Сколковский институт науки и технологий (Сколтех)

Рабочая программа дисциплины	**Геометрические методы машинного обучения**

Преподаватель	Берштейн Михаил Александрович, д.ф-м.н., доцент

Аннотация

Описание курса

Многие задачи машинного обучения имеют геометрическую природу. Общая цель машинного обучения - извлечь из данных ранее неизвестную информацию, которая отражается в структуре (геометрии) данных. Поэтому понимание формы многомерных данных играет важную роль в современной теории обучения и аналитике данных.

Данные реального мира, полученные из естественных источников, обычно занимают лишь очень небольшую часть «пространства наблюдений» и концентрируются на структурах более низкой размерности, а геометрические методы позволяют обнаружить форму этих структур по заданным данным.

Значительная часть курса относится к наиболее популярной геометрической модели многомерных данных, называемой «моделью многообразия», в соответствии с которой многомерные данные лежат на или вблизи многообразия меньшей размерности. Курс включает также топологический анализ данных, который активно используется в разведочном и интеллектуальном анализе данных и предоставляет набор топологических и геометрических инструментов для анализа многомерных, неполных и зашумленных данных. Необходимые краткие сведения по дифференциальной геометрии и топологии будут даны в курсе.

В курсе будут приведены примеры применения геометрических и топологических методов анализа данных к различным прикладным задачам. Курс рассчитан на студентов-магистров и аспирантов, интересующихся новейшими геометрическими и топологическими методами, лежащими на стыке математики и машинного обучения.

По окончанию курса, слушатели будут знать основные идеи геометрического подхода к анализу данных, владеть современными геометрическими и топологическими методами анализа данных и уметь применять их для решения основных задач машинного обучения, таких как классификация, регрессия, снижение размерности, представление и визуализация данных, кластеризация и другие. Эти знания и умения позволяет им участвовать в реальных проектах по решению сложных прикладных задач анализа данных.

1. Основная информация

Академический уровень курса	Магистратура Аспирантура
Количество кредитов	3

Предварительные требования к курсу / рекомендации

• Математика для обработки данных
• Численная линейная алгебра
• Методы оптимизации
• Машинное обучение

Мы предполагаем, что слушатель свободно владеет реальным анализом (математическим анализом), основами линейной алгебры, функциональным анализом, теорией вероятностей и статистикой, теорией графов и алгоритмами

Тип оценки - дифференцированная

Отображение оценок в процентах

A:	80
B:	70
C:	60
D:	50
E:	49
F:	0

2. Содержание курса

Тема	Краткое содержание	Лекции (час)	Семинары (час)	Лабораторные занятия (час)	Самостоятельная работа (час)
Введение	Геометрические методы в машинном обучении: мотивация, примеры, основные задачи, подходы	1			
Линейный методы в анализе данных	Методы прогнозирования в анализе данных: Анализ основных компонентов (PCA), Независимый компонентный анализ, прогнозирование результатов	3	2		4
Эвристические методы нелинейного анализа данных	Многомерное масштабирование, ядро PCA, Репликативные нейронные сети	3	2		10
Внутренняя размерность набора данных	Математические определения внутренней размерности твердого тела, оценка внутренней размерности по данным	3	2		10

Элементы дифференциальной геометрии (краткие основы)	Кривые, поверхности, касательные пространства, геодезическая линия, кривизна, многообразия, векторные поля на многообразии, риманово многообразие	2			2
Разнообразное обучение	Многообразная модель нелинейных данных. Многообразное обучение: задачи и подходы. Основные алгоритмы многообразного обучения: Локально-линейное встраивание, изометрическое отображение, Лапласиан Собственные карты, стохастический сосед Встраивание, t-SNE, UMAP, логарифмическая карта, изучение римановых многообразий, собственные карты Грассмана и Штифеля. Изучение многообразий в регрессии и оценке плотности	10		6	12
Топологический анализ данных	Элементы топологии (симплициальные комплексы, фильтрация, гомология постоянства), задачи и подходы TDA, топологические особенности в машинном обучении. Обучение	2	2	2	3

3. Результаты обучения

Результаты обучения в Сколтехе указаны в соответствии со структурой результатов обучения в Сколтехе

1. ФУНДАМЕНТАЛЬНЫЕ ЗНАНИЯ
1.1. Знание математики и естественных наук
1.2. Знание прикладной науки и техники, науки, в том числе современные методы и инструменты
2.1. ПОЗНАНИЕ И СПОСОБЫ РАССУЖДЕНИЯ
2.1.1. Аналитическое мышление и решение проблем
2.1.2. Системное мышление
2.1.3. Творческое мышление
2.2. ОТНОШЕНИЕ И ПРОЦЕСС ОБУЧЕНИЯ
2.2.5. Самосознание и стремление к самосовершенствованию, обучению на протяжении всей жизни и воспитанию
3.1. КОММУНИКАЦИЯ В МЕЖДУНАРОДНОЙ СРЕДЕ
3.1.2. Письменная, электронная и графическая коммуникация
3.1.3. Устная презентация и обсуждение
3.1.4. Вопросы, слушание и диалог
3.1.5. Общение на английском языке в научной, деловой и общественной среде
4.1. ПОНИМАНИЕ ГЛОБАЛЬНОГО СОЦИАЛЬНОГО, ЭКОЛОГИЧЕСКОГО И ДЕЛОВОГО КОНТЕКСТА
4.1.3. Понимание технических продуктов, систем и инфраструктуры отрасли
4.2. ДАЛЬНОВИДНОСТЬ — ИЗОБРЕТЕНИЕ НОВЫХ ТЕХНОЛОГИЙ ПОСРЕДСТВОМ ИССЛЕДОВАНИЙ

