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

**УТВЕРЖДЕНО**
**Директор физтех-школы электроники, фотоники и молекулярной физики**
**В.В. Иванов**

**Рабочая программа дисциплины (модуля)**

**по дисциплине:** Технологии машинного обучения и искусственного интеллекта для анализа спектров и изображений

**по направлению:** Прикладные математика и физика

**профиль подготовки:** Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы

Физтех-школа Электроники, Фотоники и Молекулярной Физики

кафедра химической физики

**курс:** 1

**квалификация:** магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

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

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Программу составил: О.Е. Родионова, д-р физ.-мат. наук

Программа обсуждена на заседании кафедры химической физики 27.05.2021

**Аннотация**

Курс "Технологии машинного обучения и искусственного интеллекта для анализа спектров и изображений" предусматривает приобретение фундаментальных знаний в области технологии машинного обучения, изучение основных методов обработки многомерных данных физического и химического анализа.

## 1. Цели и задачи

### Цель дисциплины

Целью курса является приобретение фундаментальных знаний в области технологии машинного обучения, изучение основных методов обработки многомерных данных физического и химического анализа.

### Задачи дисциплины

Формирование базовых знаний в области, интегрирующей математическое моделирование, анализ данных и планирование физико-химических экспериментов.

Обучение студентов основам проекционных методов анализа многомерных данных и применения их в физико-химическом эксперименте;

Формирование подходов к выбору эффективных математических методов для решения практических задач.

## 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

| Код и наименование компетенции | Индикаторы достижения компетенции |
|---|---|
| УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий | УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними |
| 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 |
| | ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты |

## 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны

знать:

фундаментальные основы проекционных методов.

уметь:

формулировать задачи анализа данных, собирать и подготавливать наборы данных, выбирать эффективные методы обработки, использовать программы моделирования и справочную литературу.

владеть:

навыками построения и валидации моделей, обоснованием и интерпретацией 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 | Введение и основные понятия математической статистики и матричной алгебры | 4 | | | 4 |
| 2 | Метод главных компонент | 4 | | | 4 |
| 3 | Регрессия: МЛР, РГК, ПЛС, примеры | 4 | | | 4 |
| 4 | Классификация. Основные понятия, обзор методов | 2 | | | 2 |
| 5 | Проекционные методы дискриминации – «ПЛС-дискриминации» | 2 | | | 2 |
| 6 | Проекционный метод одноклассовой классификации «SIMCA» | 2 | | | 2 |
| 7 | Введение в методы разрешения многомерных кривых (РМК) | 2 | | | 2 |
| 8 | Разбор реальных примеров решения задач колебательной спектроскопии с использованием проекционных методов | 2 | | | 2 |
| 9 | N-way задачи, обзор методов решения | 2 | | | 2 |
| 10 | Многомерный анализ изображений и гиперспектральных данных | 6 | | | 6 |
| Итого часов | | 30 | | | 30 |
| Подготовка к экзамену | | 30 час. | | | |
| Общая трудоёмкость | | 90 час., 2 зач.ед. | | | |

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

Семестр: 2 (Весенний)

1. Введение и основные понятия математической статистики и матричной алгебры

Интеллектуальный анализ данных, машинное обучение, глубокое обучение, искусственный интеллект (data mining, machine learning, deep learning, artificial intellegence). Общие определения, связь между понятиями, объекты и методы исследования, примеры.

Программные средства для обработки данных, коммерческие пакеты, свободно распространяемые программы для решения специальных задач, приложения в среде Matlab и R.

Надстройка Chemoemtics add-in, руководство по установке (примеры в лекциях 3-9 анализируются с помощью Chemometrics add-in и сопровождающих шаблонов).

Устройство данных, размерности задач. Матричная алгебра. Простейшие операции с матрицами

След, определитель, ранг матрицы. Норма вектора, угол между векторами, линейная зависимость векторов. Обратная и псевдо-обратная матрицы. Собственные векторы и собственные значения, разложение по сингулярным значениям. Подпространство и проекция на подпространство.

Математическая статистика. Основные понятия: мат. ожидание, дисперсия, ковариация и корреляция. Основные распределения: биномиальное, равномерное, нормальное, хи-квадрат

Оценка параметров и свойства оценок.

Линейная регрессия.

2. Метод главных компонент

Введение в проекционные методы. Понятие скрытых переменных. Проекционный подход для решения задач многомерного анализа данных. Примеры, иллюстрирующие многомерный подход. Два подхода к анализу данных, содержательный (физико-химический) и формальный.

Проекционные методы. Идеи, заложенные в проекционном подходе (1) позволяют работать с большими массивами данных; (2) существенно понижают размерность изучаемой системы; (3) анализируют и позволяют выделить латентные структуры в данных; (4) Позволяют отделять содержательную часть от шума.

Классы решаемых задач: (1)анализ структуры данных, только X-матрица; (2) Регрессионные.

Метод главных компонент, основные понятия графики счетов, графики нагрузок, ошибки моделирования

Работа с шаблонами для Chemometrics add-in, разбор примеров для метода главных компонент

3. Регрессия: МЛР, РГК, ПЛС, примеры

Калибровка – инструмент №1 количественного анализа данных, области практического применения калибровок.

Регрессионный анализ: от простого к сложному, одномерная (univariate) калибровка и МЛР, недостатки МЛР, коллинеарность. Факторное пространство и РГК, калибровка в пространстве главных компонент, РГК как оружие против коллинеарности, другие преимущества. Главный недостаток РГК – пространство PC не учитывает переменной Y, что может снижать качество регрессионной модели. ПЛС регрессия – мощная альтернатива РГК.

основная идея ПЛС: факторное пространство строится оптимально с учетом X и Y; ПЛС1 и ПЛС2, алгоритмы, число компонент, валидация модели, интерпретация ПЛС модели, предсказание.

Основные принципы построения «грамотной» калибровки. Выбор метода калибровки, дизайн эксперимента, отбор проб, предварительная обработка данных (pretreatment), шкалирование (scaling).

Примеры 1. «Идеальные данные». Модельные данные спектров для смеси нескольких компонентов в условиях, близких к идеальным (сигнал + нормальный шум). Цель: освоение калибровочных инструментов, сравнение различных методов, выработка навыков построения и интерпретации регрессионной модели для спектроскопических данных и ее использования. Пример 2. Реальные данные

4. Классификация. Основные понятия, обзор методов

Что такое классификация, как определить принадлежность к классу в зависимости от априорных знаний об исходных данных. Различные алгоритмы классификации с обучением (supervised) и без (unsupervised). Распознавание образов — основная процедура классификации. Различие целей и подходов при решении задач дискриминации и задач одноклассовй классификации. Определение принадлежности образца к найденным классам, определение принадлежности образца к найденным классам с возможностью нахождения выбросов, определение принадлежности образца к найденным классам с использованием переменной отклика. Оценка эффективности классификации. Связь между ошибками первого и второго рода и показателями чувствительности, специфичности и эффективности.

5. Проекционные методы дискриминации – «ПЛС-дискриминации»

Бинарная и многоклассовая дискриминация, однозначная (hard) и неоднозначная (soft) дискриминация. Этапы построения моделей: оценка сложности, валидация, способы оценки качества дискриминации в случая бинарной и многоклассовой задач. ПЛС-дискриминация как метод отбора переменных

6. Проекционный метод одноклассовой классификации «SIMCA»

Основные понятия: график расстояний, область принятия решений, график экстремальных образцов. Этапы построения моделей: поиск выбросов, оценка сложности модели, оценка чувствительности и специфичности. Различия между дискриминацией и одноклассовй классификацией.

7. Введение в методы разрешения многомерных кривых (РМК)

Основные понятия, устройство данных, задачи которые можно решать с помощью методов РМК

Постановка задачи, спектральные и концентрационные окна, условия разрешения, введение ограничений, вращательная и масштабная неопределенности.

Различные алгоритмы для решение задач РМК. Прокрустово преобразование. Эволюционный (EFA) и оконный (WFA) факторный анализ. Итерационные методы, чередующиеся наименьшие квадраты (MCR-ALS). Применение MCR-ALS для решения задач калибровки

8. Разбор реальных примеров решения задач колебательной спектроскопии с использованием проекционных методов

Задачи исследовательского анализа, калибровки и классификации, анализ промежуточных результатов, выявление причин неудачных решений и способы исправления ситуации.

Портативные приборы, применение проекционных методов для расширения возможностей приборов

9. N-way задачи, обзор методов решения

Устройство данных. Физические/ химические эксперименты, приводящие к N-way задачам.

Методы решения таких задач, достоинства и недостатки: Unfolding (развертка данных), модели Tucker 3, параллельный факторный анализ (PARAFAC), метод калибровки - N-way ПЛС

10. Многомерный анализ изображений и гиперспектральных данных

Организация данных, разница между пространственными и спектральными переменными. Основные цветовые пространства. Режимы представления изображения, растровая и векторная графика. Основные виды цифрового представления изображения. Первичная обработка изображений, применение основных функции Matlab из Image processing toolbox . Послойная и морфологическая обработка изображений, использование маски.

