

**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: Прикладные модели машинного обучения

программа аспирантуры: Биологические науки

кафедра машинного обучения и цифровой гуманитаристики

курс: 1

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 48 час.

Всего часов: 78, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

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

Программа обсуждена на заседании кафедры машинного обучения и цифровой гуманитаристики 22.05.2023

Аннотация

В курсе приводятся примеры прикладных задач классификации, регрессии, ранжирования. Рассматриваются методы прогнозирования временных рядов, поиск ассоциативных правил, нейронные сети глубокого обучения, эвристические, стохастические, нелинейные композиции.

Также кратко обсуждаются некоторые вопросы методологии машинного обучения: особенности реальных данных, межотраслевой стандарт CRISP-DM, организация вычислительных экспериментов. Выделяются многочисленные сходства и взаимосвязи между различными методами машинного обучения, обсуждаются идеи разнообразных гибридных подходов.

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.

Задачи дисциплины

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

2. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

- основные принципы и проблематику теории обучения машин;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов;
- 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;
- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- навыками теоретического анализа реальных задач, решаемых с помощью алгоритмов обучения по прецедентам.

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

3.1. 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 | Прогнозирование временных рядов | 4 | | | 6 |
| 2 | Поиск ассоциативных правил | 2 | | | 4 |
| 3 | Нейронные сети глубокого обучения | 4 | | | 6 |
| 4 | Эвристические, стохастические, нелинейные композиции. | 4 | | | 6 |
| 5 | Ранжирование | 2 | | | 4 |
| 6 | Рекомендательные системы | 2 | | | 4 |

| | | | | |
|-----------------------|----------------------------|--------------------|--|----|
| 7 | Тематическое моделирование | 4 | | 6 |
| 8 | Обучение с подкреплением | 4 | | 6 |
| 9 | Активное обучение | 4 | | 6 |
| Итого часов | | 30 | | 48 |
| Подготовка к экзамену | | 0 час. | | |
| Общая трудоёмкость | | 78 час., 2 зач.ед. | | |

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

Семестр: 2 (Весенний)

1. Прогнозирование временных рядов

- Задача прогнозирования временных рядов. Примеры приложений.
- Экспоненциальное скользящее среднее. Модель Хольта. Модель Тейла-Вейджа. Модель Хольта-Уинтерса.
- Адаптивная авторегрессионная модель.
- Следящий контрольный сигнал. Модель Тригга-Лича.
- Адаптивная селективная модель. Адаптивная композиция моделей.
- Локальная адаптация весов с регуляризацией.

2. Поиск ассоциативных правил

- Понятие ассоциативного правила и его связь с понятием логической закономерности.
- Примеры прикладных задач: анализ рыночных корзин, выделение терминов и тематики текстов.
- Алгоритм APriori. Два этапа: поиск частых наборов и рекурсивное порождение ассоциативных правил. Недостатки и пути усовершенствования алгоритма APriori.
- Алгоритм FP-growth. Понятия FP-дерева и условного FP-дерева. Два этапа поиска частых наборов в FP-growth: построение FP-дерева и рекурсивное порождение частых наборов.
- Общее представление о динамических и иерархических методах поиска ассоциативных правил.

3. Нейронные сети глубокого обучения

- Свёрточные нейронные сети (CNN). Свёрточный нейрон. Pooling нейрон. Выборка размеченных изображений ImageNet.
- Свёрточные сети для сигналов, текстов, графов, игр
- Рекуррентные нейронные сети (RNN). Обучение рекуррентных сетей: Backpropagation Through Time (BPTT).
- Сети долгой кратковременной памяти (Long short-term memory, LSTM).
- Рекуррентная сеть Gated Recurrent Unit (GRU).
- Автокодировщики. Векторные представления дискретных данных.
- Перенос обучения (transfer learning).
- Самообучение (self-supervised learning).
- Генеративные состязательные сети (GAN, generative adversarial net).

