**Санкт-Петербургский государственный университет**

**Р А Б О Ч А Я П Р О Г Р А М М А**

**УЧЕБНОЙ ДИСЦИПЛИНЫ**

Статистическое и машинное обучение

Statistical and machine learning

**Язык(и) обучения**

русский

Трудоемкость в зачетных единицах: 3

Регистрационный номер рабочей программы: 064823

2020

**Раздел 1. Характеристики учебных занятий**

**1.1. Цели и задачи учебных занятий**

Целью данной дисциплины является обучение восприятию и презентации известных научных методов и результатов в области машинного обучения.

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

Дисциплина изучается учащимися магистратуры по направлению подготовки 01.04.02 «Прикладная математика и информатика» в третьем семестре обучения.

Обучающиеся должны иметь навыки научно-исследовательской работы.

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

В результате освоения дисциплины обучающиеся должны получить умение разобраться в результатах чужих научных исследований на основе научной литературы и довести их в доступной форме на презентации. Также, должны освоить основные методы машинного обучения.

Дисциплина участвует в формировании компетенций обучающихся по образовательной программе, установленных учебным планом для данной дисциплины:

ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики,

ОПК-2 Способен совершенствовать и реализовывать новые математические методы решения прикладных задач,

ОПК-3 Способен разрабатывать математические модели и проводить их анализ при решении задач в области профессиональной деятельности,

ПКП-10 Способен развивать и обосновывать методы и алгоритмы статистического моделирования и анализа данных сложной структуры с учетом особенностей решаемой задачи.

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

Курс подразумевает интерактивное обсуждение со время занятий, опросы на семинарских занятиях (14 часов).

Ответы на вопросы на консультации (2 часа).

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

**2.1. Организация учебных занятий**

**2.1.1 профиль Статистическое моделирование**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 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 3 | 18 | 14 | 2 |  |  |  |  |  | 2 |  | 15 |  | 34 |  | 23 |  | 16 | 3 |
|  | 1-25 | 1-25 | 1-25 |  |  |  |  |  | 1-25 |  | 1-25 |  | 1-1 |  | 1-1 |  |  |  |
| ИТОГО | 18 | 14 | 2 |  |  |  |  |  | 2 |  | 15 |  | 34 |  | 23 |  | 16 | 3 |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 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 3 |  |  | экзамен, устно, традиционная форма | по графику промежуточной аттестации |  |  |

**2.2. Структура и содержание учебных занятий**

**Основной курс**

Период обучения (модуль): **Семестр 3**

|  |  |  |  |
| --- | --- | --- | --- |
| № п/п | Наименование темы (раздела, части) | Вид учебных занятий | Количество часов |
| 1 | Общие принципы машинного обучения | Лекции | 6 |
| семинары | 2 |
| под руководством преподавателя | 3 |
| по методическим материалам | 4 |
| 2 | Обучение с учителем. Регрессия | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |
| 3 | Обучение с учителем. Классификация | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |
| 4 | Обучение без учителя. Кластерный анализ | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |
| 5 | Решающие деревья. Случайный лес | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |
| 6 | Бустинг. | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |
| 7 | Нейронные сети | Лекции | 2 |
| Семинары | 2 |
| под руководством преподавателя | 2 |
| по методическим материалам | 5 |

**Более подробное содержание.**

На каждую тему даются задания обучающимся сделать доклад по следующей схеме.