4.2.1. Процесс исследования — гипотеза, доказательства и защита
4.2.2. Фундаментальные исследования, ведущие к новым научным открытиям
4.2.3. Исследования, направленные на разработку новых технологий
4.3. ВИДЕНИЕ — РАЗРАБОТКА КОНЦЕПЦИИ И ПРОЕКТИРОВАНИЕ УСТОЙЧИВЫХ СИСТЕМ
4.3.2. Определение и формулирование целей и задач
4.3.3. Разработка концепции и архитектуры продуктов и услуг на основе новых технологий и определение их влияния
4.3.4. Дисциплинарный и междисциплинарный дизайн для обеспечения устойчивости, безопасности, эстетики, работоспособности и других целей
4.3.5. Понимание технического контекста и экосистемы продукта или услуги

4. Задания и выставление оценок

Требование к физической посещаемости (% от числа занятий)	90

Тип назначения	Краткое содержание задания	% от итоговой оценки за курс
Домашние задания	Применяйте изученные алгоритмы к конкретным наборам данных	30
Финальный проект	Изучите предложенные статьи, выберите подходящий алгоритм и примените его к конкретному набору данных	35
Финальный экзамен	Подготовьте доклад на заданную тему, используя знания, полученные в ходе курса. Ответьте дополнительно вопросы.	35

5. Критерии оценки

Задание 1 Типа	Домашние задания

Пример задания 1

Примените несколько изученных алгоритмов для оценки внутренней размерности данного набора данных и сравните их результаты

Критерии оценки для задания 1

От студента ожидалось, что он продемонстрирует понимание того, как выбирать подходящие методы и приемы, как правильно использовать их в домашнем задании и как сравнивать полученные решения.
Максимальный уровень: 4

Задание 2 Типа	Финальный проект

Пример задания 2

Топологический анализ данных для диагностики COVID-19 (примените соответствующий метод для извлечения признаков TDA из КТ-изображений с COVID-19 или выделениями и примените метод глубокого обучения для задачи классификации)

Критерии оценки для задания 2

Ожидается, что студент представит презентацию проекта, идеально структурированную, с четким объяснением всех основных частей проекта. Презентация должна быть представлена в уверенной манере по отношению к аудитории. Участие студента в обсуждении и сессии вопросов и ответов должно быть активным и соответствовать цели и результатам проекта. Максимальный уровень: 4

Задание 3 Типа	Финальный экзамен

Пример задания 3

Объясните суть методов локально-линейного вложения, ISOMAP и лапласианских собственных карт и объясните разницу в подходах, лежащих в основе этих методов.

Критерии оценки для задания 3

Ожидается, что студент представит ответы на заданные вопросы с четким объяснением всех существенных частей тем вопросов. Ответы должны быть даны в уверенной манере, чтобы продемонстрировать понимание студентом сути вопросов и знание основных результатов в этой области.
Максимальный уровень: 4

6. Учебники и интернет-ресурсы

Необходимые учебники	ISBN-13 (or ISBN-10)
Burges Christopher J.C. Dimension Reduction: A Guided Tour. Foundations and Trends in Machine Learning, 2(4): 275 – 365, 2010	978-1601983787
John A. Lee Michel Verleysen. Nonlinear Dimensionality Reduction. Springer, New York, NY, 309 pp., 2007.	978-0387393506
Рекомендуемые учебники	
Li M., Chen Z.S., Bo J. Mathematical 4. Li M., Chen Z.S., Bo J. Mathematical Problems in Data Science Theoretical nd Practical Methods. Switzerland: Springer International Publishing, 213 pp., 2015	978-3-319-25125-7
Cox T.F., Cox M.A.A. Multidimensional Scaling. Chapman and Hall, 2001.	9781584880943
Jollie T. Principal Component Analysis. New-York, Springer, 2002. http://cda.psych.uiuc.edu/statistical_learning_course/Jolliffe%20I.%20Principa	9780387954424

l%20Component%20 Analysis%20(2ed.,%20Springer,%202002)(518s)_MVsa_.pdf	
Ma Y, Fu Y, eds. Manifold Learning Theory and Applications. London: CRC Press, 2011. http://www.gbv.de/dms/goettingen/689943164.pdf	9781439871096
Gorban AN, Kegl B, Wunsch D, Zinovyev AY. Principal Manifolds for Data Visualisation and Dimension Reduction. Springer, Berlin - Heidelberg - New York, 2008. http://pca.narod.ru/contentsgkwz.htm	9783540737490
Jost J. Riemannian Geometry and Geometric analysis. 6th edn. Berlin, Heidelberg: Springer-Verlag, 611 pp., 2011.	9783319618593
Hyvärinen A. Independent component analysis: recent advances. Philos Trans A Math Phys Eng Sci. 2012	9780821848159
Dimensionality Reduction in Machine Learning. Editors: Jamal Amani Rad, Snehashish Chakraverty, Kourosh Parand. Elsevier, 2025	9780443328183
Edelsbrunner, H., and Harer, J. Computational Topology – an Introduction. American Mathematical Society, 2010	978-0-8218-4925-5

Веб-ресурсы (ссылки)	Описание
https://github.com/bekemax/GMML2024	Хранилище с материалами для семинаров

7. Оборудование

Программное обеспечение
Python

Оборудование
Jupyter notebooks

8. Дополнительные примечания

Предлагаемый курс 1) имеет четкое академическое содержание и требования к получению зачетных единиц, 2) соответствует результатам обучения по программе, 3) соответствует политике и регламенту Сколтеха.