4. Эвристические, стохастические, нелинейные композиции.

- Стохастические методы: бэггинг и метод случайных подпространств
- Простое голосование (комитет большинства). Алгоритм ComBoost. Идентификация нетипичных объектов (выбросов).

- Преобразование простого голосования во взвешенное.
- Обобщение на большое число классов.
- Случайный лес.
- Анализ смещения и вариации для простого голосования.
- Смесь алгоритмов (квазилинейная композиция), область компетентности, примеры функций компетентности.
- Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический.
- Построение смеси алгоритмов с помощью EM-подобного алгоритма.

5. Ранжирование

- Постановка задачи обучения ранжированию. Примеры.
- Признаки в задаче ранжирования поисковой выдачи: текстовые, ссылочные, кликовые. TF-IDF. PageRank.
- Критерии качества ранжирования: Precision, MAP, AUC, DCG, NDCG, pFound.
- Ранговая классификация, OC-SVM.
- Попарный подход: RankingSVM, RankNet, LambdaRank.

6. Рекомендательные системы

- Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты - объекты.
- Корреляционные методы user-based, item-based. Задача восстановления пропущенных значений. Меры сходства субъектов и объектов.
- Латентные методы на основе би-кластеризации. Алгоритм Брегмана.
- Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных (LFM, Latent Factor Model). Метод стохастического градиента.
- Неотрицательные матричные разложения. Метод чередующихся наименьших квадратов ALS.
- Модель с учётом неявной информации (implicit feedback).
- Рекомендации с учётом дополнительных признаков данных. Линейная и квадратичная регрессионные модели, libFM.
- Измерение качества рекомендаций. Меры разнообразия (diversity), новизны (novelty), покрытия (coverage), догадливости (serendipity).

7. Тематическое моделирование

- Задача тематического моделирования коллекции текстовых документов.
- Вероятностный латентный семантический анализ PLSA. Метод максимума правдоподобия. EM-алгоритм. Элементарная интерпретация EM-алгоритма.
- Латентное размещение Дирихле LDA. Метод максимума апостериорной вероятности. Сглаженная частотная оценка условной вероятности.
- Небайесовская интерпретация LDA и её преимущества. Регуляризаторы разреживания, сглаживания, частичного обучения.
- Аддитивная регуляризация тематических моделей. Регуляризованный EM-алгоритм, теорема о стационарной точке (применение условий Каруша–Куна–Таккера).
- Рациональный EM-алгоритм. Онлайн-алгоритм и его распараллеливание.
- Мультимодальная тематическая модель.
- Регуляризаторы классификации и регрессии.
- Регуляризаторы декоррелирования и отбора тем.
- Внутренние и внешние критерии качества тематических моделей.

8. Обучение с подкреплением

- Задача о многоруком бандите. Жадные и эпсилон-жадные стратегии. Метод UCB (upper confidence bound). Стратегия Softmax.
- Среда для экспериментов.

- Адаптивные стратегии на основе скользящих средних. Метод сравнения с подкреплением. Метод преследования.
- Постановка задачи в случае, когда агент влияет на среду. Ценность состояния среды. Ценность действия.
- Жадные стратегии максимизации ценности. Уравнения оптимальности Беллмана.
- Метод временных разностей TD. Метод Q-обучения.
- Градиентная оптимизация стратегии (policy gradient). Связь с максимизацией log-правдоподобия.
- Постановка задачи при наличии информации о среде в случае выбора действия. Контекстный многоармный бандит.
- Линейная регрессионная модель с верхней доверительной оценкой LinUCB.
- Оценивание новой стратегии по большим историческим данным.

