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: Введение в машинное обучение
по направлению: Информатика и вычислительная техника
профиль подготовки: Компьютерные технологии и вычислительная техника
Физтех-школа Радиотехники и Компьютерных Технологий
кафедра машинного обучения и цифровой гуманитаристики
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 60 всего, в том числе:
лекции: 30 час.
семинары: 30 час.
лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

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

Программа обсуждена на заседании кафедры машинного обучения и цифровой гуманитаристики 15.05.2021

Аннотация

Курс посвящён основным понятиям и концепциям современного машинного обучения. Так, рассматриваются базовый алгоритм (алгоритмический оператор), корректирующая операция, взвешенное голосование, композиции, критерии выбора моделей, методы отбора признаков и ранжирования.

Много внимания уделяется обучению с подкреплением, в частности, жадным, адаптивным, оптимальным стратегиям; уравнению Беллмана, методам временных разностей.

Рассматриваются задачи с частичным обучением, коллаборативная фильтрация, тематическое моделирование, байесовское обучение и введение в глубинное обучение.

1. Цели и задачи

Цель дисциплины

- сформировать теоретические и практические знания в области обучения машин, современных методов восстановления зависимостей по эмпирическим данным, включая дискриминантный, кластерный и регрессионный анализ, частичное обучение.

Задачи дисциплины

- освоить методы корректной формулировки задач в терминах машинного обучения;
- овладеть навыками практического решения задач интеллектуального анализа данных.

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-2.3 Знает основные требования информационной безопасности
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные принципы и проблематику теории обучения машин,
- основные современные методы обучении по прецедентам — классификации, кластеризации и регрессии.

уметь:

- формализовать постановки прикладных задач анализа данных;
- использовать методы обучения по прецедентам для решения практических задач;
- оценивать точность и эффективность полученных решений.

владеть:

- основными понятиями теории машинного обучения.

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	Композиции классификаторов, бустинг	2	2		3
2	Критерии выбора моделей	2	2		3
3	Методы отбора признаков	4	4		3
4	Методы ранжирования	4	4		3
5	Обучение с подкреплением	4	4		3
6	Задачи с частичным обучением	4	4		3
7	Коллаборативная фильтрация	2	2		3
8	Тематическое моделирование	2	2		3
9	Байесовское обучение	4	4		3
10	Введение в глубинное обучение	2	2		3
Итого часов		30	30		30
Подготовка к экзамену		0 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

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

Семестр: 7 (Осенний)

1. Композиции классификаторов, бустинг

• Основные понятия: базовый алгоритм (алгоритмический оператор), корректирующая операция.
• Взвешенное голосование.
• Алгоритм AdaBoost. Экспоненциальная аппроксимация пороговой функции потерь. Процесс последовательного обучения базовых алгоритмов. Теорема о сходимости бустинга.
• Обобщение бустинга как процесса градиентного спуска. Теорема сходимости. Алгоритм AnyBoost.
• Простое голосование (комитет большинства). Эвристический алгоритм ComBoost. Идентификация нетипичных объектов (выбросов). Обобщение на большое число классов.
• Решающий список (комитет старшинства). Эвристический алгоритм. Стратегия выбора классов для базовых алгоритмов.
• Стохастические методы: бэггинг и метод случайных подпространств.
• Нелинейные алгоритмические композиции. Смесь экспертов, область компетентности алгоритма. Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический. Построение смесей экспертов с помощью EM-алгоритма.

2. Критерии выбора моделей

• Внутренние и внешние критерии.
• Эмпирические и аналитические оценки функционала полного скользящего контроля.
• Скользящий контроль, разновидности эмпирических оценок скользящего контроля.
• Критерий непротиворечивости.
• Регуляризация. Критерий Акаике (AIC). Байесовский информационный критерий (BIC).
• Агрегированные и многоступенчатые критерии.

3. Методы отбора признаков

• Усечённый поиск в ширину, многорядный итерационный алгоритм МГУА.
• Генетический алгоритм, его сходство с МГУА.
• Случайный поиск и Случайный поиск с адаптацией (СПА).

4. Методы ранжирования

• Постановка задачи ранжирования.
• Примеры прикладных задач.
• Признаки в задаче ранжирования поисковой выдачи: текстовые, ссылочные, кликовые.
• Критерии качества ранжирования.
• Точечный, попарный и списочный подходы.

