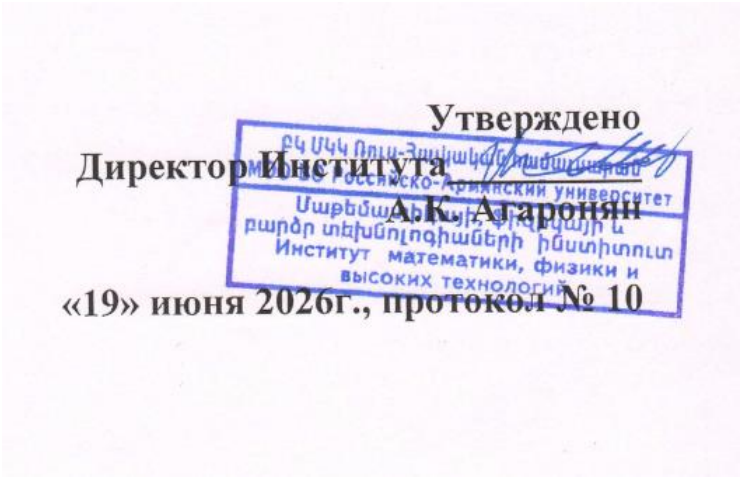
Межгосударственная образовательная организация высшего образования
«Российско-Армянский университет»



УЧЕБНО-МЕТОДИЧЕСКИЙ КОМПЛЕКС ДИСЦИПЛИНЫ

Наименование дисциплины: Б1.В.01 Статистический анализ данных

Автор: Костанян Артавазд Арамович, к.ф.м.н., ст. преподаватель

Направление подготовки: 11.04.04 Электроника и наноэлектроника
Наименование образовательной программы: Квантовая и оптическая электроника

Согласовано:

Заведующий Кафедрой общей физики и квантовых наноструктур

Айрапетян Д.Б.

(подпись)

1. АННОТАЦИЯ

1.1. Краткое описание содержания данной дисциплины;

Данный курс знакомит с базовыми методами статистического анализа данных: классические линейные методы регрессии – логистическая и линейная регрессии, классификации: логистическая регрессия и линейный дискриминантный анализ, древовидные модели, модели без учителя и т.д. Центральная задача анализа данных заключается в выборе наилучшего статистического метода для данной задачи. В течении курса студенты знакомятся также с методами перекрестной проверки, бутстрапом, ознакомятся с нелинейными методами. В рамках курса исследуются также древовидные методы, включая группировку, случайные леса, метод опорных векторов — набор подходов для выполнения как линейной, так и нелинейной классификации.

1.2. Трудоемкость в академических кредитах и часах, формы итогового контроля (экзамен/зачет);

2 з.е. (72ч.), зачет

1.3. Взаимосвязь дисциплины с другими дисциплинами учебного плана специальности (направления)

Методы машинного обучения в материаловедении, Элементы квантовой и оптической электроники, Компьютерные технологии в физике, Квантоворазмерные системы наноэлектроники.

1.4. Результаты освоения программы дисциплины:

Код компетенции *(в соответствии рабочим с учебным планом)*	**Наименование компетенции** *(в соответствии рабочим с учебным планом)*	**Код индикатора достижения компетенций** *(в соответствии рабочим с учебным планом)*	**Наименование индикатора достижений компетенций***(в соответствии рабочим с учебным планом)*
ПК-2	Способен разрабатывать эффективные алгоритмы решения сформулированных задач с использованием современных языков программирования и обеспечивать их программную реализацию	ПК-2.1 ПК-2.2	Знает методы разработки эффективных алгоритмов решения научно-исследовательских задач Умеет использовать алгоритмы решения

		ПК-2.3	исследовательских задач с использованием современных языков программирования Владеет навыками разработки стратегии и методологии исследования изделий микро- и наноэлектроники

2. УЧЕБНАЯ ПРОГРАММА

2.1. Цели и задачи дисциплины

Целью изучения дисциплины «Статистический анализ данных» является формирование знаний и навыков в области применения технологий для анализа данных. Она включает в себя разработку, оценку и адаптацию статистических моделей в предметной среде основы анализа данных, например, метод ближайших соседей, байесовские классификаторы, метод опорных векторов, древовидные методы, случайный лес, градиентный бустинг, а также комплекс прикладных программ для реализации данных технологий.

Задачами дисциплины являются:

приобрести

- умение применять методы статистического анализа для решения практических задач по обработке данных
- умение анализировать задачи статистического анализа данных и осуществлять взвешенный выбор того или иного решения

ознакомить

- студентов с компьютерными технологиями обработки многомерных неструктурированных массивов разнородной статистической информации

научить

- интерпретировать полученные результаты построенных статистических моделей.

В результате изучения данного курса студенты получат знания о принципах применения статистического анализа данных, о его методах анализа и инжиниринга,

приобретут навыки и умения применения методов решающих деревьев, случайного леса, и т.п..