Различия между гиперспектральным и многомерным анализом изображений. Визуализация данных, связь между пикселями на изображении и спектральными данными; между длинами волн спектра и пикселями на плоскости изображения. Предварительная обработка данных в пространственной и в спектральной модах. Связь между представлением результатов в пространстве главных компонент и исходном пиксельном пространстве. Обзор основных многомерных методов анализа гиперспектральных изображений.

Использование информации (БИК, ИК, КР и др. инструментальных методов) при изучении объектов культурного наследия методами многомерного и гиперспектрального анализа. Выделение и изучение отдельных фрагментов изображения. Анализ состава поперечного среза слоя краски. Анализ и классификация связующих. Послойный анализ изображения.

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

Учебная аудитория, снабженная доской, экраном, проектором.

**6.Перечень рекомендуемой литературы**

Основная литература

1. Дж. Голуб, Ч. Ван Лоу. Матричные вычисления. Пер.с англ. М.: Мир, 1999 г. -548 с.
2. Боровков А.А. Математическая статистика. Оценка параметров. Проверка гипотез. Главная редакция физико-математической литературы издательства "Наука» 1984 г.- 472 с.
3. С.А. Айвазян, В.М. Бухштабер, И.С. Енюков , Л.Д. Мешалкин Прикладная статистика: классификация и снижение размерности. М.: Финансы и статистика 1989 – 608 с.
4. Н. Дрейпер, Г.Смит Г. Прикладной регрессионный анализ в 2-х книгах. Пер. с англ. М.: Финансы и статистика, 1986 – 366 с.
5. К. Эсбенсен Анализ многомерных данных. Избранные главы. Пер. с англ. Черноголовка: ИПХФ РАН 2005- 158 с.
6. А.Л. Померанцев Хемометрика в Excel: учебное пособие, Томск: Издательство ТПУ 2014:- 434 с.

Дополнительная литература

1. R. Brereton, Chemometrica. Data analysys for the laboratory and chemical plant, J. Wiley&Sons, Chichester, 2004 (ISBN0-471-48977-8).
2. A.L. Pomerantsev, O.Ye. Rodionova, "Concept and role of extreme objects in PCA/SIMCA", J. Chemometrics, 28, 429–438 (2014)
3. A.L. Pomerantsev, O.Ye. Rodionova, "On the type II error in SIMCA method", J. Chemometrics, 28, 518-522 (2014).
4. O.Ye. Rodionova, A.V. Titova, A.L. Pomerantsev, "Discriminant analysis is an inappropriate method of authentication", Trends Anal. Chem., 78 (4), 17-22 (2016).
5. A.L. Pomerantsev, O.Ye. Rodionova, "Multiclass partial least squares discriminant analysis: Taking the right way — A critical tutorial", J. Chemometrics, 32(8): e3030 (2018).
6. A. Racz, K. Heberger, R. Rajko, J. Elek, Classification of Hungarian medieval silver coins using x-ray fluorescent spectroscopy and multivariate data analysis, Heritage Science, 1:2, (2013).
7. G. Sciutto, P. Oliveri, S. Prati, E. Catteli, I. Bonacini. R Mazzeo, “A multivariate methodology workflow for the analysis of FTIR chemical mapping applied on historical paint stratigraphies, Int. J. Anal.Chem, ID 4938145 (2017).
8. N.Navas, J. Romero_Pastro, E. Manzano, C. Cardell, Benefits of applying combined diffuse reflectance FTIR spectroscopy and principal component analysis for the study of blue tempera historical painting, Anal.Chim.Acta, 603, 141-143 (2008).
9. S.Laureti, H. Malekmohammadi, M.K. Rizwan, P. Burrascano, S. Sfarra, M. Mostacci, M. Ricci, Looking through paintings by combining hyper-spectral image and pulse-compression thermograpty, Sensors, 19, 4335 (2019)..

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

https://mipt.ru/science/labs/radiophotonics/obuchenie/lazery.php

**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 дисциплину, 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 дисциплины, уметь применять полученные знания для решения различных задач.

Успешное освоение курса требует:

– посещения всех занятий, предусмотренных учебным планом по дисциплине;

– ведения конспекта занятий;

– напряжённой самостоятельной работы студента.

Самостоятельная работа включает в себя:

– чтение рекомендованной литературы;

– проработку учебного материала, подготовку ответов на вопросы, предназначенных для самостоятельного изучения;

– решение задач, предлагаемых студентам на занятиях;

– подготовку к выполнению заданий текущей и промежуточной аттестации.

