоритм K-Means с K=3 и случайной инициализацией центроидов на месте одного из объектов. Сколько различных инициализаций возможно?  \* Один и тот же объект не может быть выбран более чем 1 раз.  ● 161700 (-)  ● 970200 - количество размещений 3 по 100  ● 53900 (-)  ● 107800 (-)  3)  Пусть дано 100 объектов на которых была запущена иерархическая кластеризация. По дендрограмме мы видим, что на каждом шаге к наибольшему кластеру добавляется по одному объекту. Является ли это признаком удачной кластеризации и почему?  ● Скорее да, алгоритм же выполнился без ошибок  ● Скорее да, нужно только умело подобрать порог отсечения  ● Скорее нет, потому что добавление по одному объекту плохо показывает иерархическую структуру в данных  ● Скорее нет, возможно в данных есть ошибка или параметры алгоритма выбраны неудачно  4)  При фиксированном min\_pts с увеличением eps  ● Количество кластеров будет увеличиваться  ● Количество кластеров будет уменьшаться  ● Аномальных объектов будет больше  ● Аномальных объектов будет меньше  5)  Выберите верные утверждения  ● Чем больше значение Dunn Index, тем разбиение считается лучше  ● Чем больше значение Daves-Bouldin Index, тем разбиение считается лучше  ● Чем больше значение Silhouette Index, тем разбиение считается лучше  ● Rand Index (Adjusted Rand Index) можно вычислить, только если количество кластеров между разбиениями совпадает |
| 4.2 | нет | 4.2.1) Выберите верные утверждения  ● Понижение размерности - это уменьшение количества признаков  ● Понижение размерности - это уменьшение количества объектов  ● Понижение размерности можно использовать для удобной визуализации  4.2.2) Выберите верные утверждения  ● Новые признаки равны линейной комбинации исходных признаков  ● Матрица ковариаций - квадратная матрица  ● Перед применением PCA признаки надо центрировать (и шкалировать)  ● Количество компонент задается исследователем  4.2.3) Выберите верные утверждения  ● Для того, чтобы воспользоваться PCA можно применить сингулярное разложение к матрице ковариаций и умножить матрицу исходных признаков на матрицу V\_k  ● Для того, чтобы воспользоваться PCA можно применить сингулярное разложение к матрице с исходными признаками и умножить их на матрицу V\_k  ● Объясненная дисперсия соответствующей компоненты равна значению собственного числа при соответствующей собственном векторе матрицы ковариации  ● Полное сингулярное разложение точно равно исходной матрице?  4.2.5) Выберите верные утверждения  ● Классический подход в многомерном шкалировании эквивалентен методу главных компонент  ● Если метод многомерного шкалирования работает хорошо, то точки на диаграмме стресса будут расположены близко к диагонали  ● В неметрическом многомерном шкалировании важно сохранение порядка а не значений попарных расстояний  4.2.6) TSNE  Выберите верные утверждения  ● Разные запуски TSNE могут привести к разным результатам  ● На одних и тех же данных PCA и TSNE дадут одинаковый результат  ● TSNE нацелен на сохранение локальной структуры в данных  ● TSNE - это метод обучения с учителем | Выберите верные утверждения  ● PCA позволяет перейти в пространство с размерностью меньшей или равной размерности исходного пространства  ● Шкалирование признаков не влияет на результат применения PCA  ● Сдвиг признаков не влияет на результат применения PCA  2)  Выберите верные утверждения  ● TSNE является методом многомерного шкалирования  ● На одних и тех же данных PCA и TSNE дадут одинаковый результат  ● TSNE нацелен на сохранение локальной структуры в данных  ● TSNE - это метод обучения с учителем  3) Выберите верные утверждения  ● Для того, чтобы воспользоваться PCA можно применить сингулярное разложение к матрице ковариаций и умножить матрицу исходных признаков на матрицу V\_k  ● Для того, чтобы воспользоваться PCA можно применить сингулярное разложение к матрице с исходными признаками и умножить их на матрицу V\_k  ● Объясненная дисперсия соответствующей компоненты равна значению собственного числа при соответствующей собственном векторе матрицы ковариации  ● Если появятся новые наблюдения, на которых PCA еще не обучался, то для их переноса в пространство меньшей размерности алгоритм нужно обучать заново  4) Выберите верные утверждения  ● Классический подход в многомерном шкалировании эквивалентен методу главных компонент  ● Если метод многомерного шкалирования работает хорошо, то точки на диаграмме стресса будут расположены близко к диагонали  ● TSNE - это метод неметрического многомерного швалирования |
| 5.1 | нет | 5.1.2) Какие пункты описывают причины недавней популярности нейронных сетей?  • Появление больших объемов данных, связанных с развитием интернет-технологий.  • Недавнее появление первых моделей нейронных сетей.  • Развитие вычислительных мощностей и ресуров.  • Все пункты из вышеперечисленных.  5.1.3) Выберите верные утверждения.  • Нейронная сеть преставляет собой суперпозицию нелинейных функций.  • Нейронная сеть состоит из большого количества нейронов, объединенных в слои.  • Результатом обучения нейронной сети является нахождение оптимального количества нейронов на каждом слое.  • Все утверждения из вышеперечисленных являются верными.  5.1.4) Как называется метод вычисления градиента, который используется при обновления весов нейронной сети в процессе обучения?  • Forward Pass.  • Forward Propagation.  • Backpropagation или алгоритм обратного распространения ошибки.  • Stochastic Gradient Descent или стохастический градиентный спуск. | 1) Выберите верные утверждения. Нейронные сети  • требуют большого количества данных при обучении.  • не переобучаются.  • требуют вычислительных ресурсов.  • всегда сходятся.  2) Функция f(x) = max(x, 0) является функцией активации, которая называется  • сигмоидой.  • гиперболическим тангенсом.  • ReLU.  • линейной.  3) Что такое Dropout?  • Это процесс генерации новых элементов для улучшения качества обучения.  • Это разновидность регуляризации для нейронных сетей.  • Добавление модифицированных данных в тренировочную выборку.  • “Отключение” части нейронов в процессе обучения, которое приводит к улучшению качества нейронной сети по аналогии с ансамблевыми моделями.  4) В чем заключается идея метода Early stopping?  Остановка процесса обучения нейронной сети  • при падении качества на валидационной выборке.  • при падение качества на тренировочной выборке.  • перед началом процессом переобучения.  • через строго заданное количество итераций. |
| 5.2 | нет | 5.2.1) Выберите верные утверждения о сверточных нейронных сетях.  • Сверточная нейронная сеть – это обычная нейронная сеть с новыми типами слоев, такими как сверточные слои и слои субдискретизации.  • В сверточных нейронных сетях не учитывается локальность признаков.  • Сверточный слой содержит значительно меньше весов для настройки в сравнении с аналогичным слоем полносвязной нейронной сети.  • В сверточных нейронных сетях отсутствуют функции активации.  5.2.2) Выберите верные утверждения о рекурентных нейронных сетях.  • Рекурентные нейронные сети позволяют хранить информацию с предыдущих шагов.  • Каждое следующее предсказание в рекурентной нейронной сети делается с учетом информации с предыдущего шага.  • LSTM – это разновидность архитектуры рекурентных нейронных сетей, способная хранить и использовать в процессе обучения долгосрочную информацию.  • Все утверждения из вышеперечисленных являются верными.  5.2.3) Выберите верные утверждения.  Генеративно-состязательная сеть (generative adversarial network)  • состоит из генератора, который создает новые данные, и дискриминатора, который пытается отличить эти данные от настоящих.  • делает возможным генерацию новых данных, похожих на настоящие, из случайного шума.  • позволяет не только генерировать новые данные из случайного шума в низкоразмерном пространстве, но и сжимать существующие объекты из набора данных до низкоразмерного представления.  5.2.4) Что такое placeholder в фреймворке tensorflow?  • tf.placeholder представляет из себя константу, которая не изменяется при выполнении вычислительного графа.  • tf.placeholder говорит о добавлении объекта в вычислительный граф, который впоследствии получит свое значение.  • tf.placeholder – это переменная в tensorflow со значением, заданным при инициализации.  5.2.5) Что такое keras?  • Это набор утилит для работы с изображениями в tensorflow.  • Это фреймворк, который раньше был частью tensorflow.  • Это надстройка над tensorflow, которая позволяет производить операции над графами tensorflow в более упрощенном виде. | 1) Предположим, что на вход полносвязной нейронной сети поступает цветное изображения размером 32x32 пикселей. Количество каналов для каждого изображения в данном случае равно 3. Какое количество весов будет соответстовать одному нейрону в первом скрытом слое нейронной сети без учета bias параметра?  • 1024  • 3072  • 96  2) Предположим, что на вход сверточного слоя поступает цветное изображения размером 32x32 пикселей. Количество каналов для каждого изображения в данном случае равно 3.  Мы использует фильтры с ядром свертки 5x5. Какое количество параметров (настраеваемых весов) содержится в одном фильтре без учета bias параметра?  • 1024  • 3072  • 75  • 675  3) Каким образом качество востанавливаемых данных зависит от размерности скрытого пространства в модели автоэнкодера при обучении на одной и той же выборке? Мы считаем каждую модель достаточно точной по отношению к величине ошибки востановления.  • При увеличении размерности скрытого пространства качество востанавливаемых данных увеличивается.  • При уменьшении размерности скрытого пространства качество востанавливаемых данных увеличивается.  • Качество востанавливаемых данных не зависит от размерности скрытого пространства.  4) В чем заключается основная идея архитектуры ResNet?  • Использование Batch normalization после сверточного слоя нейронной сети.  • Дублирование сверточных слоев по отношению к количеству каналов.  • Добавление дополнительного соединения, которое обходит цепочку преобразований и создает дополнительный поток данных, который складывается с основным.  5) Какая схема описывает стандартный процесс обучения нейронной сети в tensorflow?  • Цикл по эпохам → цикл по batches → минимизация функции потерь.  • Цикл по batches → цикл по эпохам → минимизация функции потерь.  • Цикл по эпохам → цикл по batches → максимизация качества на тестовых данных.  • Цикл по batches → цикл по эпохам → максимизация качества на тестовых данных. |
| Итоговое тестирование по программе |  |  | Вопрос 1. Какие из перечисленных случайных величин являются непрерывными?  1. Количество покупок в интернет-магазине  2. Время ожидания автобуса на остановке  3. Количество очков, выпавших на игральном кубике  4. Нет правильного варианта ответа  Вопрос 2.  Пусть скорость автомобиля описывается нормально распределенной случайной величиной с параметрами N(40,5). Какова вероятность того, что скорость автомобиля будет равна 60 км/ч?  1. 0.89  2. 0.34  3. 0  4. 0,001  Вопрос 3.  Корреляция позволяет оценить следующую зависимость между случайными величинами:  1. Нелинейную зависимость  2. Линейную зависимость  3. Корреляция не оценивает зависимость между случайными величинами  4. Корреляция оценивает только положительную зависимость между величинами  Вопрос 4.  Правило 3 сигм показывает, что:  1. Нормальное распределение имеет вид N(3σ,1)  2. 95 данных лежит в интервале +/−3σ  3. Нормальное распределение имеет вид N(0.3σ,1)  4. 99.7 данных лежит в интервале +/−3σ  Вопрос 5. Что из нижеперечисленных задач является задачей обучения с учителем?  1. Прогнозирование курса доллара к рублю  2. Оценка вероятности клиента интернет-магазина сделать покупку в следующем месяце  3. Поиск подозрительных транзакций по банковской карте  4. Сегментация абонентов сотовой связи по их активности  Вопрос 6.  Какая функция ошибки используется для оптимизации коэффициентов линейной регрессии градиентным методом?  1. Средняя абсолютная ошибка, MAE  2. Для линейной регрессии имеется аналитическое решение, которое всегда предпочтительнее использовать  3. Средняя квадратичная ошибка, MSE  Вопрос 7.  Предположим, что вы решаете задачу бинарной классификации об отказе оборудования не предприятии. Вы должны предсказывать 1, если оборудование откажет в течение ближайших шести месяцев и 0, если этого не случится. Вы знаете, что отказы оборудования – редкое событие. Какую метрику НЕ стоит использовать для оценки качества модели?  1. Accuracy  2. Recall  3. F-measure  4. Precision  Вопрос 8.  Сколько раз обучается модель при кросс-валидации на 5 фолдах?  1. 1  2. 4  3. 5  4. ни разу  Вопрос 9. Выберите наиболее верное из следующих увтерждений:  1. Кросс-валидация борется с переобучением  2. Кросс-валидация позволяет оценить обощающую способность модели  3. Кросс-валидация оценивает качество, используя ансамбль, построенный с помощью исходной модели.  4. Вместо обычной валидации на отложенной выборке, всегда лучше применять кросс-валидацию.  Вопрос 10.  Выберите верные утверждения.  1. Решающие деревья склонны к переобучению.  2.Результатом классификации нового объекта в решающем дереве является класс, содержащийся в листе этого дерева.  3. Решающие деревья не используются в качестве регрессионных моделей.  Вопрос 11.  Выберите верные утверждения:  В градиентном бустинге над случайными деревьями:  1. Базовые алгоритмы строятся независимо.  2. Каждый следующий алгоритм пытается исправить ошибки предыдущего.  3. Необходимо использовать глубокие деревья.  4. Модель может переобучаться при увеличении числа решающих деревьев.  Вопрос 12. Какие из пунктов описывают достоинства случайного леса?  1. Существуют методы оценки предсказательной способности признаков в модели.  2. Случайный лес не склонен к переобучению при увеличении числа базовых алгоритмов.  3. В случайном лесе обычно используются глубокие решающие деревья.  4. В этой модели решающие деревья строятся независимо, что делает возможным использование методов паралелльных вычислений.  Вопрос 13.  Какие из данных проблем могут решаться с помощью кластеризации?  1. Сегментация посетителей интернет магазина  2. Прогнозирование спроса на товары на следующей неделе  3. Выявление аномальных наблюдений  4. Выделение групп единомышленников по результатам опроса  Вопрос 14.  Пусть дано 100 объектов на которых была запущена иерархическая кластеризация. По дендрограмме мы видим, что на каждом шаге к наибольшему кластеру добавляется по одному объекту. Является ли это признаком удачной кластеризации и почему?  1. Скорее да, алгоритм же выполнился без ошибок  2. Скорее да, нужно только умело подобрать порог отсечения  3. Скорее нет, потому что добавление по одному объекту плохо показывает иерархическую структуру в данных  4. Скорее нет, возможно в данных есть ошибка или параметры алгоритма выбраны неудачно  Вопрос 15. Выберите верные утверждения  1. Чем больше значение Dunn Index, тем разбиение считается лучше  2. Чем больше значение Daves-Bouldin Index, тем разбиение считается лучше  3. Чем больше значение Silhouette Index, тем разбиение считается лучше  4. Rand Index (Adjusted Rand Index) можно вычислить, только если количество кластеров между разбиениями совпадает  Вопрос 16. Выберите верные утверждения  1. Новые признаки равны линейной комбинации исходных признаков  2. Матрица ковариаций - квадратная матрица  3. Перед применением PCA признаки надо центрировать (и шкалировать)  4. Количество компонент задается исследователем  Вопрос 17.  Выберите верные утверждения  1. Классический подход в многомерном шкалировании эквивалентен методу главных компонент  2. Если метод многомерного шкалирования работает хорошо, то точки на диаграмме стресса будут расположены близко к диагонали  3. В неметрическом многомерном шкалировании важно сохранение порядка а не значений попарных расстояний  Вопрос 18. Как называется метод вычисления градиента, который используется при обновления весов нейронной сети в процессе обучения?  1. Forward Pass.  2. Forward Propagation.  3. Backpropagation или алгоритм обратного распространения ошибки.  4. Stochastic Gradient Descent или стохастический градиентный спуск.  Вопрос 19. Выберите верные утверждения о сверточных нейронных сетях.  1. Сверточная нейронная сеть – это обычная нейронная сеть с новыми типами слоев, такими как сверточные слои и слои субдискретизации.  2. В сверточных нейронных сетях не учитывается локальность признаков.  3. Сверточный слой содержит значительно меньше весов для настройки в сравнении с аналогичным слоем полносвязной нейронной сети.  4. В сверточных нейронных сетях отсутствуют функции активации.  Вопрос 20. Выберите верные утверждения. Нейронные сети  1. Требуют большого количества данных при обучении.  2. Не переобучаются.  3. Требуют вычислительных ресурсов.  4. Всегда сходятся. |

