

**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 |
| по направлению: | Информатика и вычислительная техника |
| профиль подготовки: | Физтех-школа Прикладной Математики и Информатики кафедра банковских информационных технологий |
| курс: | 4 |
| квалификация: | бакалавр |

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

Аудиторных часов: 45 всего, в том числе:

лекции: 45 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: В.В. Кантор, преподаватель

Программа обсуждена на заседании кафедры банковских информационных технологий 12.06.2020

Аннотация

Дисциплина продолжает цикл дисциплин, связанных с основами анализа данных, информационных технологий и программирования.

В курсе изучаются основные методы машинного обучения и их математическое обоснование. Затрагиваются задачи классификации текстов на два и более классов, многоотечной и иерархической классификации. Определяется класс задач АОТ, решаемых с помощью методов классификации последовательностей.

Рассматриваются методы снижения размерности, основанные как на матричных разложениях, так и на обучении векторам слов. Также разбираются методы поиска глобального минимума на поверхности как подзадача машинного обучения. Теоретический материал курса подкрепляется практическими занятиями по использованию популярных инструментов по изучаемой тематике.

1. Цели и задачи

Цель дисциплины

Систематизировать и углубить знания студентов в области методов машинного обучения и анализа данных, а также развить понимание связи их теоретических основ с решением практических задач.

Задачи дисциплины

1. Создать понимание задач машинного обучения, мотивации к их решению и практических приложений этих задач.
2. Познакомить с теоретической основой методов, используемых для решения этих задач.
3. Выработать у студентов базовые практические навыки постановки и решения задач машинного обучения.
4. Довести до сведения студентов актуальные задачи и некоторые последние достижения в области машинного обучения.

2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|---|
| ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач | ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности |
| ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре) | ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры |

3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

Формулировки классических задач анализа данных и машинного обучения и теоретические основы методов их решения.

уметь:

Решать задачи машинного обучения и видеть их в возникающих в профессиональной деятельности ситуациях.

владеть:

Навыками сведения практической задачи к стандартным задачам машинного обучения и реализации пригодного к применению решения.

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 | Введение | 8 | | | 8 |
| 2 | Алгоритмы машинного обучения | 8 | | | 8 |
| 3 | Работа с признаками | 8 | | | 8 |
| 4 | Постановка задачи и оценка качества моделей | 7 | | | 7 |
| 5 | Прикладные задачи | 7 | | | 7 |
| 6 | Краткий обзор последних достижений в области машинного обучения | 7 | | | 7 |
| Итого часов | | 45 | | | 45 |
| 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. Введение

Основные понятия: классификация, регрессия, кластеризация, переобучение, кросс-валидация, learning curves, bias-variance trade-off. Карта курса, анонс заданий.

1. Напоминание простых алгоритмов классификации, регрессии и кластеризации: метод ближайших соседей, центроидный классификатор, K-means.

Библиотека sklearn. Обзор реализованных алгоритмов, документации и интерфейсов.

2. Напоминание статистики и методов оптимизации: оценка параметров распределений, свойства оценок, бутстреп, градиентные методы оптимизации (первого и второго порядка), негладкие и дискретные функции, поиск глобального экстремума.

2. Алгоритмы машинного обучения

1. Линейная классификация и регрессия: функции потерь и регуляризаторы, метод стохастического градиента и другие методы настройки параметров. Онлайн-обучение. Библиотека Vowpal Wabbit. Логистическая регрессия, максимизация энтропии и расстояния Кульбака-Лейблера, экспоненциальное семейство распределений. SVM: условная, безусловная и двойственная задачи, используемые методы оптимизации, ядра, l2-loss и l1-penalized модификации. Semi-supervised SVM и логистическая регрессия.

2. Решающие и регрессионные деревья: общая идея, критерии информативности, ID3, Бинаризация признаков, пост-пруннинг и пре-пруннинг, C4.5 и CART. *Unsupervised decision trees.

3. Байесовские методы классификации и регрессии. Наивный байесовский классификатор. Выбор семейства распределений. Оптимальное байесовское решающее правило. Восстановление плотности распределений.

4. Нейросети: сети прямого распространения, метод обратного распространения ошибки, рекуррентные нейросети, сверточные нейросети, глубокое обучение. Знакомство с библиотеками Theano, Lasagne, Nolearn, keras, kaffa.
5. Композиции алгоритмов: бустинг (адаптивный и градиентный), бэггинг, блендинг, стекинг. Градиентный бустинг над деревьями и случайный лес. Библиотека XGBoost. Ансамбли деревьев в sklearn и R: особенности реализации.
6. Алгоритмы кластеризации: K-means, иерархическая, EM-алгоритм, MeanShift, DBScan, AffinityPropagation
7. Анализ временных рядов: виды тренда и сезонности, простые модели их анализа, ARMA, ARIMA, работа с нестационарными временными рядами
8. Обучение с подкреплением (обзор)
9. Графические модели: марковские поля и байесовские сети. Условные случайные поля. (обзор)
10. Байесовский вывод (обзор)

