**Правительство Российской Федерации**

**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Введение в машинное обучение

Introduction to Machine Learning

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 4

Регистрационный номер рабочей программы: 027261

Санкт-Петербург

2020

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Целью курса «Введение в машинное обучение» является введение в обширную область прикладной математики, занимающейся моделированием и оптимизацией в условиях неполных и неточных данных. Одними из наиболее важных аспектов предлагаемого курса и его отличительной особенностью является широта охвата и практическая направленность. Большой охват достигаются за счет меньшей подробности изложения: многие утверждения приводится без строгих доказательств. Практическая направленность выражается как в том, что рассматриваются работающие методы, доказавшие свою состоятельность в реальных условиях, так и в организации практических занятий вокруг экспериментов с реальными данными.

Еще одной отличительной особенностью и целью курса является составление у слушателя целостной картины актуальных исследований в области машинного обучения. Поэтому в курсе рассматриваются не только ставшие классическими, но и перспективные направления развития машинного обучения. Материалы для этой части курса взяты и обновляются по результатам ведущих мировых конференций (ICML, CIKM, NIPS, и т.д.).

**1.2. Требования подготовленности обучающегося к освоению содержания учебных занятий (пререквизиты)**

Программа курса предназначена обучающимся 1 курса магистратуры и рассчитана на обучающихся, изучавших:

• общий курс математического анализа;

• линейную алгебру;

• методы вычислений;

• углубленное программирование;

• теорию вероятностей;

• математическую статистику.

Обязательным условием к обучающимся курса является способность к самостоятельному программированию и доступ к соответствующей технике. Для чтения рекомендованной литературы нужен английский язык.

**1.2.1. Требуемые компетенции**

* ПКА-1 – способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий;
* ПКП-7 – способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений;
* ПКП-8 – способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков.

**1.3. Перечень результатов обучения (learning outcomes)**

**1.3.1. Компетенции, развиваемые дисциплиной**

Курс способствует формированию следующих компетенций:

* ОПК-1 – способен находить, формулировать и решать актуальные проблемы фундаментальной и прикладной информатики и информационных технологий;
* ПКА-1 – способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий;
* ПКП-1 – способность проводить научные исследования на основе существующих методов в конкретной области профессиональной деятельности;
* ПКП-7 – способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений;
* ПКП-8 – способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков.

**1.3.2. Знания, умения и навыки, формируемые дисциплиной**

После прослушивания курса, предполагается, что обучающийся получит следующие знания и умения:

1. понимание задач машинного обучения, отличия предмета от смежных областей;
2. навыки по моделированию практических задач и перевод их на язык оптимизации;
3. способность оценить потенциал использования того или иного подхода в условиях поставленной задачи и предоставленных данных;
4. навыки практической реализации решений на одном из языков программирования (Octave/Python/Java);
5. навыки анализа полученных результатов и возможных путей их улучшения;
6. ориентирование в актуальных проблемах машинного достаточное для формулирования собственных исследовательских задач.

**1.4. Перечень и объём активных и интерактивных форм учебных занятий**

Интерактивная форма учебных занятий (20 часов в течение семестра) заключается в обсуждении в аудитории вопросов по пройденному материалу, домашним заданиям и самостоятельно изученным научным статьям.

**Раздел 2. Организация, структура и содержание учебных занятий**

**2.1. Организация учебных занятий**

**2.1.1 Основной курс**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Трудоёмкость, объёмы учебной работы и наполняемость групп обучающихся | | | | | | | | | | | | | | | | | | |
| Код модуля в составе дисциплины,  практики и т.п. | Контактная работа обучающихся с преподавателем | | | | | | | | | | | | Самостоятельная работа | | | | Объём активных и интерактивных  форм учебных занятий | Трудоёмкость |
| лекции | семинары | консультации | практические  занятия | лабораторные работы | контрольные работы | коллоквиумы | текущий контроль | промежуточная  аттестация | итоговая аттестация | под руководством преподавателя | в присутствии  преподавателя | сам. раб. с использованием  методических материалов | текущий контроль (сам.раб.) | промежуточная аттестация (сам.раб.) | итоговая аттестация  (сам.раб.) |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | | | | | | | | | | | | | |
| Форма обучения: очная | | | | | | | | | | | | | | | | | | |
| Семестр 1 | 15 | 15 | 2 |  |  |  |  |  | 2 |  |  |  | 74 |  | 36 |  | 20 | 4 |
|  | 1-30 | 10-25 | 2-100 |  |  |  |  |  | 2-100 |  |  |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 15 | 15 | 2 |  |  |  |  |  | 2 |  |  |  | 74 |  | 36 |  | 20 | 4 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Виды, формы и сроки текущего контроля успеваемости и промежуточной аттестации | | | | | | |
| Код модуля в составе дисциплины, практики и т.п. | Формы текущего контроля успеваемости | | Виды промежуточной аттестации | | Виды итоговой аттестации  (только для программ итоговой аттестации и дополнительных образовательных программ) | |
| Формы | Сроки | Виды | Сроки | Виды | Сроки |
| ОСНОВНАЯ ТРАЕКТОРИЯ | | | | | | |
| Форма обучения: очная | | | | | | |
| Семестр 1 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

Период обучения (модуль): семестр 1.

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование темы (раздела, части) | Вид учебных занятий | Количество часов |
| I. | Введение | лекции | 1 |
| семинары | 1 |
| II. | Линейные модели | лекции | 2 |
| семинары | 2 |
| по методическим материалам | 16 |
| III. | Методы уменьшения размерности | лекции | 2 |
| семинары | 2 |
| по методическим материалам | 10 |
| IV. | Метод опорных векторов | лекции | 2 |
| семинары | 2 |
| по методическим материалам | 8 |
| V. | Кластеризация | лекции | 1 |
| семинары | 1 |
| по методическим материалам | 8 |
| VI. | Вероятностные методы | лекции | 1 |
| семинары | 1 |
| по методическим материалам | 8 |
| VII. | Нейронные сети | лекции | 3 |
| семинары | 3 |
| по методическим материалам | 10 |
| VIII. | Деревья принятия решения | лекции | 1 |
| семинары | 1 |
| по методическим материалам | 4 |
| IX. | Работа с текстами на естественном языке | лекции | 1 |
| семинары | 1 |
| по методическим материалам | 6 |
| X. | Приватность в анализе данных и машинном обучении | лекции | 1 |
| семинары | 1 |
| по методическим материалам | 4 |
| XI. | Промежуточная аттестация | консультация | 2 |
| экзамен | 36 |
| Итого | | | 144 |

**2.2.1. Содержание учебных занятий**

Модуль 1 (Введение)

1. Основные задачи машинного обучения и его роль в индустрии.
2. Особенности применение машинного обучения для достижения краткосрочных и долгосрочных целей.
3. Концепции Artificial intelligence, Augmented Intelligence и Artificial General Intelligence.

Модуль 2 (Линейные модели)

1. Задача линейное регрессии и метод градиентного спуска.
2. Особенности реализации градиентного спуска: полный, стохастический, пакетный.
3. Задача логистической регрессии.
4. О роли регуляризация в машинном обучении на примере линейных моделей.
5. Работа с признаками для линейных моделей: нормировка, взаимодействия, производные признаки.
6. Интерпретация линейных моделей.

Модуль 3 (Методы уменьшения размерности)

1. Метод главных компонент.
2. Метод сингулярного разложения.
3. Метод локально-чувствительных хэшей.
4. Факторизационные машины.

Модуль 4 (Метод опорных векторов)

1. Метод опорных векторов, случай hard margin.
2. Метод опорных векторов, случай soft margin.
3. Прямая и двойственная постановка задачи SVM.
4. Использование ядер в методе опорных векторов.

Модуль 5 (Кластеризация)

1. Кластеризация больших банных с помощью K-Means.
2. Кластеризация на базе плотности DBSCAN.
3. Кластеризация графов Affinity Propagation.

Модуль 6 (Вероятностные методы)

1. Метод наивного Байесовского классификатора.
2. Сэмплирование Томпсона.
3. Решение задачи оптимизации с помощью Гауссовых процессов.

Модуль 7 (Нейронные сети)

1. Общие понятия искусственных нейронных сетей.
2. Сверточные сети для анализа изображений.
3. Рекурентные сети для анализа последовательностей.

Модуль 8 (Деревья принятия решения)

1. Структура и алгоритмы построения решающих деревьев.
2. Ансамбли решающих деревьев: случайный лес и градиентный бустинг.

Модуль 9 (Работа с текстами на естественном языке)

1. Основные задачи NLP.
2. Предобработка текстов (токенизация, стеминг, лематизация TF-IDF).
3. Тематическое моделирование PLSA/LDA.
4. Нейронные сети при работе с текстами.

Модуль 10 (Приватность в анализе данных и машинном обучении)

1. Разные подходы к обеспечению приватности: k-anonymity и differential privacy.
2. Примеры реализации DP-моделей в Google, Apple и Microsoft.

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Методические материалы включают в себя следующие типы материалов — литература, Интернет-ресурсы, учебные пособия, с опорой на которые проводится аудиторная работа.