**8.2.**  **описание показателей и критериев оценивания, шкалы оценивания**

Входная/выходная диагностика участников представляет собой измерение соответствующих программе компетенций участников перед началом и по завершении обучения:

1. Входная/выходная диагностика представляет собой тест, который направлен на выявление начального уровня подготовки и уровня подготовки после завершения программы

2. критерии и система оценки результатов входной/выходной диагностики представлены ниже.

Шкала оценивания для входной диагностики

|  |  |  |
| --- | --- | --- |
| Критерии | Шкала | Описание |
| зачтено | Обучающийся успешно выполнил входной тест: 50%-100% верных ответов на вопросы теста | Потенциальный обучающийся успешно выполнил входной тест и может быть допущен к обучению по программе |
| не зачтено | Обучающийся не выполнил успешно входной тест: 0%-49% верных ответов на вопросы теста | Потенциальный обучающийся не выполнил успешно входной тест и не может быть допущен к обучению по программе |

Шкала оценивания для промежуточного контроля по темам/модулям

|  |  |  |
| --- | --- | --- |
| Критерии | Шкала | Описание |
| зачтено | Обучающийся проходит тест для самоконтроля до успешного завершения теста: 50%-100% верных ответов на вопросы теста | Данный вид тестов не оценивается |

Шкала итогового тестирования по темам/модулям

|  |  |  |  |
| --- | --- | --- | --- |
| Критерии | Шкала | Описание | Уровень сформированности компетенции |
| зачтено | 91%-100% баллов за тестирование | Обучающийся успешно выполнил тест по теме | Профессиональный уровень |
| зачтено | 76%-90% баллов за тестирование | Обучающийся успешно выполнил тест по теме | Продвинутый уровень |
| зачтено | 61%-75% баллов за тестирование | Обучающийся успешно выполнил тест по теме | Базовый уровень |
| не зачтено | 50%-60% баллов за тестирование | Обучающийся не выполнил успешно тестирование, необходимо повторное назначение теста и консультация преподавателя | Начальный уровень |
| не зачтено | 0%-49% баллов за промежуточное тестирование и практическую работу | Обучающийся не выполнил успешно промежуточное тестирование, необходимо повторное назначение теста и консультация преподавателя | Не владеет, компетенция не сформирована |

Шкала оценивания для выходной диагностики

|  |  |  |  |
| --- | --- | --- | --- |
| Критерии | Шкала | Описание | Уровень сформированности компетенции |
| зачтено | 91%-100% верных ответов на вопросы итогового теста  Успешно выполнено не менее 80% практических заданий курса | Обучающийся успешно выполнил итоговый тест | Профессиональный уровень |
| зачтено | 76%-90% верных ответов на вопросы итогового теста | Обучающийся успешно выполнил итоговый тест | Продвинутый уровень |
| зачтено | 60%-75% верных ответов на вопросы итогового теста | Обучающийся успешно выполнил итоговый тест по теме | Базовый уровень |
| не зачтено | 50%-59% верных ответов на вопросы итогового теста | Обучающийся не выполнил успешно итоговый тест необходимо повторное назначение теста и консультация преподавателя | Начальный уровень |
| не зачтено | 0%-49% верных ответов на вопросы теста | Обучающийся не выполнил успешно итоговый тест | Не владеет, компетенция не сформирована |

