

**Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»**

УТВЕРЖДЕНО
Проректор по учебной работе

А.А. Воронов

	Рабочая программа дисциплины (модуля)
по дисциплине:	Приложения машинного обучения
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: 8 (весенний) - Экзамен

Аудиторных часов: 60 всего, в том числе:

лекции: 30 час.

семинары: 30 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Подготовка к экзамену: 30 час.

Всего часов: 135, всего зач. ед.: 3

Количество контрольных работ, заданий: 2

Программу составили:

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

Р.Г. Нейчев, ассистент

Программа обсуждена на заседании кафедры машинного обучения и цифровой гуманитаристики 15.05.2021

Аннотация

Курс посвящён разнообразным приложениям машинного обучения. Вначале рассматриваются стандартные задачи (классификация, регрессия, кластеризация) и простые модели: kNN, naïve bayes, linear regression, линейные ансамбли деревьев и т.д. Отдельно обсуждаются инструменты, с помощью которых эти задачи можно решать: питоновские библиотеки pandas и sklearn.

Много внимания уделяется бустингу и его обобщениям. Линейные модели рассматриваются на примерах работы с признаками из текстов и с one-hot-encoding.

Даётся обзор актуальных архитектур нейронных сетей для решения задач, в частности, компьютерного зрения. Также обсуждаются приложения методов обучения без учителя, рекомендательные системы и работа с данными разной природы.

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

правильно формулировать задачу 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 обучения 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	Введение: основные понятия и простые методы	4	4		7
2	Решающие деревья и ансамбли	6	6		10
3	Линейные модели	6	6		10
4	Нейронные сети	6	6		6
5	Обучение без учителя	4	4		6
6	Обзор приложений машинного обучения	4	4		6
Итого часов		30	30		45
Подготовка к экзамену		30 час.			
Общая трудоёмкость		135 час., 3 зач.ед.			

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

Семестр: 8 (Весенний)

1. Введение: основные понятия и простые методы

Основные понятия. Примеры использования машинного обучения. Ключевые понятия. Supervised и unsupervised learning. Стандартные задачи (классификация, регрессия, кластеризация). Простые модели (kNN, naïve bayes, linear regression), кратко о тех моделях, которые часто используются на практике - линейные и ансамбли деревьев (основная идея). Оценка качества - кросс-валидация, кривые обучения, переобучение и недообучение, как детектировать, истории из практики. Как возникают и как решаются оптимизационные задачи в машинном обучении. Немного об инструментах: Python, numpy, scipy, matplotlib

Метрики, признаки и инструменты. Метрики качества в стандартных задачах. Извлечение признаков (на примере текста, изображений, звука) и предобработка признаков (на примере работы с разреженными и категориальными признаками). Разбор примеров задач: с обсуждением метрик качества, способов оценки качества, необходимых данных и извлекаемых признаков. Инструменты, с помощью которых эти задачи можно решать: питоновские библиотеки pandas и sklearn. Демонстрация: pandas, sklearn: datasets, metrics, cross_validation, trees.

2. Решающие деревья и ансамбли

Решающие деревья

- как работает уже построенное решающее дерево;
- задача классификации и регрессии;
- рекурсивное построение деревьев:

- критерии информативности, information gain - misclassification, энтропийный критерий, индекс Gini;
- дискретизация / бинаризация признаков, работа с категориальными признаками;
- работа с пропущенными значениями;
- стрижка деревьев (pruning);
- преимущества и недостатки деревьев;
- оценка важности признаков;
- технические заметки (ID3, C4.5, C5.0, CART)

Ансамбли решающих деревьев

- Bias-Variance Trade-off
 - Бэггинг (Bagging = Bootstrap Aggregation), связь корреляция между ответами моделей и качеством модели в бэггинге.
 - Улучшения бэггинга: RSM, Pasting, случайный лес (Random Forest), Extremely Randomized Trees (превращение неустойчивости деревьев из недостатка в преимущество)
 - Бустинг (Boosting), AdaBoost и обобщения
 - Stacking и Blending
- Boosting, state-of-the-art алгоритмы
- тонкости реализации boosting
 - обобщение до Gradient Tree Boosting / GBDT / GBM / MART
 - эвристики оптимизации и state-of-the-art алгоритмы (xgboost, lightgbm, ...)