5. Обучение с подкреплением

• Задача о многоруком бандите. Жадные и эпсилон-жадные стратегии. Среда для экспериментов. Метод сравнения с подкреплением. Метод преследования.
• Адаптивные стратегии на основе скользящих средних.
• Уравнения Беллмана. Оптимальные стратегии. Динамическое программирование. Метод итераций по ценностям и по стратегиям.
• Методы временных разностей: TD, SARSA, Q-метод. Многошаговое TD-прогнозирование. Адаптивный полужадный метод VDBE.

6. Задачи с частичным обучением

• Постановка задачи Semisupervised Learning, примеры приложений.
• Простые эвристические методы: self-training, co-training, co-learning.
• Адаптация алгоритмов кластеризации для решения задач с частичным обучением. Кратчайшиё незамкнутый путь. Алгоритм Ланса-Уильямса. Алгоритм k-средних.
• Трансдуктивный метод опорных векторов TSVM.
• Алгоритм Expectation-Regularization на основе многоклассовой регуляризированной логистической регрессии.

7. Коллаборативная фильтрация

• Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты—объекты.
• Корреляционные методы user-based, item-based.
• Латентные методы на основе би-кластеризации. Алгоритм Брегмана.
• Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных. Метод стохастического градиента.
• Неотрицательные матричные разложения. Вероятностный латентный семантический анализ PLSA. EM-алгоритм для PLSA.
• Эксперименты на данных конкурса «Интернет-математика» 2005.

8. Тематическое моделирование

• Задачи тематического моделирования, коллекции текстовых документов и матрица документы—слова. Перплексия как мера качества тематической модели. Задача тематического поиска.
• Униграммная модель документа. Метод максимума правдоподобия и метод максимума апостериорной вероятности. Применение метода множителей Лагранжа.
• Вероятностный латентный семантический анализ PLSA. EM-алгоритм. Инкрементное добавление новых документов (folding-in). Задача с частичным обучением.

• Латентное размещение Дирихле. Сглаженная частотная оценка вероятности. Сэмплирование Гиббса. Оптимизация гиперпараметров.

• Робастная тематическая модель с фоновой и шумовой компонентой. Эксперименты по сравнению робастных и регуляризованных моделей.

9. Байесовское обучение

• Понятие условной независимости, графические модели.
• Байесовские сети.
• Марковские поля.
• Скрытые марковские модели.
• Условные случайные поля.

10. Введение в глубинное обучение

• Рекуррентные нейросети, сверточные нейросети
• Примеры прикладных задач, успешно решаемых с помощью глубинного обучения.
• Ограниченная машина Больцмана.

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

Учебная аудитория, оснащенная мультимедиапроектором и экраном.

6.Перечень рекомендуемой литературы

Основная литература

1. Математическая статистика [Текст] : [учебник для вузов] / А. А. Боровков .— [3-е изд., испр.] .— М. : Физматлит, 2007 .— 704 с.

2. Математическая статистика [Текст] : учеб. пособие для вузов / А. А. Натан, О. Г. Горбачев, С. А. Гуз ; Моск. физико-техн.ин-т (гос.ун-т .— М : МЗ Пресс, 2004, 2005 .— 160 с.

3. Python и машинное обучение [Текст], крайне необходимое издание по новейшей предсказательной аналитике для более глубокого понимания методологии машинного обучения/С. Рашка, -М., ДМК Пресс, 2017

4. Python для сложных задач: наука о данных и машинное обучение [Текст], [учеб. пособие для вузов] /Дж. Вандер Плас ; [пер. с англ. И. Пальти]. -СПб., Питер, 2018

Дополнительная литература

1. Математическая статистика [Текст] : оценка параметров,проверка гипотез:учеб.пособие для вузов:доп.М-вом образования СССР / А. А. Боровков .— М. : Наука, 1984 .— 472 с.

2. Саттон Р.С., Барто Э.Г. Обучение с подкреплением. — БИНОМ, 2011. (с сайта библиотеки МФТИ).

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

1. http://www.machinelearning.ru – профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

2. http://shad.yandex.ru – сайт школы анализа данных Яндекса.