Шкала оценивания практических заданий

|  |  |  |  |
| --- | --- | --- | --- |
| Критерии | Шкала | Описание | Допуск к итоговой аттестации/выходной диагностики |
| зачтено | 50%-100% верных ответов по всем практическим заданиям программы | Обучающийся успешно выполнил практические задания | Да |
| не зачтено | 0%-49% верных ответов по всем практическим заданиям программы | Обучающийся не выполнил успешно практические задания в нужном объеме | Нет  Необходимо повторное выполнение заданий и консультация преподавателя |

**8.3.**  **примеры контрольных заданий по модулям или всей образовательной программе**

Примеры контрольных заданий по программе:

Вопрос 1. Какие из перечисленных случайных величин являются непрерывными?

1. Количество покупок в интернет-магазине

2. Время ожидания автобуса на остановке

3. Количество очков, выпавших на игральном кубике

4. Нет правильного варианта ответа

Вопрос 2. Пусть скорость автомобиля описывается нормально распределенной случайной величиной с параметрами N(40,5). Какова вероятность того, что скорость автомобиля будет равна 60 км/ч?

1. 0.89

2. 0.34

3. 0

4. 0,001

Вопрос 3. Корреляция позволяет оценить следующую зависимость между случайными величинами:

1. Нелинейную зависимость

2. Линейную зависимость

3. Корреляция не оценивает зависимость между случайными величинами

4. Корреляция оценивает только положительную зависимость между величинами

Вопрос 4. Правило 3 сигм показывает, что:

1. Нормальное распределение имеет вид N(3σ,1)

2. 95 данных лежит в интервале +/−3σ

3. Нормальное распределение имеет вид N(0.3σ,1)

4. 99.7 данных лежит в интервале +/−3σ

Вопрос 5. Какие из нижеприведенных задач являются задачами обучения с учителем?

1. Прогнозирование курса доллара к рублю

2. Оценка вероятности клиента интернет-магазина сделать покупку в следующем месяце

3. Поиск подозрительных транзакций по банковской карте

4. Сегментация абонентов сотовой связи по их активности

Вопрос 6. Какая функция ошибки используется для оптимизации коэффициентов линейной регрессии градиентным методом?

1. Средняя абсолютная ошибка, MAE

2. Для линейной регрессии имеется аналитическое решение, которое всегда предпочтительнее использовать

3. Средняя квадратичная ошибка, MSE

Вопрос 7. Предположим, что вы решаете задачу бинарной классификации об отказе оборудования не предприятии. Вы должны предсказывать 1, если оборудование откажет в течение ближайших шести месяцев и 0, если этого не случится. Вы знаете, что отказы оборудования – редкое событие. Какую метрику НЕ стоит использовать для оценки качества модели?

1. Accuracy

2. Recall

3. F-measure

4. Precision

Вопрос 8. Сколько раз обучается модель при кросс-валидации на 5 фолдах?

1. 1

2. 4

3. 5

4. ни разу

Вопрос 9. Выберите наиболее верное из следующих утверждений:

1. Кросс-валидация борется с переобучением

2. Кросс-валидация позволяет оценить обобщающую способность модели

3. Кросс-валидация оценивает качество, используя ансамбль, построенный с помощью исходной модели.

4. Вместо обычной валидации на отложенной выборке, всегда лучше применять кросс-валидацию.

Вопрос 10. Выберите верные утверждения.

1. Решающие деревья склонны к переобучению.

2.Результатом классификации нового объекта в решающем дереве является класс, содержащийся в листе этого дерева.

3. Решающие деревья не используются в качестве регрессионных моделей.

Вопрос 11. Выберите верные утверждения:

В градиентном бустинге над случайными деревьями:

1. Базовые алгоритмы строятся независимо.

2. Каждый следующий алгоритм пытается исправить ошибки предыдущего.

3. Необходимо использовать глубокие деревья.

4. Модель может переобучаться при увеличении числа решающих деревьев.

Вопрос 12. Какие из пунктов описывают достоинства случайного леса?

1. Существуют методы оценки предсказательной способности признаков в модели.

2. Случайный лес не склонен к переобучению при увеличении числа базовых алгоритмов.

3. В случайном лесе обычно используются глубокие решающие деревья.

4. В этой модели решающие деревья строятся независимо, что делает возможным использование методов паралелльных вычислений.

Вопрос 13. Какие из данных проблем могут решаться с помощью кластеризации?

1. Сегментация посетителей интернет магазина

2. Прогнозирование спроса на товары на следующей неделе

3. Выявление аномальных наблюдений

4. Выделение групп единомышленников по результатам опроса

Вопрос 14. Пусть дано 100 объектов на которых была запущена иерархическая кластеризация. По дендрограмме мы видим, что на каждом шаге к наибольшему кластеру добавляется по одному объекту. Является ли это признаком удачной кластеризации и почему?