3. Линейные модели

Линейные модели. Идея линейной классификации. Настройка параметров линейного классификатора: функции потерь, оптимизационные задачи. Gradient Descent и Stochastic Gradient Descent. Регуляризация: l1, l2, elastic net. Стандартные линейные классификаторы. Линейная регрессия: выражение для вычисления весов, регуляризация (гребневая регрессия и лассо). Примеры применения линейных моделей: работа с признаками из текстов и с one-hot-encoding (заодно упомянуть про hashing trick). Библиотеки для построения линейных моделей: sklearn.linear_model, liblinear, vowpal wabbit.

Логистическая регрессия и SVM. Логистическая функция потерь, как к ней можно прийти (из требований к виду функции и из желания оценивать величины от 0 до 1, похожие на вероятности). Log loss. Максимизация ширины разделяющей полосы, оптимизационная задача в SVM для задачи классификации. Безусловная оптимизационная задача. Двойственная задача с выводом. Kernel trick. Радиальное ядро (RBF).

Дополнительные темы

SVM для регрессии. Мультиклассовые SVM и логистическая регрессия. Примеры использования. Одноклассовый SVM. Примеры использования. (Опционально) Semi-supervised модификации линейных моделей (S3VM, entropy regularizer).

4. Нейронные сети

Нейронные сети как суперпозиция моделей. Исторический экскурс.

Математическая модель нейрона, проблема XOR.

Механизм обратного распространения ошибки (backpropagation). Идея и математика обучения нейронных сетей.

Механизмы оптимизации. Стохастический градиент и его вариации (adagrad, momentum, nesterov momentum, adadelta, rmsprop, adam).

Обзор слоев и функций активации в нейронных сетях (полносвязный, сверточный, dropout, batchnorm etc.)

Проблема переобучения, регуляризация нейронных сетей.

Сверточные нейронные сети для задачи анализа изображений: принцип работы, методы обучения.

Обзор актуальных архитектур нейронных сетей для решения задач компьютерного зрения (Computer Vision).

Рекуррентные нейронные сети.

Обзор классической RNN-cell, LSTM, GRU.

Рекуррентные нейронные сети в задаче анализа сигналов и естественного языка.

Генеративные модели на основе RNN.

Механизм внимания (Attention mechanism) в задаче машинного перевода и других задачах.

Сверточные нейронные сети в задачах обработки текста, сравнение с рекуррентными нейронными сетями.

5. Обучение без учителя

Преобразование признаков

Dimensionality Reduction: PCA, SVD, t-SNE, MDS

Embedding Manifold (overview)

Latent Models: LDA

Задача кластеризации

1. Статистические алгоритмы: EM, k-means+ dbscan+?

2. Графовые алгоритмы кластеризации, выделение связных компонент. (Алгоритм FOREL)

3. Агломеративная кластеризация, Алгоритм Ланса-Вильямса, построение дендрограммы. Определение числа кластеров.

Свойства сжатия/растяжения, монотонности и редуктивности.

Дополнительные темы

1. Самоорганизующаяся карта Кохонена

другие подходы к визуализации

2. RBM

3. Автоэнкодеры

6. Обзор приложений машинного обучения

Рекомендательные системы

Работа с текстами и тематическое моделирование

Работа с изображениями

Работа с данными в индустрии

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.

Дополнительная литература

1. 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.

2. Прикладная статистика. Принципы и примеры [Текст] : [учеб. пособие для вузов] / Д. Кокс, Э. Снелл ; пер. с англ. Е. В. Чепурина ; под ред. Ю. К. Беляева .— М. : Мир, 1984 .— 200 с.

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%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

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.

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: 8 (весенний) - Экзамен

Разработчики:

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

Р.Г. Нейчев, ассистент

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 «Приложения машинного обучения» обучающийся должен:

знать:

основные принципы и проблематику теории обучения машин, основные современные методы обучения 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 прошлого занятия.