3. http://www.machinelearning.ru/wiki/index.php?title=%D0%9C%D0%B0%D1%88%D0%B8%D0%BD%D0%BD%D0%BE%D0%B5_%D0%BE%D0%B1%D1%83%D1%87%D0%B5%D0%BD%D0%B8%D0%B5_%28%D0%BA%D1%83%D1%80%D1%81_%D0%BB%D0%B5%D0%BA%D1%86%D0%B8%D0%B9%2C_%D0%9A.%D0%92.%D0%92%D0%BE%D1%80%D0%BE%D0%BD%D1%86%D0%BE%D0%B2%29

4. https://b-ok.cc/book/4988060/2077e6
5. http://bookre.org/reader?file=728687&pg=1
6. https://fileskachat.com/download/60768_01627121f7737502321aceda021ad5f9.html
7. http://bookre.org/reader?file=579056
8. http://bookre.org/reader?file=437034
9. http://bookre.org/reader?file=448640
10. http://bookre.org/reader?file=727780
11. http://bookre.org/reader?file=788986
12. http://bookre.org/reader?file=445105
13. http://bookre.org/reader?file=445920
14. http://bookre.org/reader?file=637236

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

В процессе самостоятельной работы обучающихся предполагается использование таких программных средств, как WEKA, IPython Notebook и др.

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

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:

– проработку учебного материала (по конспектам лекций, учебной и научной литературе);

– подготовку к практическим занятиям, выполнение домашних теоретических и практических заданий.

ПРИЛОЖЕНИЕ

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника
профиль подготовки: Компьютерные технологии и вычислительная техника
Физтех-школа Радиотехники и Компьютерных Технологий
кафедра машинного обучения и цифровой гуманитаристики
курс: 4
квалификация: бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Разработчик: К.В. Воронцов, д-р физ.-мат. наук, профессор

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-2.3 Знает основные требования информационной безопасности
ПК-1 Способен ставить, формализовывать и решать задачи, в том числе разрабатывать и исследовать математические модели изучаемых явлений и процессов, системно анализировать научные проблемы, получать новые научные результаты	ПК-1.1 Способен находить, анализировать и обобщать информацию об актуальных результатах исследований в рамках тематической области своей профессиональной деятельности
	ПК-1.2 Способен выдвигать гипотезы, строить математические модели для описания изучаемых явлений и процессов, оценивать качество разработанной модели

2. Показатели оценивания компетенций

В результате изучения дисциплины «Введение в машинное обучение» обучающийся должен:

знать:

- основные принципы и проблематику теории обучения машин,
- основные современные методы обучении по прецедентам — классификации, кластеризации и регрессии.

уметь:

- формализовать постановки прикладных задач анализа данных;
- использовать методы обучения по прецедентам для решения практических задач;
- оценивать точность и эффективность полученных решений.

владеть:

- основными понятиями теории машинного обучения.

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 прошлой лекции или в конце занятия по пройденной теме.

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

1. В чём особенности внутренних и внешних критериев?
2. Разновидности внешних критериев.
3. Разновидности критерия скользящего контроля.
4. Что такое критерий непротиворечивости? В чём его недостатки?
5. Основная идея отбора признаков методом добавлений и исключений.
6. Основная идея отбора признаков методом поиска в глубину.
7. Основная идея отбора признаков методом поиска в ширину.
8. Что такое МГУА?
9. Основная идея отбора признаков с помощью генетического алгоритма.
10. Основная идея отбора признаков с помощью случайного поиска.
11. Приведите пример выборки, которую невозможно классифицировать без ошибок с помощью линейного алгоритма классификации. Какова минимальная длина выборки, обладающая данным свойством? Какие существуют способы модифицировать линейный алгоритм так, чтобы данная выборка стала линейно разделимой?
12. Почему любая булева функция представима в виде нейронной сети? Сколько в ней слоёв?

13. Метод обратного распространения ошибок. Основная идея. Основные недостатки и способы их устранения.

14. Как можно выбирать начальное приближение в градиентных методах настройки нейронных сетей?

15. Как можно ускорить сходимость в градиентных методах настройки нейронных сетей?

16. Как выбирать число слоёв в градиентных методах настройки нейронных сетей?

17. Как выбирать число нейронов скрытого слоя в градиентных методах настройки нейронных сетей?