**3.1.2 Методическое обеспечение самостоятельной работы**

Самостоятельная работа обучающегося, как вид деятельности, стимулирующий активность, самостоятельность, познавательный интерес с целью поиска необходимой информации, приобретения знаний, использования этих знаний для решения учебных, научных и профессиональных задач, представляет собой важную составляющую учебного процесса, которой отводится не менее половины учебного времени при очной форме обучения. Время, отводимое на самостоятельную работу, должно использоваться обучающимися для наиболее полного освоения учебной дисциплины. Следовательно, организация эффективной внеаудиторной самостоятельной работы в процессе обучения требует создания условий, призванных обеспечить рациональное и планомерное управление учебной деятельностью, протекающей в отсутствие преподавателя.

К числу методических пособий относятся:

- Задания для самостоятельного выполнения;

- Литература по теме курса;

- Сайт поддержки курса в сети интернет, на котором слушатели курса публикуют свои самостоятельные работы и имеют возможность задавать вопросы преподавателю и друг другу.

Роль преподавателя в организации самостоятельной работы состоит в координации действий обучающихся в освоении дисциплины, в методическом и организационном обеспечении учебного процесса. Взаимодействие между преподавателем и обучающимся осуществляется в форме консультаций, а также через сайт поддержки курса. Преподаватели также оказывают помощь обучающимся по планированию и организации самостоятельной работы.

**3.1.3 Методика проведения текущего контроля успеваемости и промежуточной аттестации и критерии оценивания**

**3.1.3.1. Методика проведения текущего контроля успеваемости и промежуточной аттестации**

Текущая успеваемость контролируется в рамках решения практических задач по блокам курса (4 задачи на семестр). Обучающиеся, успешно сдавшие решения всех 4-х задач, допускаются к экзамену, который проводится в формате устного собеседования. На собеседовании необходимо продемонстрировать свои знания как минимум в 3 различных модулях.

**3.1.3.2. Критерии оценивания итогового процента освоения дисциплины**

Практические задачи оцениваются по шкале от 0 (решение отсутствует или обладает существенными недостатками) до 1 (решение существенными недостатками не обладает). Результирующая оценка находится в диапазоне от 0 до 100 и вычисляется по формуле MAX(0, (n/N – 0.6)) \* 2.5 \* 100, где n – суммарный балл, набранный обучающимся на момент аттестации, N – максимально возможный суммарный балл за курс. Например, обучающийся, успешно сдавший 60% заданий, получает 0 баллов, 80% – 50, 100% – 100 баллов.

На теоретическом экзамене ответ на каждый вопрос про конкретный модуль и на дополнительные вопросы оценивается по шкале от 0 (нет ответа) до 10 (очень хороший ответ), далее оценка усредняется. Результат переводится в диапазон от 0 до 100.

Итоговый процент выполнения целей изучения дисциплины вычисляется как минимум из оценки за практические задачи и оценки за теоретический экзамен. Далее применяется следующее правило выставления оценки:

|  |  |  |
| --- | --- | --- |
| Итоговый процент  выполнения, % | Оценка СПбГУ при  проведении экзамена | Оценка  ECTS |
| 90-100 | отлично | A |
| 80-89 | хорошо | B |
| 70-79 | хорошо | C |
| 61-69 | удовлетворительно | D |
| 50-60 | удовлетворительно | E |
| менее 50 | неудовлетворительно | F |

**3.1.4 Методические материалы для проведения текущего контроля успеваемости и промежуточной аттестации (контрольно-измерительные материалы, оценочные средства)**

**3.1.4.1. Формируемые дисциплиной компетенции**

**ОПК-1** – способен находить, формулировать и решать актуальные проблемы фундаментальной и прикладной информатики и информационных технологий.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКА-1** – способен демонстрировать фундаментальные знания математических и естественных наук, программирования и информационных технологий.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКП-1** – способность проводить научные исследования на основе существующих методов в конкретной области профессиональной деятельности.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКП-7** – способен использовать основные концептуальные положения функционального, логического, объектно-ориентированного и визуального направлений программирования, методы, способы и средства разработки программ в рамках этих направлений.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**ПКП-8** – способен использовать современные методы разработки и реализации конкретных алгоритмов математических моделей на базе языков.

□ Формируется дисциплиной.

✓ Развивается дисциплиной.

□ Полностью сформирована по результатам освоения дисциплины.

Шкала оценивания: линейная, определяется долей успешно выполненных заданий, проверяющих данную компетенцию.

**3.1.4.2. Контрольно-измерительные материалы (примеры)**

***Пример списка вопросов для устного экзамена:***

1. Основные задачи машинного обучения и его роль в индустрии
2. Линейные модели
3. Методы уменьшения размерности
4. Факторизационные машины
5. Метод опорных векторов
6. Методы кластеризации
7. Вероятностные методы анализа
8. Искусственные нейронные сети
9. Деревья принятия решения и их ансамбли
10. Работа с текстами на естественном языке

***Проверяемые компетенции*:** УКБ-1, УКБ-3

***Критерии оценивания*:** обучающемуся даётся два билета и задаётся несколько дополнительных вопросов по курсу. Ответ на каждый вопрос билета и на дополнительные вопросы оценивается по шкале от 0 (нет ответа) до 10 (очень хороший ответ), далее оценка усредняется. Результат переводится в диапазон от 0 до 100.