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

- Задачи обучения по прецедентам. Supervised, unsupervised и semi-supervised обучение. Понятия переобучения и обобщающей способности. Скользящий контроль (cross-validation).
- Метрические алгоритмы классификации. Обобщённый метрический классификатор, понятие отступа. Метод ближайших соседей (kNN) и его обобщения. Подбор числа k по критерию скользящего контроля. Отбор эталонных объектов. алгоритм СТОЛП. Функция конкурентного сходства (FRiS).
- Построение метрик и отбор признаков. Стандартные метрики. Оценивание качества метрики. Проклятие размерности. Жадный алгоритм отбора признаков.
- Логические закономерности. Статистический критерий информативности I с (φ, X, l) : смысл и способы вычисления. Энтропийный критерий информативности — информационный выигрыш $IGain$ с (φ, X, l) . Многоклассовые варианты критериев. Индекс Gini. Задача перебора конъюнкций. “Градиентный” алгоритм синтеза конъюнкций и его частные случаи: жадный алгоритм, стохастический локальный поиск, стабилизация, редукция.

5. Бинаризация признаков, алгоритм выделения информативных зон. Решающие списки. Решающие деревья: принцип работы. Разбиение пространства объектов на подмножества, выделяемые конъюнкциями терминальных вершин. Алгоритм ID3. Пре-прунинг и пост-прунинг. RandomForest.
6. Линейная классификация. Непрерывные аппроксимации пороговой функции потерь. Метод минимизации аппроксимированного эмпирического риска. SG, SAG. Связь минимизации аппроксимированного эмпирического риска и максимизации совместного правдоподобия данных и модели. Регуляризация (l1, l2, elasticnet). Вероятностный смысл регуляризаторов. Примеры различных функций потерь и классификаторов. Эвристический вывод логистической функции потерь.
7. Метод опорных векторов. Оптимизационная задача с ограничениями в виде неравенств и безусловная. Опорные векторы. Kerneltrick. Оптимизационная задача в S3VM и SVR. SVM и беспризнаковое машинное обучение на примере ядер графов и классификации вершин графа.
8. Задача снижения размерности пространства признаков. Идея метода главных компонент (PCA). Связь PCA и сингулярного разложения матрицы признаков (SVD). Вычисление SVD в пространствах высокой размерности методом стохастического градиента (SG SVD).
9. Многомерная линейная регрессия. Геометрический и аналитический вывод. Регуляризация в задаче регрессии. Непараметрическая регрессия. Формула Надарая-Ватсона. Регрессионные деревья.
10. Байесовская классификация и регрессия. Функционал риска и функционал среднего риска. Оптимальный байесовский классификатор и теорема о минимизации среднего риска. Наивный байесовский классификатор.
11. Восстановление плотности: параметрический и непараметрический подход. Метод парзеновского окна. Параметрический подход на примере нормального дискриминантного анализа. Линейный дискриминант Фишера.
12. Задача прогнозирования временного ряда, примеры задач. Адаптивные алгоритмы прогнозирования: экспоненциальное сглаживание, модели Брауна, Тейла-Вейджа, Хольта-Винтерса. Преимущества и недостатки адаптивных алгоритмов прогнозирования.
13. Модели ARMA, ARIMA, а также регрессионные методы решения задачи прогнозирования временного ряда. Композиции адаптивных алгоритмов: селекция, композиция, LAMP, агрегирующий алгоритм.
14. Задача кластеризации. Агломеративная и дивизионная кластеризация. Алгоритмы k-Means, k-Means++. Кластеризация с помощью EM-алгоритма (без вывода M-шага). Формула Ланса-Уилльямса.

Пример экзаменационного билета:

1. Решающие списки: принцип работы, схема алгоритма построения по обучающей выборке, стратегии выбора классов при построении. Примеры задач, не решаемых решающими списками.
2. Метод опорных векторов. Оптимизационная задача с ограничениями в виде неравенств и безусловная. Опорные векторы. Kerneltrick. Оптимизационная задача в S3VM и SVR.
3. Логистическая регрессия. Принцип максимума правдоподобия и логарифмическая функция потерь. Метод стохастического градиента в логистической регрессии.

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

При проведении экзамена обучающемуся предоставляется 30 минут на подготовку. Опрос обучающегося по билету не должен превышать двух астрономических часов.

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины.