

**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

Рабочая программа дисциплины (модуля)

по дисциплине:	Математические основы машинного обучения. Дополнительные главы
по направлению:	Прикладная математика и физика
профиль подготовки:	Радиотехника и компьютерные технологии Физтех-школа Радиотехники и Компьютерных Технологий кафедра проблем передачи информации и анализа данных
курс:	4
квалификация:	бакалавр

Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 15 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 15 час.

Всего часов: 45, всего зач. ед.: 1

Программу составил: В.В. Вьюгин, д-р физ.-мат. наук, профессор

Программа обсуждена на заседании кафедры проблем передачи информации и анализа данных 05.06.2023

Аннотация

В курсе рассматриваются ключевые понятия и современные математические методы машинного обучения, а также их приложения в различных областях. В первой части курса рассматриваются статистические методы машинного обучения. Решается задача классификации при предположении, что данные распределены независимо по одному и тому же неизвестному для нас распределению. Вводится понятие VC-размерности Вапника — Червоненкиса и приводятся оценки вероятности ошибки обобщения для построенного классификатора. Изучаются алгоритмы построения машин на опорных векторах. Вводится и изучается понятие среднего Радемахера, приводятся оценки вероятности ошибки в терминах среднего Радемахера. Изучается понятие гильбертова пространства, порожденного воспроизводящим ядром. Вторая часть курса содержит обзор основных понятий и классических алгоритмов Литлстоуна и Вармута, Вовка, Фройнда и Шапире. Получены оценки регрета этих алгоритмов. Приведены также их приложения для построения алгоритма бустинга AdaBoost и универсального портфеля (финансовых инструментов) Ковера. Характерная черта данного подхода — отсутствие предположения о существовании стохастических механизмов, генерирующих данные. Вместо этого, рассматриваются конкурирующие классы методов прогнозирования, на основании которых строится оптимальный метод, который прогнозирует не хуже, чем любой метод из заданного класса. Для успешного освоения курса слушателю желательно знать основные разделы общей теории алгоритмов и обладать элементарными знаниями в области теории вероятностей. Слушателям курса предлагается большое количество задач по указанным темам.

1. Цели и задачи

Цель дисциплины

Дать краткий обзор основных математических методов и алгоритмов машинного обучения, наиболее широко обсуждаемых в мировой научной литературе последних лет.

Задачи дисциплины

- освоение задач статистической теории машинного обучения, задач классификации и регрессии с опорными векторами, теории обобщения Вапника—Червоненкиса и алгоритмов построения разделяющих гиперплоскостей;
- освоение задач адаптивного прогнозирования в режиме онлайн в теоретико-игровой и сравнительной постановке: игры с рандомизированными предсказаниями, предсказания использованием экспертных стратегий.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений
ПК-4 Способен критически оценивать применимость используемых методик и методов	ПК-4.2 Знает источники происхождения и умеет производить оценку погрешности измерений и достоверности экспериментальных результатов

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны знать:

- статистическую теорию машинного обучения, методы классификации и регрессии с опорными векторами, теорию обобщения Вапника—Червоненкиса и алгоритмы построения разделяющих гиперплоскостей;
- методы теории адаптивного прогнозирования в режиме онлайн в теоретико-игровой и сравнительной постановке: игры с рандомизированными предсказаниями, предсказания с использованием экспертных стратегий;
- основные понятия теории игр.

уметь:

- применять основные математические методы и алгоритмы теории машинного обучения;
- применять методы теории адаптивного прогнозирования в режиме онлайн.

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач.

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	Агрегирующий алгоритм Вовка	3	3		3
2	Универсальные предсказания	3	3		3
3	Элементы сравнительной теории машинного обучения	3	3		3
4	Элементы теории игр	3	3		3
5	Элементы теории классификации и регрессии с опорными векторами	3	3		3
Итого часов		15	15		15
Подготовка к экзамену		0 час.			
Общая трудоёмкость		45 час., 1 зач.ед.			

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

Семестр: 8 (Весенний)

1. Агрегирующий алгоритм Вовка

Экспоненциально выпуклые функции потерь. Агрегирующий алгоритм для конечного числа экспертов.

Агрегирующий алгоритм для бесконечного пространства экспертов. Агрегирующий псевдоалгоритм, функция подстановки.

Агрегирующий алгоритм для конечного числа экспертов. Игра с логарифмической функцией потерь. Построение функции подстановки.