3. Работа с признаками

1. Извлечение и генерация признаков на примере практических задач: анализ текстов, изображений, звука. Взаимодействия признаков.
2. Отбор признаков: по статистическим критериям, отбор жадными алгоритмами, отбор генетическими алгоритмами.
3. Преобразование признаков: главные компоненты, независимые компоненты, матричные разложения, факторизационные машины, вероятностное тематическое моделирование, автоэнкодеры, обучение представлений, manifold learning

4. Постановка задачи и оценка качества моделей

1. Сведение практических задач к стандартным задачам машинного обучения. Особенности реализации кросс-валидации.
2. Сбор и чистка выборки, выбор задачи с учетом трудностей подготовки обучающей выборки и особенностей реализации.
3. Функционалы качества (log loss, AUC ROC, AUC PRC, accuracy, precision, recall, внутрикластерное и межкластерное расстояние, MAE, RSME, RAE, коэффициент детерминации), их свойства, вероятностный смысл и интерпретируемость. Особенности максимизации различных функционалов качества.
4. Вероятностная интерпретация различных методов построения классификаторов. Общие сведения о структурной минимизации риска и обобщающей способности алгоритмов.

5. Прикладные задачи

1. Бизнес-аналитика: прогнозирование оттока и спроса.
2. Страхование и банковская сфера: кредитный скоринг и детектирование мошенничества.
3. Информационный поиск: PageRank, learning to rank, re-ranking
4. Рекомендательные системы: user-based и item-based подходы, SVD и LDA, графовые методы. Netflix, YouTube.
5. Реклама: прогноз CTR, прогноз вероятностей просмотров, рекомендации рекламных предложений. Многоармные бандиты.
6. Анализ текстов, изображений и видео, звука.

6. Краткий обзор последних достижений в области машинного обучения

Знаковые проекты ML-проектов.

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

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

6.Перечень рекомендуемой литературы

Основная литература

1. Введение в методы машинного обучения с подкреплением, учебное пособие /А. И. Панов; Министерство науки и высшего образования Российской Федерации ; Московский физико-технический институт (национальный исследовательский университет). Москва, МФТИ, 2019

Дополнительная литература

1. Нейронные сети [Текст] : полный курс / С. Хайкин ; пер. с англ. Н. Н. Куссуль, А. Ю. Шелестова ; под ред. Н. Н. Куссуль .— 2-е изд., испр. — М. : Вильямс, 2006 .— 1103 с.

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

<http://statweb.stanford.edu/~tibs/ElemStatLearn/> (п.1 из основного списка литературы)
<http://www.cs.ubc.ca/~murphyk/MLbook/>
<http://scikit-learn.org>

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

На занятиях используются мультимедийные технологии, включая демонстрацию презентаций.

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

Методические рекомендации позволяют студенту оптимальным образом организовать процесс обучения. В структуре учебного плана значительное время отводится на самостоятельное изучение данной дисциплины. В рабочей программе приведено примерное распределение часов аудиторной и внеаудиторной нагрузки по различным темам данной дисциплины.

Для успешного освоения данной дисциплины студенту необходимо:

- посещать занятия;
- выполнять задания;
- сдать дифференцированный зачет по дисциплине.

ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

по направлению: Информатика и вычислительная техника

профиль подготовки: Физтех-школа Прикладной Математики и Информатики
кафедра банковских информационных технологий

курс: 4

квалификация: бакалавр

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

Разработчик: В.В. Кантор, преподаватель

1. Компетенции, формируемые в процессе изучения дисциплины

| Код и наименование компетенции | Индикаторы достижения компетенции |
|--|---|
| ОПК-4 Способен осуществлять сбор и обработку научно-технической и (или) технологической информации для решения фундаментальных и прикладных задач | ОПК-4.2 Знает основные источники научно-технической и (или) технологической информации в области профессиональной деятельности |
| ОПК-5 Способен участвовать в проведении фундаментальных и прикладных исследований и разработок, самостоятельно осваивать новые теоретические, в том числе, математические методы исследований и работать на современной экспериментальной научно-исследовательской, измерительно-аналитической и технологической аппаратуре) | ОПК-5.3 Способен к профессиональной эксплуатации современной экспериментальной научно-исследовательской (измерительно-аналитической и технологической) аппаратуры |

2. Показатели оценивания компетенций

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

знать:

Формулировки классических задач анализа данных и машинного обучения и теоретические основы методов их решения.

уметь:

Решать задачи машинного обучения и видеть их в возникающих в профессиональной деятельности ситуациях.

владеть:

Навыками сведения практической задачи к стандартным задачам машинного обучения и реализации пригодного к применению решения.

