

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»**

**УТВЕРЖДЕНО**

**Директор физтех-школы  
прикладной математики и  
информатики**

**А.М. Райгородский**

|                            |  |
|----------------------------|--|
|                            | <b>Рабочая программа дисциплины (модуля)</b>   |
| <b>по дисциплине:</b>      | Статистическая теория машинного обучения   |
| <b>по направлению:</b>     | Прикладная математика и информатика  |
| <b>профиль подготовки:</b> | А1360: Передовые методы искусственного интеллекта<br>Физтех-школа Прикладной Математики и Информатики<br>кафедра математических основ управления |
| <b>курс:</b>               | 4  |
| <b>квалификация:</b>       | бакалавр   |

Семестр, формы промежуточной аттестации: 7 (осенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Программу составил: Н.А. Пучкин

Программа обсуждена на заседании кафедры математических основ управления 12.02.2024

## Аннотация

В курсе рассматриваются ключевые понятия и методы статистической теории машинного обучения, исследующей проблему надежности восстановления зависимостей по эмпирическим данным. Прежде всего ставится задача классификации, вводятся понятия РАС-обучения, агностического РАС-обучения, переобучения, обобщающей способности алгоритмов. Даются понятия размерности Вапника-Червоненкиса и средних по Радемахеру, играющие важную роль в анализе алгоритмов, минимизирующих эмпирическую ошибку. Обсуждаются неравенства концентрации меры, включая неравенство Хеффдинга, неравенство МакДиармида и неравенство Бернштейна. Проводится анализ алгоритмов машинного обучения, таких как метод k ближайших соседей, метод опорных векторов, персептрон, нейронные сети. Курс содержит обсуждение базовых вопросов статистической теории машинного обучения, разбор задач. Для успешного освоения курса слушателю необходимо владеть основами теории вероятностей.

### 1. Цели и задачи

#### Цель дисциплины

- изучение основных понятий и методов статистической теории машинного обучения.

#### Задачи дисциплины

- освоение студентами базовых знаний в области машинного обучения;
- приобретение теоретических знаний в области Байесовской теории машинного обучения;
- оказание консультаций и помощи студентам в решении теоретических и практических задач.

### 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

| Код и наименование компетенции  | Индикаторы достижения компетенции   |
|---|---|
| ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности | ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности                     |
|   | ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области |
| ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию   | ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации                                       |
|   | ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива                |

### 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные понятия, законы, методы статистической теории машинного обучения;
- основные свойства соответствующих математических объектов;
- аналитические и численные подходы и методы для решения типовых прикладных задач.

уметь:

- понять поставленную задачу;
- использовать свои знания для решения фундаментальных и прикладных задач теории машинного обучения;
- оценивать корректность постановок задач;
- строго доказывать или опровергать утверждение;
- самостоятельно находить алгоритмы решения задач теории машинного обучения, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

владеть:

- навыками освоения большого объема информации и решения задач;
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов и методов теории машинного обучения;
- предметным языком дискретной математики и навыками грамотного описания решения задач и представления полученных результатов.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

| №                     | Тема (раздел) дисциплины                        | Трудоемкость по видам учебных занятий, включая самостоятельную работу, час. |          |                 |                |
|-----------------------|---|---|----------|-----------------|----------------|
|                       |   | Лекции  | Семинары | Лаборат. работы | Самост. работа |
| 1                     | Постановка задачи классификации. РАС-обучение.  | 10  | 5        |                 | 10             |
| 2                     | Метод опорных векторов.                         | 10  | 5        |                 | 10             |
| 3                     | Анализ избранных алгоритмов машинного обучения. | 5   | 10       |                 | 10             |
| 4                     | Минимизация эмпирического риска.                | 5   | 10       |                 | 15             |
| Итого часов           |   | 30  | 30       |                 | 45             |
| Подготовка к экзамену |   | 30 час.   |          |                 |                |
| Общая трудоёмкость    |   | 135 час., 3 зач.ед.   |          |                 |                |

##### 4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

Семестр: 7 (Осенний)

###### 1. Постановка задачи классификации. РАС-обучение.

Постановка задачи обучения. Явление переобучения. РАС-обучаемость и агностическая РАС-обучаемость. Минимаксные порядки. Необучаемость класса всех функций. Принцип равномерной сходимости. Агностическая обучаемость конечных классов. Функция роста. Оценка на предсказательную способность алгоритма через функцию роста в бесшумном случае. Размерность Вапника-Червоненкиса. Лемма Зауэра. Среднее по Радемахеру.