2.2. Трудоемкость дисциплины и виды учебной работы (в академических часах и зачетных единицах) ***(удалить строки, которые не будут применены в рамках дисциплины)***

Виды учебной работы	Всего, в акад. часах	Распределение по семестрам
		3 сем
1	2	3
1. Общая трудоемкость изучения дисциплины по семестрам, в т. ч.:	72	72
1.1. Аудиторные занятия, в т. ч.:	16	16
1.1.1. Лекции	10	10
1.1.2. Практические занятия, в т. ч.	6	6
1.2. Самостоятельная работа, в т. ч.:	56	56
1.3. Консультации		
Итоговый контроль (Экзамен, Зачет, диф. зачет - указать)	Зачет	Зачет

2.3. Содержание дисциплины

2.3.1. Тематический план и трудоемкость аудиторных занятий (модули, разделы дисциплины и виды занятий) по рабочему учебному плану

Разделы и темы дисциплины	Всего (ак. часов)	Лекции(ак. часов)	Практ. Занятия (ак. часов)
1	2=3+4	3	4
Тема 1. Введение	1	2	
Тема 2. Задача предмета анализа данных	1		2
Тема 3. Линейная регрессия	2	2	
Тема 4. Классификации	2		
Тема 5. Методы повторной выборки	2	2	2
Тема 6. Регуляризация линейной модели	2		
Тема 7. Нелинейные модели	1	2	
Тема 8. Древовидные модели	1		2
Тема 9. Модели без учителя	2	2	
Тема 10. Многократное тестирование	2		
ИТОГО	16	10	6

2.3.2. Краткое содержание разделов дисциплины в виде тематического плана

Тема 1. Введение

Что такое статистический анализ данных? Зачем и кака оценивать функцию предсказания?

Тема 2. Задача предмета

Компромисс между точностью прогнозирования и интерпретируемостью модели. Контролируемая и неконтролируемая модели. Проблемы регрессии и классификации. Оценка точности модели. Измерение качества соответствия. Компромисс между смещением и дисперсией.

Тема 3. Линейная регрессия

Простая линейная регрессия. Оценка коэффициентов. Оценка точности коэффициентов. Оценка точности модели. Множественная линейная регрессия. Оценка коэффициентов регрессии. Другие соображения по регрессионной модели. Качественные предикторы. Расширения линейной модели. Сравнение линейной регрессии с K-ближайшими соседями

Тема 4. Классификации

Обзор классификации. Логистическая регрессия. Оценка коэффициентов регрессии. Множественная логистическая регрессия. Мультиномиальная логистическая регрессия. Генеративные модели для классификации. Линейный дискриминантный анализ. Квадратичный дискриминантный анализ. Наивный байесовский алгоритм. Аналитическое сравнение. Обобщенные линейные модели. Обобщенные линейные модели в большей общности

Тема 5. Методы повторной выборки

Перекрестная проверка. Подход к набору проверки. Перекрестная проверка с исключением одного. Перекрестная проверка k-кратности. Компромисс смещения и дисперсии для перекрестной проверки k-кратности. Перекрестная проверка в задачах классификации. Бутстрап.

Тема 6. Регуляризация линейной модели

Выбор подмножества. Пошаговый выбор. Выбор оптимальной модели. Методы сжатия. Регрессия гребня. Лассо. Выбор параметра настройки. Методы сокращения размерности. Регрессия главных компонентов. Частичные наименьшие квадраты. Данные больших размерностей. Регрессия в больших размерностях. Интерпретация результатов в больших размерностях.

Тема 7. Нелинейные модели

Полиномиальная регрессия. Ступенчатые функции. Базисные функции. Сплайны регрессии. Кусочные полиномы. Ограничения и сплайны. Представление базиса сплайна. Выбор количества и расположения узлов. Сравнение с полиномиальной регрессией. Сглаживающие сплайны. Обзор сглаживающих сплайнов. Выбор параметра сглаживания. Локальная регрессия. Обобщенные аддитивные модели (GAM).

Тема 8. Древовидные модели

Основы деревьев решений. Деревья регрессии. Деревья классификации. Деревья против линейных моделей. Преимущества и недостатки деревьев. Бэггинг. Случайные леса. Бустинг. Деревья аддитивной регрессии Байеса. Краткое изложение методов ансамбля деревьев

Тема 9. Модели без учителя

Проблема модели без учителя. Анализ главных компонентов. Что такое главные компоненты? Объяснение доли дисперсии. Подробнее о PCA. Другие применения главных компонентов. Пропущенные значения и заполнение матриц. Методы кластеризации. Кластеризация методом k-средних. Иерархическая кластеризация.

Тема 10. Многократное тестирование

Краткий обзор проверки гипотез. Проверка гипотезы. Ошибки типа I и типа II. Проблема множественного тестирования. Частота ошибок по семействам. Что такое частота ошибок по семействам? Подходы к контролю частоты ошибок по семействам.