1. Скорее да, алгоритм же выполнился без ошибок

2. Скорее да, нужно только умело подобрать порог отсечения

3. Скорее нет, потому что добавление по одному объекту плохо показывает иерархическую структуру в данных

4. Скорее нет, возможно в данных есть ошибка или параметры алгоритма выбраны неудачно

Вопрос 15. Выберите верные утверждения

1. Чем больше значение Dunn Index, тем разбиение считается лучше

2. Чем больше значение Daves-Bouldin Index, тем разбиение считается лучше

3. Чем больше значение Silhouette Index, тем разбиение считается лучше

4. Rand Index (Adjusted Rand Index) можно вычислить, только если количество кластеров между разбиениями совпадает

Вопрос 16. Выберите верные утверждения

1. Новые признаки равны линейной комбинации исходных признаков

2. Матрица ковариаций - квадратная матрица

3. Перед применением PCA признаки надо центрировать (и шкалировать)

4. Количество компонент задается исследователем

Вопрос 17. Выберите верные утверждения

1. Классический подход в многомерном шкалировании эквивалентен методу главных компонент

2. Если метод многомерного шкалирования работает хорошо, то точки на диаграмме стресса будут расположены близко к диагонали

3. В неметрическом многомерном шкалировании важно сохранение порядка а не значений попарных расстояний

Вопрос 18. Как называется метод вычисления градиента, который используется при обновления весов нейронной сети в процессе обучения?

1. Forward Pass.

2. Forward Propagation.

3. Backpropagation или алгоритм обратного распространения ошибки.

4. Stochastic Gradient Descent или стохастический градиентный спуск.

Вопрос 19. Выберите верные утверждения о сверточных нейронных сетях.

1. Сверточная нейронная сеть – это обычная нейронная сеть с новыми типами слоев, такими как сверточные слои и слои субдискретизации.

2. В сверточных нейронных сетях не учитывается локальность признаков.

3. Сверточный слой содержит значительно меньше весов для настройки в сравнении с аналогичным слоем полносвязной нейронной сети.

4. В сверточных нейронных сетях отсутствуют функции активации.

Вопрос 20. Выберите верные утверждения. Нейронные сети

1. Требуют большого количества данных при обучении.

2. Не переобучаются.

3. Требуют вычислительных ресурсов.

4. Всегда сходятся.

**8.4.**  **тесты и обучающие задачи (кейсы), иные практикоориентированные формы заданий**

Примеры тестов представлены в п. 8.1.

В программе используются преимущественно задания с автоматизированной проверкой, которые являются обязательными для выполнения по каждой теме (см. п. 7).

**8.5.**  **описание процедуры оценивания результатов обучения**

К итоговой аттестации допускаются слушатели, выполнившие не менее 50% всех практических заданий и сдавшие итоговые тесты по отдельным темам (60% верных ответов по каждому тесту по теме), предусмотренные учебной программой.

В ходе итогового электронного тестирования слушателю предлагается ответить на 20 вопросов. Тестирование считается успешно завершённым, если слушателем даны верные ответы на не менее чем 60% вопросов теста.

**9.Организационно-педагогические условия реализации программы**

**9.1. Кадровое обеспечение программы**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **№**  **п/п** | **Фамилия, имя, отчество (при наличии)** | **Место основной работы и должность, ученая степень и ученое звание (при наличии)** | **Ссылки на веб-страницы с портфолио (при наличии)** | **Фото в формате jpeg** | **Отметка о полученном согласии на обработку персональных данных** |
| **1** | Пестров Никита Николаевич | Data Science Lead в Habidatum  Основатель CocoaHeads Moscow  Сооснователь мобильного приложения easy ten | <https://www.basispythoncourse.ru/#rec237674990> | Пестров Никита, Instructor | Coursera | + |
| **2** | Жебрак Александр Михайлович | CTO в Insilico Medicine | <https://www.basispythoncourse.ru/#rec237674990> | Конференция по Data Science и Big Data в Москве | + |
| 3 | Шестаков Андрей Владимирович | Руководитель группы предиктивной аналитики Mail.Ru Group | <https://www.basispythoncourse.ru/#rec237674990> | Шестаков Андрей Владимирович, Instructor | Coursera | + |
| 4 | Широкова Елена Сергеевна | Data Scientist в команде Big Data Мегафон | <https://www.basispythoncourse.ru/#rec237674990> | elena-shirokova (Elena Shirokova) · GitHub | + |