Агрегирующий алгоритм для конечного числа экспертов. Простая игра на предсказания. Построение функции подстановки.

Агрегирующий алгоритм для конечного числа экспертов. Игра с квадратичной функцией потерь. Построение функции подстановки.

Многомерная он-лайн регрессия с помощью агрегирующего алгоритма. Алгоритм многомерной линейной регрессии, Оценки ошибки предсказания.

Универсальный портфель. Применение агрегирующего алгоритма для построения универсального портфеля.

2. Универсальные предсказания

Задача универсального прогнозирования в режиме он-лайн: статистический подход.

Калибруемость прогнозов. Алгоритм вычисления хорошо калибруемых прогнозов.

Прогнозирование с произвольным ядром.

3. Элементы сравнительной теории машинного обучения

Алгоритм оптимального распределения потерь в режиме он-лайн.

Задача нахождения оптимальных решений с учетом экспертных стратегий. Алгоритм экспоненциального взвешивания экспертных решений.

Рандомизированные прогнозы.

Усиление простых классификаторов – Boosting. Алгоритм AdaBoost.

4. Элементы теории игр

Антагонистические игры двух игроков. Достаточное условие существования седловой точки.

Достаточное условие существования седловой точки. Смешанные расширения матричных игр. Минимаксная теорема.

Чистые стратегии. Решение матричной игры типа $2 \times M$. Решение игры типа $N \times M$.

Бесконечные игры с рандомизированными предсказаниями. Построение выигрышной стратегии предсказателя с использованием минимаксной теоремы.

Хорошо калибруемые предсказания. Универсальная калибруемость со счетным числом правил выбора.

5. Элементы теории классификации и регрессии с опорными векторами

Постановка задачи классификации. Байесовский классификатор. Линейные классификаторы: персептрон. Алгоритм Розенблатта. Теорема Новикова о сходимости.

Теория обобщения Вапника–Червоненкиса. Верхняя оценка вероятности ошибки классификации через VC-размерность класса функций классификации.

VC-размерность, определение, основное свойство. VC-размерность класса всех линейных (однородных) классификаторов.

Метод опорных векторов. Оптимальная гиперплоскость. Алгоритм построения оптимальной гиперплоскости. Оценка вероятности ошибки обобщения через число опорных векторов.

SVM - метод в пространстве признаков, примеры. Ядра.

Случай неотделимой выборки. Вектор переменных мягкого отступа. Оценка вероятности ошибки обобщения. Оптимизационная задача для классификации с ошибками в квадратичной норме.

Случай неотделимой выборки. Вектор переменных мягкого отступа. Оценка вероятности ошибки обобщения. Оптимизационная задача для классификации с ошибками в линейной норме. Оптимизационная задача для классификации с ошибками в форме задачи линейного программирования.

Задача многомерной регрессии. Простая линейная регрессия. Гребневая регрессия.

Задача многомерной регрессии. Регрессия с опорными векторами. Ошибка обобщения при регрессии. Решение задачи гребневой регрессии с помощью SVM.

Гребневая регрессия в прямой форме и в двойственной форме как частный случай регрессии с опорными векторами в случае квадратичной функции потерь. Нелинейная многомерная гребневая регрессия (с ядром).

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

Компьютерный класс, доска, ноутбук и мультимедийное оборудование (проектор или плазменная панель).

6. Перечень рекомендуемой литературы

Основная литература

1. Математические основы машинного обучения и прогнозирования, Электронная версия печатной публикации / В. В. Вьюгин. — Москва, МЦНМО, 2014

Дополнительная литература

фонд литературы кафедры

1. Cristianini N., Shawe-Taylor J. An introduction to support vector machines (and other kernel-based learning method). Cambridge UP, 2000. 187 p. [имеется в библиотечном фонде кафедры]
2. Cesa-Bianchi N., Lugosi G. Prediction, Learning and Games. New York: Cambridge University Press, 2006. 394 p. [имеется в библиотечном фонде кафедры]

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

Не используются

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

Программное обеспечение и информационные технологии не требуются.