***Примеры практических задач:***

**Задача 1: реализация градиентного спуска**

Для выданных входных данных решить задачу линейной регрессии. Задача предполагает реализацию градиентного спуска и подсчёта метрик оценки качества модели.

Задача выполняется каждым обучающимся самостоятельно, срок сдачи – 2 недели, результатом является программный код и таблица с подсчитанными метриками оценки качества модели.

***Проверяемые компетенции*:** ПКА-1, ПКП-7, ПКП-8

***Критерии оценивания*:** по шкале от 0 (решение отсутствует или обладает существенными недостатками) до 1 (решение существенными недостатками не обладает).

**Задача 2: реализация факторизационной машины**

Для выданных входных данных решить задачу предсказания оценки пользователя. Задача предполагает реализацию факторизационной машины 2-го порядка с квадратичной функцией потерь и подсчёта метрик оценки качества модели.

Задача выполняется каждым обучающимся самостоятельно, срок сдачи – 2 недели, результатом является программный код и таблица с подсчитанными метриками оценки качества модели.

***Проверяемые компетенции*:** ПКА-1, ПКП-7, ПКП-8

***Критерии оценивания*:** по шкале от 0 (решение отсутствует или обладает существенными недостатками) до 1 (решение существенными недостатками не обладает).

**Задача 3: реализация метода кластеризации Affinity Propagation**

Для выданных входных данных решить задачу кластеризации. Задача предполагает реализацию метода Affinity Propagation и подсчёта метрик оценки качества модели.

Задача выполняется каждым обучающимся самостоятельно, срок сдачи – 2 недели, результатом является программный код и таблица с подсчитанными метриками оценки качества модели.

***Проверяемые компетенции*:** ПКА-1, ПКП-7, ПКП-8

***Критерии оценивания*:** по шкале от 0 (решение отсутствует или обладает существенными недостатками) до 1 (решение существенными недостатками не обладает).

**Задача 4: создание фреймворка для обучения нейронных сетей**

Написать свой небольшой фреймворк для обучения нейронных сетей. Применить написанный фреймворк для задачи классификации.

Задача выполняется каждым обучающимся самостоятельно, срок сдачи – 2 недели, результатом является программный код и таблица с подсчитанными метриками оценки качества модели.

***Проверяемые компетенции*:** ПКА-1, ПКП-7, ПКП-8

***Критерии оценивания*:** по шкале от 0 (решение отсутствует или обладает существенными недостатками) до 1 (решение существенными недостатками не обладает).

**3.1.5 Методические материалы для оценки обучающимися содержания и качества учебного процесса**

Для оценки обучающимися содержания и качества учебного процесса применяется анкетирование в соответствии с методикой и графиком, утвержденными в установленном порядке.

**3.2. Кадровое обеспечение**

**3.2.1 Образование и (или) квалификация штатных преподавателей и иных лиц, допущенных к проведению учебных занятий**

К проведению лекционных занятий должны привлекаться преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

**3.2.2 Обеспечение учебно-вспомогательным и (или) иным персоналом**

Учебно-вспомогательный и инженерно-технический персонал должен иметь соответствующее образование и обладать навыками организации работы с пользовательскими программными продуктами в локальной сети компьютерного класса и в Интернете.

**3.3. Материально-техническое обеспечение**

**3.3.1 Характеристики аудиторий (помещений, мест) для проведения занятий**

Требуется стандартно оборудованная аудитория с проектором.

**3.3.2 Характеристики аудиторного оборудования, в том числе неспециализированного компьютерного оборудования и программного обеспечения общего пользования**

Доска для письма маркером, мультимедийный проектор.

**3.3.3 Характеристики специализированного оборудования**

Не требуется.

**3.3.4 Характеристики специализированного программного обеспечения**

Не требуется.

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Маркеры для доски, губка

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning. Springer, 2009. – ЭР по подписке СПбГУ: <https://proxy.library.spbu.ru:2096/book/10.1007%2F978-0-387-84858-7>.
2. Нестеров Ю.Е. «Методы выпуклой оптимизации», М.: Издательство МЦНМО, 2010. – ЭР открытого доступа в сети Интернет: <https://docplayer.ru/409076-Metody-vypukloy-optimizacii.html>.

**3.4.2 Список дополнительной литературы**

Нет.

**3.4.3 Перечень иных информационных источников**

Ресурсы сети Интернет.

**Раздел 4. Разработчики программы**

Азимов Рустам Шухратуллович, магистр, ассистент кафедры информационно-аналитических систем, [st013567@student.spbu.ru](mailto:st013567@student.spbu.ru).