**9.2.Учебно-методическое обеспечение и информационное сопровождение**

|  |  |
| --- | --- |
| **Учебно-методические материалы** | |
| Методы, формы и технологии | Методические разработки,  материалы курса, учебная литература |
| Проблемно-развивающие методы  Исследовательский метод  Объяснительно-иллюстративный метод  Метод проблемного изложения  Частично-поисковый метод  Метод контроля и самоконтроля  Самостоятельная работа по освоению теоретического материала  Самоконтроль через ответы на вопросы по видеолекциям и выполнение практических заданий для самоконтроля  Промежуточный контроль (тесты по темам для промежуточной аттестации)  Итоговый контроль (итоговое тестирование по курсу)  Консультативная поддержка преподавателя и менторов в режиме онлайн. | Плас Дж. Вандер Python для сложных задач: наука о данных и машинное обучение. —СПб.: Питер, 2018. —576 с.  Рашка, С. Python и машинное обучение: крайне необходимое пособие по новейшей предсказательной аналитике, обязательное для более глубокого понимания методологии машинного обучения руководство / Москва : ДМК Пресс, 2017. —418 с. |

|  |  |
| --- | --- |
| **Информационное сопровождение** | |
| Электронные  образовательные ресурсы | Электронные  информационные ресурсы |
| Система автоматизированной проверки заданий на языке Python  Материалы программы на образовательной платформе | Официальная документация библиотеки scikit-learn <https://scikit-learn.org>  Официальная документация библиотеки scikit-learn <https://scikit-learn.org>  Официальная документация библиотеки NumPy <https://numpy.org/>  Официальная документация библиотеки Matplotlib <https://matplotlib.org/> |

**9.3.Материально-технические условия реализации программы**

|  |  |
| --- | --- |
| Вид занятий | Наименование оборудования,  программного обеспечения |
| Теоретический компонент | Операционная система семейств Windows, Mac Os, Linux.  Браузер:  • Firefox 64  • Chrome 61  • Safari 11  • Opera 57  Рекомендуется от 128 кбит/сек исходящего потока.  Рекомендуется от 256 кбит/сек входящего потока. |
| Дополнительные теоретические материалы | Операционная система семейств Windows, Mac Os, Linux.  Браузер:  • Firefox 64  • Chrome 61  • Safari 11  • Opera 57  Рекомендуется от 128 кбит/сек исходящего потока.  Рекомендуется от 256 кбит/сек входящего потока. |
| Аттестационный компонент | Операционная система семейств Windows, Mac Os, Linux.  Браузер:  • Firefox 64  • Chrome 61  • Safari 11  • Opera 57  Рекомендуется от 128 кбит/сек исходящего потока.  Рекомендуется от 256 кбит/сек входящего потока. |
| Практический компонент | Операционная система семейств Windows, Mac Os, Linux.  Браузер:  • Firefox 64  • Chrome 61  • Safari 11  • Opera 57  Рекомендуется от 128 кбит/сек исходящего потока.  Рекомендуется от 256 кбит/сек входящего потока.  Установленный пакет программ (распространяемый бесплатно) Anaconda (https://www.anaconda.com/) |

**III.Паспорт компетенций (Приложение 2)**

Автономная некоммерческая организация дополнительного профессионального образования

«СофтЛайн Эдюкейшн»