9. Активное обучение

- Постановка задачи машинного обучения. Основные стратегии: отбор объектов из выборки и из потока, синтез объектов.
- Сэмплирование по неуверенности. Почему активное обучение быстрее пассивного.
- Сэмплирование по несогласию в комитете. Сокращение пространства решений.
- Сэмплирование по ожидаемому изменению модели.
- Сэмплирование по ожидаемому сокращению ошибки.
- Синтез объектов по критерию сокращения дисперсии.
- Взвешивание по плотности.
- Оценивание качества активного обучения.
- Введение изучающих действий в стратегию активного обучения. Алгоритмы ϵ -active и EG-active.
- Применение обучения с подкреплением для активного обучения. Активное томпсоновское сэмплирование.

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

Учебная аудитория, оснащенная мультимедиапроектором и экраном.

5. Перечень рекомендуемой литературы

Основная литература

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

Дополнительная литература

6. 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%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

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

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

8. 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;
- подготовку к дифференцированному зачету.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

программа аспирантуры: Биологические науки
Физтех-школа Биологической и Медицинской Физики
кафедра машинного обучения и цифровой гуманитаристики

курс: 1

Семестр, формы промежуточной аттестации: 2 (весенний) - Дифференцированный зачет

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

1. Показатели оценивания компетенций

В результате изучения дисциплины «Прикладные модели машинного обучения» обучающийся должен:

знать:

- основные принципы и проблематику теории обучения машин;
- основные методы и алгоритмы решения задач обучения по прецедентам;
- основные области применения этих методов и алгоритмов;
- 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;
- навыками самостоятельной работы при решении типовых задач;
- культурой постановки и моделирования практически значимых задач;
- навыками теоретического анализа реальных задач, решаемых с помощью алгоритмов обучения по прецедентам.

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

1. Задача прогнозирования временных рядов. Примеры приложений.
2. Адаптивная авторегрессионная модель.
3. Адаптивная селективная модель. Адаптивная композиция моделей.
4. Понятие ассоциативного правила и его связь с понятием логической закономерности.
5. Примеры прикладных задач: анализ рыночных корзин, выделение терминов и тематики текстов.
6. Алгоритм APriori. Два этапа: поиск частых наборов и рекурсивное порождение ассоциативных правил. Недостатки и пути усовершенствования алгоритма APriori.
7. Общее представление о динамических и иерархических методах поиска ассоциативных правил.
8. Свёрточные сети для сигналов, текстов, графов, игр.
9. Автокодировщики. Векторные представления дискретных данных.
10. Перенос обучения (transfer learning).
11. Самообучение (self-supervised learning).
12. Стохастические методы: бэггинг и метод случайных подпространств.
13. Преобразование простого голосования во взвешенное.
14. Обобщение на большое число классов.
15. Случайный лес.
16. Выпуклые функции потерь. Методы построения смесей: последовательный и иерархический.
17. Построение смеси алгоритмов с помощью ЕМ-подобного алгоритма.
18. Постановка задачи обучения ранжированию. Примеры.
19. Мультимодальная тематическая модель.
20. Регуляризаторы классификации и регрессии.
21. Регуляризаторы декоррелирования и отбора тем.
22. Внутренние и внешние критерии качества тематических моделей.
23. Задача о многоруком бандите. Жадные и эпсилон-жадные стратегии. Метод UCB (upper confidence bound). Стратегия Softmax.
24. Среда для экспериментов.