Показателем владения материалом служит умение без конспекта отвечать на вопросы по темам дисциплины.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к преподавателю.

Возможен промежуточный контроль знаний студентов в виде решения задач в соответствии с тематикой занятий.

**ПРИЛОЖЕНИЕ**

# ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)

**по направлению:** Прикладные математика и физика
**профиль подготовки:** Физика перспективных технологий: альтернативная энергетика, научное программирование и функциональные материалы
Физтех-школа Электроники, Фотоники и Молекулярной Физики
кафедра химической физики
**курс:** 1
**квалификация:** магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

**Разработчик:** О.Е. Родионова, д-р физ.-мат. наук

**1. Компетенции, формируемые в процессе изучения дисциплины**

| Код и наименование компетенции | Индикаторы достижения компетенции |
|---|---|
| УК-1 Способен осуществлять критический анализ проблемных ситуаций на основе системного подхода, вырабатывать стратегию действий | УК-1.1 Анализирует проблемную ситуацию как систему, выявляя ее составляющие и связи между ними |
| 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 |
| | ПК-1.3 Способен применять теоретические и (или) экспериментальные методы исследований к конкретной научной задаче и интерпретировать полученные результаты |

**2. Показатели оценивания компетенций**

В результате изучения дисциплины «Технологии машинного обучения и искусственного интеллекта для анализа спектров и изображений» обучающийся должен:

**знать:**

фундаментальные основы проекционных методов.

**уметь:**

формулировать задачи анализа данных, собирать и подготавливать наборы данных, выбирать эффективные методы обработки, использовать программы моделирования и справочную литературу.

**владеть:**

навыками построения и валидации моделей, обоснованием и интерпретацией 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 прошлой лекции или в конце занятия по пройденной теме.

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

Вопросы к экзамену:

1. 1. Особенности проекционных методов, понятие скрытых переменных, способы представления данных.
2. Метод главных компонент, основные свойства, размерность системы, эффективный и математический ранг системы.
3. Представление данных в пространстве главных компонент, графики счетов и нагрузок.
4. Способы оценки числа главных компонент.
5. Регрессионные задачи, сравнение множественной регрессии, регрессии на главные компоненты и регрессии на латентные переменные.
6. Основные свойства проекционных регрессионных методов, способы оценки числа компонент.
7. Выбросы и экстремальные образцы, учет, поиск удаление.

8. Основные принципы построения «грамотной» калибровки.

9. Различные способы валидации регрессионных моделей, их преимущества и недостатки.

10. Классификация, цель, классификация с обучением и без обучения.

11. Основные различия между дискриминантным анализом и задачами одноклассовой классификации, строгая и нестрогая классификация.

12. Метод SIMCA- основные шаги, характеристика образцов в проекционном пространстве, область принятия решений, выбросы.

13. Метод SIMCA- способы оценки сложности модели, и показатели качества построенной модели.

14. PLS-DA – построение дискриминантной модели при бинарной классификации, способ выбора числа латентных переменных, оценка качества моделирования.

15. Дискриминнатные модели при наличии 3-х и более классов. Методы построения моделей.

16. N-way задачи, устройство данных, физические/ химические эксперименты, приводящие к N-way задачам.

17. N-way задачи, обзор методов решения таких задач.

18. Многомерные методы разрешения кривых, постановка задачи, спектральные и концентрационные окна, условия разрешения.

19. Многомерные методы разрешения кривы, введение ограничений, вращательная и масштабная неопределенность.

20. Многомерные методы разрешения кривых, обзор основных методов.

21. Метод чередующиеся наименьшие квадраты (MCR-ALS).

22. Анализ изображений. Устройство данных, разница между пространственными и спектральными переменными. Основные цветовые пространства, первичная обработка изображений.

23. Визуализация данных, связь между пикселями на изображении и спектральными данными; между длинами волн спектра и пикселями на плоскости изображения.

24. Связь между представлением результатов в пространстве главных компонент и исходном пиксельном пространстве.

Примеры экзаменационных билетов.

Пример 1.

1. Выбросы и экстремальные образцы, учет, поиск удаление.
2. Основные принципы построения «грамотной» калибровки.

Пример 2.

1. Различные способы валидации регрессионных моделей, их преимущества и недостатки.
2. Классификация, цель, классификация с обучением и без обучения.

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 устного экзамена обучающемуся предоставляется 1 час на подготовку. Опрос обучающегося на экзамене не должен превышать одного астрономического часа.