1. Какая практическая задача решается. Пример данных, на основе которых предполагается ее решать.
2. Если обучение без учителя, то используется базовая модель данных. В этом случае, скорее всего, будет максимизироваться функция правдоподобия. Если с учителем, то используется модель (алгоритм) предсказания и мера для ошибки предсказания, которая будет минимизироваться.
3. Дальше, теоретически, просто оптимизационная задача и обсуждение метода ее решения. Например, в случае без учителя это м.б. EM-алгоритм. В случае с учителем – метод стохастического градиента. При этом, если исходно в задаче были условия, при сведении задачи к безусловной оптимизации используются теорема Лагранжа или теорема Куна-Такера.
4. Обсуждение свойств метода оптимизации. Улучшение алгоритма за счет специфики задачи, эвристических приемов.
5. Возможная регуляризация, которую можно рассматривать как просто изменение оптимизационной задачи в той же модели с целью получать оценки параметров с лучшими свойствами; в частности, для получения нулевых оценок в случае добавления модуля параметра.
6. Изменение (усложнение или упрощение) рассматриваемой модели данных или предсказывающего алгоритма и переход к пункту 3.
7. Примеры (могут перемежаться с теорией).

**Раздел 3. Обеспечение учебных занятий**

**3.1. Методическое обеспечение**

**3.1.1 Методические указания по освоению дисциплины**

Успешное освоение дисциплины возможно благодаря посещению лекций, участию в обсуждениях докладов на занятиях, ознакомлении с хорошо написанными статьями и применению полученных знаний для анализа реальных данных.

**3.1.2 Методическое обеспечение самостоятельной работы**

Список литературы по изучаемым темам.

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

В процессе занятий, обучающийся должен ознакомиться с литературой по данным ему темам, сделать доклад со слайдами, продемонстрировать примеры применения методов на языке R или Python. После доклада необходимо сделать конспект с учетом замечаний, поступивших во время доклада. Допустимо объединение в группы по каждой теме с распределением ролей (доклад, конспект, демонстрация на R или Python).

Аппарат контроля за усвоением материала включает в себя экзамен по итогам курса.

Экзамен проводится в устной форме.

Для подготовки к экзамену дается три дня, каждый применяет несколько методов машинного обучения к реальным данным и делает отчет, включающий сравнение методов между собой и их теоретическое обоснование.

На экзамене оценка ставится на основе сделанных слайдов, доклада, конспекта, а также ответа на вопросы билета, где обучающийся должен продемонстрировать умение применить методы к реальным данным, с объяснениями их работы.

Наличие слайдов, доклада, конспекта дает базовую оценку «удовлетворительно». Успешное применение заданного метода повышает оценку до «хорошо». Умение объяснить теоретические и практические аспекты метода повышает оценку до «отлично».

A (отлично) – Все (доклад, конспект) выполнены полностью и в срок; при ответе на экзамене объяснены теоретические и практические аспекты методов.

B (хорошо) – Доклад сделан в срок, конспект подготовлен полностью; при ответе на экзамене объяснены теоретические и практические аспекты методов.

C (хорошо) – Доклад сделан в срок, конспект подготовлен полностью; при ответе на экзамене объяснены теоретические и практические аспекты методов, но при объяснении теоретических аспектов возникла сложности, которые были преодолены в процессе выступления.

D (удовлетворительно) – Доклад сделан в срок, конспект подготовлен полностью; при ответе на экзамене объяснены практические аспекты методов, но выявилось недостаточное понимание теоретических аспектов, которые не было преодолено в процессе выступления.

E (удовлетворительно) – Доклад сделан в срок, конспект подготовлен полностью; при ответе на экзамене выявилось недостаточное понимание теоретических и практических аспектов методов, которые не было преодолено в процессе выступления.

F (неудовлетворительно) – Либо доклад не был сделан, либо конспект не подготовлен.

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

**Примеры тем докладов**

1. Обучение с учителем. Регрессия. Регуляризация в регрессии – зачем, почему. Разные подходы.
2. Обучение с учителем. Дискриминантный анализ. Логистическая регрессия. Feature selection и extraction.
3. Обучение с учителем. Метод опорных векторов. Выбор модели с помощью кросс-валидации.
4. Вычислительные аспекты оптимизации. Гладкие функционалы и пр. Метод стохастического градиента как метод оптимизации. Примеры на основе одного из предыдущих методов.
5. Решающие деревья. Random Forest.
6. Композиция методов. Бустинг.
7. Нейронные сети. Общая структура (особый класс функций для оптимизации). Back propagation как вычислительный подход.
8. Нейронные сети для изображений.
9. Обучение без учителя. Разделение смеси распределений. Кластеризация.
10. Активное обучение.
11. Тематическое моделирование (это обучение без учителя).