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

Перечень вопросов для промежуточного контроля:

1. Обучить классификатор текстов по теме. Обучение должно происходить без загрузки всей выборки в память. В качестве обучающей выборки предлагается датасет в несколько гигабайт. Дополнительное задание: сделать задание с помощью Vowpal Wabbit, настроив веса стохастическим градиентным спуском, а затем дообучив их BFGS.
2. Реализовать простую рекомендательную систему на основе SVD и LDA.
3. Реализовать базовое решение для контекста с Kaggle.com.
4. Имея реализации нескольких алгоритмов, решающих задачу, построить на их основе новый алгоритм, превосходящий каждый из них по качеству.

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

Вопросы по методам машинного обучения и библиотекам:

1. Линейная классификация и регрессия: функции потерь и регуляризаторы, метод стохастического градиента и другие методы настройки параметров. Онлайн обучение. Библиотека Vowpal Wabbit.
2. Логистическая регрессия, максимизация энтропии и расстояния Кульбака-Лейблера, экспоненциальное семейство распределений.
3. SVM: условная, безусловная и двойственная задачи, используемые методы оптимизации, ядра, l2-loss и l1-penalized модификации. Semi-supervised SVM и логистическая регрессия.

4. Решающие и регрессионные деревья: общая идея, критерии информативности, ID3, Бинаризация признаков, пост-прунинг и пре-прунинг, C4.5 и CART.
5. Байесовские методы классификации и регрессии. Наивный байесовский классификатор. Выбор семейства распределений. Оптимальное байесовское решающее правило. Восстановление плотности распределений.
6. Нейросети: сети прямого распространения, метод обратного распространения ошибки, рекуррентные нейросети, сверточные нейросети, глубокое обучение. Библиотеки Theano, Lasagne, Nolearn, keras, kaffa.
7. Композиции алгоритмов: бустинг (адаптивный и градиентный), бэггинг, блендинг, стекинг. Градиентный бустинг над деревьями и случайный лес. Библиотека XGBoost. Ансамбли деревьев в sklearn и R: особенности реализации.
8. Алгоритмы кластеризации: K-means, иерархическая, EM-алгоритм, MeanShift, DBScan, AffinityPropagation
9. Анализ временных рядов: виды тренда и сезонности, простые модели их анализа, ARMA, ARIMA, работа с нестационарными временными рядами
10. Извлечение и генерация признаков на примере практических задач: анализ текстов, изображений, звука. Взаимодействия признаков. Отбор признаков: по статистическим критериям, отбор жадными алгоритмами, отбор генетическими алгоритмами.
11. Преобразование признаков: главные компоненты, независимые компоненты, матричные разложения, факторизационные машины, вероятностное тематическое моделирование, автоэнкодеры, обучение представлений, manifold learning
12. Функционалы качества (log loss, AUC ROC, AUC PRC, accuracy, precision, recall, внутрикластерное и межкластерное расстояние, MAE, RSME, RAE, коэффициент детерминации), их свойства, вероятностный смысл и интерпретируемость. Особенности максимизации различных функционалов качества.

Вопросы по приложениям:

1. Бизнес-аналитика: прогнозирование оттока и спроса.
2. Страхование и банковская сфера: кредитный скоринг и детектирование мошенничества.
3. Информационный поиск: PageRank, learning to rank, re-ranking
4. Рекомендательные системы: user-based и item-based подходы, SVD и LDA, графовые методы. Netflix, YouTube.
5. Реклама: прогноз CTR, прогноз вероятностей просмотров, рекомендации рекламных предложений. Многоармие бандиты.
6. Машинное обучение в анализе текстов, изображений и видео, звука.

Критерии оценивания

Оценка «отлично (10)» выставляется обучающемуся, если он показал 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 обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений, но при этом были допущены небольшие неточности, которые были самостоятельно обнаружены и исправлены;

оценка «отлично (8)» выставляется обучающемуся, если он показал всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений, но при этом были допущены небольшие неточности, которые после указания экзаменатора были самостоятельно исправлены;

оценка «хорошо (7)» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает неточности в ответе или делает несущественные ошибки при решении задач;

оценка «хорошо (6)» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает небольшие ошибки в ответе и (или) при решении задач;

оценка «хорошо (5)» выставляется обучающемуся, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но отвечает неуверенно и (или) допускает ошибки при решении задач;

оценка «удовлетворительно (4)» выставляется обучающемуся, показавшему фрагментарный, разрозненный характер знаний, неточные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, если 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 обучающемуся, показавшему фрагментарный, разрозненный характер знаний, неточные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, не владеющему некоторыми разделами учебной программы, но умеющему применять полученные знания по образцу в стандартной ситуации;

оценка «неудовлетворительно (2)» выставляется обучающемуся, 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 обучающемуся, показавшему полное незнание учебной программы дисциплины.

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

Для сдачи курса необходимо успешно сдать практическую и теоретическую часть дифференцированного зачета (могут сдаваться отдельно). На практической нужно самостоятельно решить задачу на знание методов и библиотек. На теоретической части ответить на два билета по курсу – один из списка вопросов по методам, один из списка вопросов по приложениям.