

**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:	Статистическая теория машинного обучения
по направлению:	Информатика и вычислительная техника
профиль подготовки:	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра математических основ управления
курс:	1
квалификация:	магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Аудиторных часов: 45 всего, в том числе:

лекции: 30 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 60 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составил: Н.А. Пучкин

Программа обсуждена на заседании кафедры математических основ управления 03.04.2024

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 Хефдинга, 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-1 Владеет системой фундаментальных научных знаний в области информатики и вычислительной техники	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области информатики и вычислительной техники
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области информатики и вычислительной техники и способен их применять при решении задач профессиональной деятельности
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
ПК-3 Владеет основами ведения научной дискуссии и формы устного научного высказывания	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания

ПК-5 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований	ПК-3.2 Умеет вести корректную дискуссию в области информационных технологий задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

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. Минимизация эмпирического риска.	10	5		20
2	Метод опорных векторов.	10	5		20
3	Анализ избранных алгоритмов машинного обучения.	10	5		20
Итого часов		30	15		60
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 1 (Осенний)

1. Постановка задачи классификации. РАС-обучение. Минимизация эмпирического риска.

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

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.

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

1. 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

1. 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/1ci86mgxjnx96/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)

Не предусмотрено.

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

Студент, изучающий курс "Статистическая теория машинного обучения", 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;
- подготовку к экзамену.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций. Показателем владения материалом служит умение решать задачи. Для формирования умения применять теоретические знания на практике студенту необходимо решать как можно больше задач. При решении задач каждое действие необходимо аргументировать, ссылаясь на известные теоретические сведения. Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору или преподавателю, ведущему практические занятия.

Литература для самостоятельной работы:

1. Вьюгин В.В.. Математические основы машинного обучения и прогнозирования. – М.: МЦНМО, 2013.
2. Bousquet, O., Boucheron, S., and Lugosi, G.: Introduction to statistical learning theory. in: Advanced Lectures on Machine Learning. pp. 169--207 (2004)
3. Steinwart, I.: On the influence of the kernel on the consistency of support vector machines. Journal of Machine Learning Research. 2, 67--93 (2001)
4. Rakhlin, A., Sridharan, K., Tewari, A.: Online learning: Beyond regret. In Proceedings of the 24rd Annual Conference on Learning Theory, v/ 19 of JMLR Workshop and Conference Proceedings, pages 559--594, 2011. longer version available as arXiv:1011.3168 (2011)
5. S Bubeck and N. Cesa-Bianchi, Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. In Foundations and Trends in Machine Learning, Vol 5: No 1, 1-122, 2012.

6. S. Shalev-Shwartz and S. Ben-David. «Understanding Machine Learning: From Theory to Algorithms». Cambridge University Press, USA, 2014.
7. S. Boucheron, O. Bousquet, G. Lugosi, «Theory of Classification: a Survey of Some Recent Advances», ESAIM: PS 9 323-375 (2005), DOI: 10.1051/ps:2005018

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника
профиль подготовки: Прикладная математика и информатика
Физтех-школа Прикладной Математики и Информатики
кафедра математических основ управления
курс: 1
квалификация: магистр

Семестр, формы промежуточной аттестации: 1 (осенний) - Экзамен

Разработчик: Н.А. Пучкин

1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Владеет системой фундаментальных научных знаний в области информатики и вычислительной техники	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области информатики и вычислительной техники
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области информатики и вычислительной техники и способен их применять при решении задач профессиональной деятельности
ПК-2 Понимает и способен применить в научно-исследовательской и прикладной деятельности основные законы естествознания, современный математический аппарат и алгоритмы, современные информационно-коммуникационные технологии	ПК-2.1 Знает основы научно-исследовательской деятельности в области информационных технологий, владеет знанием основ философии и методологии науки; знанием методов научных исследований и навыками их проведения
	ПК-2.2 Умеет применять полученные знания в области фундаментальных научных основ теории информации и решать стандартные задачи в собственной научно-исследовательской деятельности
	ПК-2.3 Имеет практический опыт научно-исследовательской деятельности в области информационно-коммуникационных технологий
	ПК-2.4 Владеет методами и алгоритмами решения задач цифровой обработки сигналов, использования сети Интернет, аннотирования, реферирования, библиографического поиска, опыт работы с научными источниками
ПК-3 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания
	ПК-3.2 Умеет вести корректную дискуссию в области информационных технологий задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий
ПК-1 Готов к включению в профессиональное сообщество; способен проводить под научным руководством локальные исследования на основе существующих методов в конкретной области профессиональной деятельности	ПК-1.1 Знает принципы построения научной работы, методы сбора и анализа полученного материала, способы аргументации; владеет навыками подготовки научных обзоров, публикаций, рефератов и библиографий по тематике проводимых исследований на русском и английском языке
	ПК-1.2 Умеет решать научные задачи с пониманием существующих подходов к верификации моделей программного обеспечения в связи с поставленной целью и в соответствии с выбранной методикой
	ПК-1.3 Имеет практический опыт выступлений и научной аргументации при анализе объекта научной и профессиональной деятельности

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. РАС-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)» 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 (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;

оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений;

оценка «хорошо (7)» 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 (6)» выставляется студенту, если он 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)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;

оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;

оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач;

оценка «неудовлетворительно (1)» 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 экзамена обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций. Экзамен проводится путем организации специального опроса, проводимого в устной форме.