Требования к презентации и докладу выкладываются в интернет на странице курса.

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

Для оценки содержания и качества учебного процесса может применяться анкетирование или опрос в соответствии с методикой и графиком, утверждаемым 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 преподаванию дисциплины могут быть допущены преподаватели, имеющие диплом о высшем образовании по соответствующему направлению.

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

Не предусмотрено.

**3.3. Материально-техническое обеспечение**

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

Учебные аудитории для проведения учебных занятий, оснащенные стандартным оборудованием, используемым для обучения в СПбГУ в соответствии с требованиями материально-технического обеспечения.

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

Стандартное оборудование, используемое для обучения в СПбГУ. MS Windows, MS Office, Mozilla FireFox, Google Chrome, Acrobat Reader DC, WinZip, Антивирус Касперского.

**3.3.3 Характеристики специализированного оборудования**

Не предусмотрено.

**3.3.4 Характеристики специализированного программного обеспечения**

R, RStudio, Python

**3.3.5 Перечень и объёмы требуемых расходных материалов**

Белая бумага формата А4 для печати на принтере из расчета 10 листов на человека.

**3.4. Информационное обеспечение**

**3.4.1 Список обязательной литературы**

1. Балдин, Евгений Михайлович. Компьютерная типография LATEX (+ CD) - СПб : БХВ-Петербург, 2008.

2. А. Шипунов, Е. Балдин, П. Волкова, А. Коробейников, С. Назарова, С. Петров, В. Суфиянов. Наглядная статистика. Используем R!. ДМК-Пресс, 2012. 298 стр. - ЭБС «Лань» по подписке СПбГУ: <https://proxy.library.spbu.ru/login?url=https://e.lanbook.com/>

3. M. Sugiyama Introduction to Statistical Machine Learning. Springer, 2016 <https://proxy.library.spbu.ru:2068/book/9780128021217/introduction-to-statistical-machine-learning> - ЭР по подписке СПбГУ

**3.4.2 Список дополнительной литературы**

Не предусмотрен.

**3.4.3 Перечень иных информационных источников**

* Сайт Научной библиотеки им. М. Горького СПбГУ:
* <http://www.library.spbu.ru/>
* Электронный каталог Научной библиотеки им. М. Горького СПбГУ:
* <http://www.library.spbu.ru/cgi-bin/irbis64r/cgiirbis_64.exe?C21COM=F&I21DBN=IBIS&P21DBN=IBIS>
* Перечень электронных ресурсов, находящихся в доступе СПбГУ:
* <http://cufts.library.spbu.ru/CRDB/SPBGU/>
* Перечень ЭБС, на платформах которых представлены российские учебники, находящиеся в доступе СПбГУ:
* <http://cufts.library.spbu.ru/CRDB/SPBGU/browse?name=rures&resource_type=8>
* ACM Digital Library: <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/12>
* Institute of Electrical and Electronics Engineers (IEEE): <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/375>
* MathSciNet - электронная коллекция Американского математического сообщества (AMS): <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/415>
* O’Reilly: <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/483>
* **Zentralblatt MATH:** <http://cufts.library.spbu.ru/CRDB/SPBGU/resource/86>

• <http://statmod.ru/wiki/>

<http://www.machinelearning.ru>

**Раздел 4. Разработчики программы**

Голяндина Нина Эдуардовна, к.ф.-м.н., доцент Кафедры статистического моделирования СПбГУ, n.golyandina@spbu.ru, тел. +7(812) 428-42-31