ПАСПОРТ КОМПЕТЕНЦИИ

**Дополнительная профессиональная программа повышения квалификации**

**«Python для анализа данных»**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1. | Наименование компетенции | | Анализ больших данных с использованием существующей в организации методологической и технологической инфраструктуры | |
| 2. | Указание типа компетенции | общекультурная/  универсальная |  | |
| общепрофессиональная |  | |
| профессиональная |  | |
| профессионально-специализированная | Да | |
| 3. | Определение, содержание и основные сущностные характеристики компетенции | | Под компетенцией понимается способность постановки и нахождения путей решения прикладных задач по созданию и совершенствованию программного кода с использованием специализированных программных средств на примере инструментов экосистемы Python  знать:  - синтаксис основных библиотек для анализа данных – NumPy, SciPy, Pandas, Sklearn и построения нейронных сетей – Keras, TensorFlow:  -синтаксис основных библиотек для визуализации данных – Matplotlib, Plotly;  - особенности работы в среде Jupyter notebook и Colab;  - основные алгоритмы обучения на размеченных данных;  - основные алгоритмы обучения на неразмеченных данных;  - основные архитектуры нейросетей;  уметь:  - использовать библиотеки Python для решения задач регрессии, классификации, понижения размерности  - использовать библиотеки Keras и TensorFlow для построения нейросети  - использовать библиотеки Matplotlib и Plotly для визуализации данных.  владеть:  - навыками создания программного кода на Python, который визуализирует данные;  - навыками создания программного кода на Python, который решает задачи регрессии и классификации  - навыками создания кода, который решает задачу построения нейросети с использованием фреймворка Keras. | |
| 4. | Дескриптор знаний, умений и навыков по уровням | | Уровни сформированности компетенции  обучающегося | Индикаторы |
|  | | Начальный уровень  (Компетенция недостаточно развита. Частично проявляет навыки, входящие в состав компетенции. Пытается, стремится проявлять нужные навыки, понимает их необходимость, но у него не всегда получается.) | Знает: основные методы библиотек для визуализации и анализа данных, основные строительные блоки нейросетей  Умеет: применять Python для вызова библиотек с параметрами по умолчанию  Владеет: навыками разработки и запуска кода в Jupyter notebook |
|  | | Базовый уровень  (Уверенно владеет навыками, способен, проявлять соответствующие навыки в ситуациях с элементами неопределённости, сложности.) | Знает: не только основные методы библиотек для визуализации и анализа данных, но и основные параметры их вызова: основные строительные блоки нейросетей  Умеет: настраивать параметры вызова библиотек; строить простейшие нейросети.  Владеет: эффективными навыками разработки и запуска кода в Jupyter notebook |
|  | | Продвинутый  (Владеет сложными навыками, способен активно влиять на происходящее, проявлять соответствующие навыки в ситуациях повышенной сложности.) | Знает: методы библиотек для визуализации и анализа данных; параметры их вызова; основные строительные блоки нейросетей,  Умеет: настраивать параметры вызова библиотек и выбирать оптимальные параметры; строить стандартные нейросети;  Владеет: эффективными навыками разработки и запуска кода в Jupyter notebook |
|  | | Профессиональный  (Владеет сложными навыками, создает новые решения для сложных проблем со многими взаимодействующими факторами, предлагает новые идеи и процессы, способен активно влиять на происходящее, проявлять соответствующие навыки  в ситуациях повышенной сложности.) | Знает: продвинутые методы библиотек для визуализации и анализа данных; тонкие параметры их вызова; основные строительные блоки нейросетей,  Умеет: настраивать параметры вызова библиотек и выбирать оптимальные параметры; строить продвинутые нейросети;  Владеет: эффективными навыками разработки и запуска кода в Jupyter notebook  Умеет реализовывать код из статей, отражающих последние теоретические достижения в области анализа данных.  В состоянии предложить новые идеи решения сложных задач анализа данных. Принимает профессиональные решения в условиях неполной определенности, при недостаточном документальном, нормативном и методическом обеспечении. |
| 5. | Характеристика взаимосвязи данной компетенции с другими компетенциями/ необходимость владения другими компетенциями для формирования данной компетенции | | Для формирования данной компетенции слушатель должен владеть:  • компетенциями цифровой грамотности, в том числе навыками владения ПК на уровне среднего или продвинутого пользователя;  • компетенции в области алгоритмизации, в том числе знаниями и умениями в области алгоритмизации (знание основных алгоритмических структур, умения выражать их на псевдокоде и в виде блок-схем) и программирования, базовыми знаниями в области языка Python | |
| 6. | Средства и технологии оценки | | Входное, промежуточное, итоговое тестирование по темам/модулям, выходная диагностика | |

**VI.Иная информация о качестве и востребованности образовательной программы** (результаты профессионально-общественной аккредитации образовательной программы, включение в системы рейтингования, призовые места по результатам проведения конкурсов образовательных программ и др.) (при наличии)

Программа в асинхронном режиме обучения прошла апробацию в компании "Mail.ru Group" — крупнейшем холдинге в России по дневной аудитории (Mediascope Web Index Desktop+Mobile, Россия 0+, 12–64 года, сентябрь 2019 г.). Обучение по программе проходили IT-специалисты компании. Рекомендательное письмо может быть предоставлено по запросу.

**V.Рекомендаций к программе от работодателей**: наличие не менее двух писем и/или подтверждения на цифровой платформе Государственной системы предоставления ПЦС от работодателей о рекомендации образовательной программы для реализации в рамках Государственной системы предоставления ПЦС на формирование у трудоспособного населения компетенций цифровой экономики с указанием востребованности результатов освоения программы в сфере деятельности соответствующих компаний и готовности к рассмотрению заявок наиболее успешно освоивших образовательную программу граждан на прохождение стажировки и (или) собеседования на предмет трудоустройства путем проставления отметки в профиле программы

К паспорту прилагаются рекомендации от работодателей в формате .pdf

**VI.Указание на возможные сценарии профессиональной траектории граждан** по итогам освоения образовательной программы (в соответствии с приложением)

|  |  |
| --- | --- |
| **Цели получения персонального цифрового сертификата** | |
| **текущий статус** | **цель** |
| **Развитие компетенций в текущей сфере занятости** | |
| работающий по найму в организации, на предприятии | сохранение текущего рабочего места |
| работающий по найму в организации, на предприятии | развитие профессиональных качеств |
| работающий по найму в организации, на предприятии | повышение заработной платы |
| работающий по найму в организации, на предприятии | смена работы без изменения сферы профессиональной деятельности |

**VII.Дополнительная информация**

нет

**VIII.Приложенные Скан-копии**

1. Рекомендательные письма

2. Скан утвержденной образовательной программы

3. Скан утвержденного паспорта образовательной программы

4. Паспорт образовательной программы в формате word

**Генеральный директор**

**АНО ДПО «СофтЛайн Эдюкейшн» \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ В.Э. Разуваев**

М.П.