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. В результате изучения дисциплины студент должен знать основные определения, понятия, аксиомы, методы доказательств.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы,
- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- выполнение лабораторных работ, для осознания связей между теорией и практическими навыками.
- подготовку к дифференцированному зачету.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению:	Прикладные математика и физика
профиль подготовки:	Радиотехника и компьютерные технологии Физтех-школа Радиотехники и Компьютерных Технологий кафедра проблем передачи информации и анализа данных
курс:	4
квалификация:	бакалавр
Семестр, формы промежуточной аттестации: 8 (весенний) - Дифференцированный зачет	
Разработчик:	В.В. Вьюгин, д-р физ.-мат. наук, профессор

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
ПК-3 Способен выбирать и применять подходящее оборудование, инструменты и методы исследований для решения задач в избранной предметной области	ПК-3.2 Знает области и критерии применимости используемых теоретических подходов и умение оценивать точность приближенных аналитических методов вычислений
ПК-4 Способен критически оценивать применимость используемых методик и методов	ПК-4.2 Знает источники происхождения и умеет производить оценку погрешности измерений и достоверности экспериментальных результатов

2. Показатели оценивания компетенций

В результате изучения дисциплины «Математические основы машинного обучения. Дополнительные главы» обучающийся должен:

знать:

- статистическую теорию машинного обучения, методы классификации и регрессии с опорными векторами, теорию обобщения Вапника—Червоненкиса и алгоритмы построения разделяющих гиперплоскостей;
- методы теории адаптивного прогнозирования в режиме онлайн в теоретико-игровой и сравнительной постановке: игры с рандомизированными предсказаниями, предсказания с использованием экспертных стратегий;
- основные понятия теории игр.

уметь:

- применять основные математические методы и алгоритмы теории машинного обучения;
- применять методы теории адаптивного прогнозирования в режиме онлайн.

владеть:

- навыком освоения большого объема информации;
- навыками постановки научно-исследовательских задач.

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

1. Задача классификации. Линейные классификаторы: персептрон.
2. Метод опорных векторов.
3. Оптимальная гиперплоскость.
4. Алгоритм построения оптимальной гиперплоскости.
5. Случай неразделимой выборки.
6. Вектор переменных мягкого отступа.
7. Оптимизационная задача для классификации с ошибками.
8. Регрессия с опорными векторами.
9. Решение задачи регрессии с помощью SVM.
10. Прогнозирование с произвольным ядром
11. Алгоритм экспоненциального взвешивания экспертных решений.
12. Агрегирующий алгоритм Вовка.
13. Конечное множество экспертов.
14. Агрегирующий алгоритм для квадратичной функции потерь.
15. Антагонистические игры двух игроков.

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

1. Верхние оценки вероятности ошибки классификации. VC-размерность.
2. SVM-метод в пространстве признаков.
3. Задача многомерной регрессии.
4. Простая линейная регрессия.
5. Гребневая регрессия.
6. Универсальное прогнозирование в режиме онлайн.
7. Калибруемость прогнозов.
8. Алгоритм вычисления калибруемых прогнозов.
9. Алгоритм оптимального распределения, потерь в режиме онлайн.
10. Усиление простых классификаторов - бустинг.
11. Агрегирующий алгоритм для логарифмической функции потерь.
12. Многомерная онлайн регрессия.
13. Достаточное условие существования седловой точки.
14. Смешанные расширения матричных игр.
15. Минимаксная теорема.

Критерии оценивания

Оценка отлично 10 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины, проявляющему интерес к данной предметной области, продемонстрировавшему умение уверенно и творчески применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 9 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений.

Оценка отлично 8 баллов - выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, правильное обоснование принятых решений, с некоторыми недочетами.

Оценка хорошо 7 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но недостаточно грамотно обосновывает полученные результаты.

Оценка хорошо 6 баллов - выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности.

Оценка хорошо 5 баллов - выставляется студенту, если он в основном знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач достаточно большое количество неточностей.

Оценка удовлетворительно 4 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он освоил основные разделы учебной программы, необходимые для дальнейшего обучения, и может применять полученные знания по образцу в стандартной ситуации.

Оценка удовлетворительно 3 балла - выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, допускающему ошибки в формулировках базовых понятий, нарушения логической последовательности в изложении программного материала, слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и с трудом применяет полученные знания даже в стандартной ситуации.

Оценка неудовлетворительно 2 балла - выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных принципов и не умеет использовать полученные знания при решении типовых задач.

Оценка неудовлетворительно 1 балл - выставляется студенту, который не знает основного содержания учебной программы дисциплины, допускает грубейшие ошибки в формулировках базовых понятий дисциплины и вообще не имеет навыков решения типовых практических задач.

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

Во время проведения дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций.

Дифференцированный зачет может проводиться по итогам текущей успеваемости и сдачи заданий, или путем организации специального опроса, проводимого в устной форме.