18. В чём заключается метод оптимального прореживания нейронной сети? Какие недостатки стандартного алгоритма обратного распространения ошибок позволяет устранить метод ODB?

19. Что такое графические модели?

20. Чем отличаются частотный и байесовский подход к теории вероятности?

21. Что такое байесовские сети? В чем их отличие от марковских полей?

22. Когда переменные называются условно независимыми?

23. Как устроена сверточная нейронная сеть?

24. Какие архитектуры глубоких нейронных сетей Вы знаете?

25. Дать определение алгоритмической композиции (помнить формулу). Какие типы корректирующих операций вы знаете?

26. Какие типы голосования вы знаете? Какой из них наиболее общий? (помнить формулу)

27. Как обнаружить объекты-выбросы при построении композиции классификаторов для голосования по большинству?

28. Как обеспечивается различность базовых алгоритмов при голосовании по большинству?

29. Как обеспечивается различность базовых алгоритмов при голосовании по старшинству?

30. Какие две эвристики лежат в основе алгоритма AdaBoost?

31. Как обнаружить объекты-выбросы в алгоритме AdaBoost?

32. Достоинства и недостатки алгоритма AdaBoost.

33. Основная идея алгоритма AnyBoost.

34. Основная идея метода bagging.

35. Основная идея метода случайных подпространств.

36. Приведите примеры выпуклых функций потерь. Почему свойство выпуклости помогает строить смеси экспертов?

37. Каковы основные цели кластеризации?

38. Основные типы кластерных структур. Приведите для каждой из этих структур пример алгоритма кластеризации, который для неё НЕ подходит.

39. В чём заключается алгоритм кратчайшего незамкнутого пути? Как его использовать для кластеризации? Как с его помощью определить число кластеров? Всегда ли это возможно?

40. Основная идея алгоритма ФорЭл.

41. Как вычисляются центры кластеров в алгоритме ФорЭл, если объекты — элементы метрического (не обязательно линейного векторного) пространства?

42. Какие существуют функционалы качества кластеризации и для чего они применяются?

43. Основные отличия алгоритма k-средних и EM-алгоритма. Кто из них лучше и почему?

44. Основная идея иерархического алгоритма Ланса-Вильямса.

45. Какие основные типы расстояний между кластерами применяются в алгоритме Ланса-Вильямса?

46. Какие расстояния между кластерами, применяемые в алгоритме Ланса-Вильямса, лучше и почему?

47. Что такое дендрограмма? Всегда ли её можно построить?

48. Какой функционал качества оптимизируется сетью Кохонена? (помнить формулу)

49. Как устроена самооганизующаяся карта Кохоненеа?

50. Как интерпретируются карты Кохонена?

51. Почему задачи с частичным обучением выделены в отдельный класс? Приведите примеры, когда методы классификации и кластеризации дают неадекватное решение задачи с частичным обучением.

52. Как приспособить графовые алгоритмы кластеризации для решения задачи с частичным обучением?

53. Как приспособить EM-алгоритм для решения задачи с частичным обучением?

54. Какие способы решения задачи с частичным обучением Вы знаете?

55. Постановка задачи обучения ранжированию.

56. Какие способы решения задач ранжирования Вы знаете?

57. Какие функционалы качества ранжирования Вы знаете?

58. Постановка задачи коллаборативной фильтрации.

59. В чем особенности корреляционных методов решения задачи коллаборативной фильтрации? Какие корреляционные методы Вы знаете?

60. В чем особенности латентных методов решения задачи коллаборативной фильтрации? Какие латентные методы Вы знаете?

61. Постановка задачи тематического моделирования коллекции текстовых документов.

62. Какие подходы к тематическому моделированию Вы знаете?

63. Какие подходы к регуляризации тематических моделей Вы знаете?

Пример вопросов билета:

1. Основная идея алгоритма AnyBoost.
2. Какие архитектуры глубоких нейронных сетей Вы знаете?
3. Чему равна вероятность того, что объект x попадет в очередную Bootstrap-выборку, если выбор элементов происходит без возвращения?

Критерии оценивания

Оценка «зачёт» - выставляется студенту, показавшему, как минимум, фрагментарный, разрозненный характер знаний базовых понятий и программного материала, который хотя бы слабо владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и применяет полученные знания , как минимум, в стандартной ситуации.

Оценка «незачёт» - 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.

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

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