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

1. Экспоненциальное скользящее среднее. Модель Хольта. Модель Тейла-Вейджа. Модель Хольта-Уинтерса.
2. Следящий контрольный сигнал. Модель Тригга-Лича.
3. Локальная адаптация весов с регуляризацией.
4. Алгоритм FP-growth. Понятия FP-дерева и условного FP-дерева. Два этапа поиска частых наборов в FP-growth: построение FP-дерева и рекурсивное порождение частых наборов.
5. Свёрточные нейронные сети (CNN). Свёрточный нейрон. Pooling нейрон. Выборка размеченных изображений ImageNet.
6. Рекуррентные нейронные сети (RNN). Обучение рекуррентных сетей: Backpropagation Through Time (BPTT).
7. Сети долгой кратковременной памяти (Long short-term memory, LSTM).
8. Рекуррентная сеть Gated Recurrent Unit (GRU).
9. Генеративные состязательные сети (GAN, generative adversarial net).
10. Простое голосование (комитет большинства). Алгоритм ComBoost. Идентификация нетипичных объектов (выбросов).
11. Анализ смещения и вариации для простого голосования.
12. Смесь алгоритмов (квазилинейная композиция), область компетентности, примеры функций компетентности.
13. Признаки в задаче ранжирования поисковой выдачи: текстовые, ссылочные, кликовые. TF-IDF. PageRank.
14. Критерии качества ранжирования: Precision, MAP, AUC, DCG, NDCG, pFound.
15. Ранговая классификация, OC-SVM.
16. Попарный подход: RankingSVM, RankNet, LambdaRank.
17. Задачи коллаборативной фильтрации, транзакционные данные и матрица субъекты - объекты.
18. Корреляционные методы user-based, item-based. Задача восстановления пропущенных значений. Меры сходства субъектов и объектов.
19. Латентные методы на основе би-кластеризации. Алгоритм Брегмана.
20. Латентные методы на основе матричных разложений. Метод главных компонент для разреженных данных (LFM, Latent Factor Model). Метод стохастического градиента.
21. Неотрицательные матричные разложения. Метод чередующихся наименьших квадратов ALS.
22. Модель с учётом неявной информации (implicit feedback).
23. Рекомендации с учётом дополнительных признаков данных. Линейная и квадратичная регрессионные модели, libFM.
24. Измерение качества рекомендаций. Меры разнообразия (diversity), новизны (novelty), покрытия (coverage), догадливости (serendipity).
25. Задача тематического моделирования коллекции текстовых документов.
26. Вероятностный латентный семантический анализ PLSA. Метод максимума правдоподобия. EM-алгоритм. Элементарная интерпретация EM-алгоритма.
27. Латентное размещение Дирихле LDA. Метод максимума апостериорной вероятности. Сглаженная частотная оценка условной вероятности.
28. Небайесовская интерпретация LDA и её преимущества. Регуляризаторы разреживания, сглаживания, частичного обучения.
29. Аддитивная регуляризация тематических моделей. Регуляризованный EM-алгоритм, теорема о стационарной точке (применение условий Каруша–Куна–Таккера).
30. Рациональный EM-алгоритм. Онлайн-версия EM-алгоритма и его распараллеливание.
31. Адаптивные стратегии на основе скользящих средних. Метод сравнения с подкреплением. Метод преследования.
32. Постановка задачи в случае, когда агент влияет на среду. Ценность состояния среды. Ценность действия.
33. Жадные стратегии максимизации ценности. Уравнения оптимальности Беллмана.
34. Метод временных разностей TD. Метод Q-обучения.
35. Градиентная оптимизация стратегии (policy gradient). Связь с максимизацией log-правдоподобия.
36. Постановка задачи при наличии информации о среде в случае выбора действия. Контекстный многорукий бандит.

37. Линейная регрессионная модель с верхней доверительной оценкой LinUCB.
38. Оценивание новой стратегии по большим историческим данным.
39. Постановка задачи машинного обучения. Основные стратегии: отбор объектов из выборки и из потока, синтез объектов.
40. Сэмплирование по неувренности. Почему активное обучение быстрее пассивного.
41. Сэмплирование по несогласию в комитете. Сокращение пространства решений.
42. Сэмплирование по ожидаемому изменению модели.
43. Сэмплирование по ожидаемому сокращению ошибки.
44. Синтез объектов по критерию сокращения дисперсии.
45. Взвешивание по плотности.
46. Оценивание качества активного обучения.
47. Введение изучающих действий в стратегию активного обучения. Алгоритмы ϵ -active и EG-active.
48. Применение обучения с подкреплением для активного обучения. Активное томпсоновское сэмплирование.

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.

4. 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 проводится путем организации специального опроса, проводимого в устной форме.