###### 2. Метод опорных векторов.

Метод опорных векторов в случае разделимой выборки. Обобщающая способность метода опорных векторов в случае разделимой выборки. Метод опорных векторов в случае неразделимой выборки. Переменные мягкого отступа. Обобщающая способность метода опорных векторов в случае неразделимой выборки. Метод опорных векторов в пространстве признаков. Пространства, порожденные воспроизводящим ядром (RKHS). Теорема о представителе. Обобщающая способность метода опорных векторов в случае разделимой выборки в пространстве признаков. Положительно и отрицательно определенные ядра и их свойства. Теорема Мерсера.

###### 3. Анализ избранных алгоритмов машинного обучения.

Условие малого шума Маммена-Цыбакова. Оценка предсказательной способности алгоритма в условиях малого шума. Метод  $k$  ближайших соседей. Быстрые порядки для plug-in классификаторов. Схемы сжатия выборок. Оценка скорости обучения в классе со схемой сжатия размера  $k$ . Схемы сжатия выборок с потерями. Оценка скорости обучения в классе со схемой сжатия с потерями размера  $k$ . Персептрон. Верхняя оценка числа итераций алгоритма в случае линейно разделимой выборки. Нейронные сети. Оценка обобщающей способности нейронных сетей.

#### 4. Минимизация эмпирического риска.

Оценка на предсказательную способность алгоритма через среднее по Радемахеру. Число покрытия и число упаковки. Оценка на среднее по Радемахеру через число покрытия. Фундаментальная теорема PAC-обучения.

### 5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Необходимое оборудование для лекций: компьютер и мультимедийное оборудование (проектор, звуковая система).

### 6. Перечень рекомендуемой литературы

#### Основная литература

Колмогоровская сложность и алгоритмическая случайность [Текст] : учеб. пособие для вузов / В. В. Вьюгин ; М-во образования и науки РФ, Моск. физ.-техн. ин-т (гос. ун-т), Ин-т проблем информации им. А. А. Харкевича .— М. : МФТИ, 2012 .— 140 с.

#### Дополнительная литература

Вероятность [Текст] : в 2 т. : учебник для вузов / А. Н. Ширяев .— 4-е перераб. и доп. — М. : МЦНМО, 2007, 2011 .— Т. 2 : Суммы и последовательности случайных величин - стационарные, мартингалы, марковские цепи. - 2007, 2011. - 416 с.

### 7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)

<https://www.dropbox.com/sh/1ci86mgxjjnxc96/AABaLLkJ2dnclXx-C2ar6cbSa?dl=0>  
<http://www.iitp.ru/ru/userpages/>

### 8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)

На лекциях используется компьютер и мультимедийное оборудование (проектор, звуковая система),

### 9. Методические указания для обучающихся по освоению дисциплины (модуля)

Студент, изучающий дисциплину, должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

Успешное освоение дисциплины требует:

- посещения студентом всех видов аудиторных занятий;
- ведения конспекта в ходе лекционных занятий;
- качественной самостоятельной подготовки к практическим занятиям, активной работы на них;
- активной самостоятельной и аудиторной работы студента;
- своевременной сдачи преподавателю заданий по аудиторным видам работ.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

**по направлению:** Прикладная математика и информатика  
**профиль подготовки:** АІ360: Передовые методы искусственного интеллекта  
Физтех-школа Прикладной Математики и Информатики  
кафедра математических основ управления  
**курс:** 4  
**квалификация:** бакалавр  
Семестр, формы промежуточной аттестации: 7 (осенний) - Экзамен  
**Разработчик:** Н.А. Пучкин

## 1. Компетенции, формируемые в процессе изучения дисциплины

| Код и наименование компетенции  | Индикаторы достижения компетенции   |
|---|---|
| ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности | ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности                     |
|   | ОПК-2.2 Знает и умеет применять численные математические методы и прикладное программное обеспечение для решения научных задач в профессиональной области |
| ПК-2 Способен самостоятельно или в качестве члена (руководителя) малого коллектива организовывать и проводить научные исследования и их апробацию   | ПК-2.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации                                       |
|   | ПК-2.2 Способен планировать и проводить научные исследования самостоятельно или в качестве члена (руководителя) малого научного коллектива                |

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Статистическая теория машинного обучения» обучающийся должен:

### знать:

- фундаментальные понятия, законы, методы статистической теории машинного обучения;
- основные свойства соответствующих математических объектов;
- аналитические и численные подходы и методы для решения типовых прикладных задач.

### уметь:

- понять поставленную задачу;
- использовать свои знания для решения фундаментальных и прикладных задач теории машинного обучения;
- оценивать корректность постановок задач;
- строго доказывать или опровергать утверждение;
- самостоятельно находить алгоритмы решения задач теории машинного обучения, в том числе и нестандартных, и проводить их анализ;
- самостоятельно видеть следствия полученных результатов.

### владеть:

- навыками освоения большого объема информации и решения задач;
- навыками самостоятельной работы и освоения новых дисциплин;
- культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов и методов теории машинного обучения;
- предметным языком дискретной математики и навыками грамотного описания решения задач и представления полученных результатов.

## 3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

## 4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся

1. Постановка задачи классификации. Байесовский классификатор. Примеры классификаторов: перцептрон, нейронные сети.
2. РАС-теория ошибок. Теория обобщения Вапника – Червоненкиса. Верхние оценки ошибки классификации.
3. VC-размерность. Лемма Вапника – Червоненкиса (Сауэра, Шелаха)

4. VC-размерность класса линейных классификаторов. Примеры вычисления VC-размерности других классов функций.
5. Теория обобщения для задач классификации с помощью пороговых решающих правил. Число покрытия для классов функций. Оценка ошибки обобщения через число покрытия.
6. Пороговая размерность. Оценка ошибки обобщения через пороговую размерность.
7. Покрытия и упаковки в метрических пространствах. Теорема Алона, Бен-Давида, Хауслера и Чеза-Бьянки.
8. Средние по Радемахеру. Равномерная оценка отклонения эмпирического среднего от математического ожидания для класса функций.
9. Неравенство Мак-Диармонда и его применения.
10. Среднее Радемахера композиции.
11. Средние по Радемахеру и другие меры емкости классов функций (VC-размерность, число покрытия).
12. Оценка ошибки обобщения с помощью среднего по Радемахеру.
13. Алгоритм построения оптимальной разделяющей гиперплоскости. Задача оптимизации. Опорные векторы.
14. SVM-метод в пространстве признаков. Пространства, порожденные воспроизводящим ядром (RKHS) и их свойства.
15. Построение канонического RKHS.
16. Теорема о представителе.
17. Случай неразделимой выборки. Вектор переменных мягкого отступа. Оценка ошибки в случае неразделимой выборки.
18. Задача оптимизации для классификации с ошибками в квадратичной норме.
19. Задача оптимизации для классификации с ошибками в линейной норме.
20. Многомерная регрессия с помощью SVM. Гребневая регрессия.
21. Конформные предсказания. Метаалгоритм. Примеры мер неконформности.

Примеры экзаменационных билетов:

Билет 1

1. PAC-теория ошибок. Теория обобщения Вапника – Червоненкиса. Верхние оценки ошибки классификации.
2. Задача оптимизации для классификации с ошибками в квадратичной норме.

Билет 2

1. VC-размерность. Лемма Вапника – Червоненкиса.
2. Оценка ошибки обобщения через пороговую размерность.

#### Критерии оценивания

Оценка "Отлично" (10) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы, код оформлен в едином удобочитаемом стиле.

Оценка "Отлично" (9) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач, реализованы оптимальные алгоритмы.

Оценка "Отлично" (8) - полностью и вовремя решены все задачи без ошибок. Продemonстрирован грамотный подход к решению задач.

Оценка "Хорошо" (7) - полностью решены все задачи. Допущены несущественные ошибки.

Оценка "Хорошо" (6) - полностью решено большинство задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Хорошо" (5) - полностью решено две трети задач. В некоторых задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (4) - полностью решено более половины задач. В остальных задачах допущены и не исправлены ошибки, либо некоторые задачи решены частично.

Оценка "Удовлетворительно" (3) - полностью решено более половины задач.

Оценка "Неудовлетворительно" (2) - решено менее половины задач.

Оценка "Неудовлетворительно" (1) - не решено ни одной задачи.

## **5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Экзамен может проводиться по итогам текущей успеваемости и сдачи заданий и других видов работ, предусмотренных программой дисциплины и (или) путем организации специального опроса, проводимого в устной и (или) письменной форме.

При проведении устного экзамена обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося по билету не должен превышать одного астрономического часа.

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, конспектами лекций или другими материалами.