2.3.3. Краткое содержание практических занятий

Разработка и представление результатов, на основе изученных метдодов в формате ноутбуков (Jupyter Notebook) разработанных средствами языка Python.

1. Практическое занятие 1: простая линейная регрессия, множественная линейная регрессия, многомерное качество соответствия, качественные предикторы.
2. Практическое занятие 2: логистическая регрессия, линейный дискриминантный анализ, квадратичный дискриминантный анализ, наивный байесовский алгоритм, классификацияя k-NN, регрессия Пуассона
3. Практическое занятие 3: приближение набора проверки, перекрестная проверка. перекрестная проверка k-кратности, бутстрап.

4. Практическое занятие 4: методы выбор подмножеств: пошаговый выбор, выбор оптимального подмножества, регрессия гребня и лассо, регрессия главных компонентов, частичные наименьшие квадраты
5. Практическое занятие 5: полиномиальная регрессия и ступенчатые функции, сплайны, сглаживающие сплайны, обобщенные аддитивные модели, GAM регрессии, тест ANOVA и аддитивные модели
6. Практическое занятие 6: деревья решений, деревья регрессии и классификации, бэггинг, бустинг, деревья аддитивной регрессии Байеса
7. Практическое занятие 7: анализ главных компонентов, кластеризация методом k-средних, иерархическая кластеризация.
8. Практическое занятие 8: проверка гипотезы, ошибки типа I и типа II, частота ошибок по семействам, частота ложных открытий

2.3.4. Материально-техническое обеспечение дисциплины

Компьютерная аудитория, проектор

2.4. Модульная структура дисциплины с распределением весов по формам контролей

Формы контролей	**Вес формы (форм) текущего контроля в результирующей оценке текущего контроля (по модулям)**		**Вес формы промежуточного контроля в итоговой оценке промежуточного контроля**		**Вес итоговой оценки промежуточного контроля в результирующей оценке промежуточных контролей**		**Вес итоговой оценки промежуточного контроля в результирующей оценке промежуточных контролей (семестровой оценке)**	**Веса результирующей оценки промежуточных контролей и оценки итогового контроля в результирующей оценке итогового контроля**
Вид учебной работы/контроля	**М1[1]**	**М2**	**М1**	**М2**	**М1**	**М2**		
Контрольная работа *(при наличии)*			0.5	0.5				
Устный опрос *(при наличии)*								
Тест *(при наличии)*								
Лабораторные работы *(при наличии)*	0.5	0.5						
Письменные домашние задания *(при наличии)*								
Реферат *(при наличии)*								
Эссе *(при наличии)*								
Проект *(при наличии)*								
Решение задач	0.5	0.5						

[1] Учебный Модуль

Веса результирующих оценок текущих контролей в итоговых оценках промежуточных контролей					0.5	0.5		
Веса оценок промежуточных контролей в итоговых оценках промежуточных контролей								
Вес итоговой оценки 1-го промежуточного контроля в результирующей оценке промежуточных контролей							0.5	
Вес итоговой оценки 2-го промежуточного контроля в результирующей оценке промежуточных контролей							0.5	
Вес результирующей оценки промежуточных контролей в результирующей оценке итогового контроля								0.5
Вес итогового контроля (Экзамен/зачет) в результирующей оценке итогового контроля								0.5
	∑ = 1	**∑ = 1**	**∑ = 1**	**∑ = 1**	**∑ = 1**	**∑ = 1**	**∑ = 1**	**∑ = 1**

3. Теоретический блок

3.1. Материалы по теоретической части курса

3.1.1. *Учебник*: G.James, D.Witten, T.Hastie, R.Tibshirani, J.Taylor, An Introduction to Statistical Learning With Applications in Python, Springer International Publishing (2023), https://www.statlearning.com

3.1.2. *Краткие конспекты лекций (слайды)*: https://www.statlearning.com/resources-python

4. Фонды оценочных средств

4.1. Планы практических занятий

1. Решения регрессионных задач с использованием простой и множественной линейной регрессий
2. Решение задач классификации: логистическая регрессия, линейный дискриминантный анализ, наивный байесовский алгоритм, классификацияя k-NN
3. Задача приближениея набора проверки, перекрестная проверка, перекрестная проверка k-кратности, бутстрап

4. Имплементация метода выбора подмножеств: пошаговый выбор, выбор оптимального подмножества, регрессия гребня и лассо, регрессия главных компонентов, частичные наименьшие квадраты
5. Задача регрессии с использованием полиномиальной регрессии и ступенчатых функций, сплайны, сглаживающие сплайны, имплементация и ознокомление с тестом ANOVA
6. Классификация и регрессия с деревом регрессии и классификации, бэггинг, бустинг, деревья аддитивной регрессии Байеса
7. Реализация анализа главных компонентов, кластеризация методом k-средних, иерархическая кластеризация
8. Проверка гипотезы, оценка ошибок типа I и типа II, частота ошибок по семействам

4.2. Материалы по практической части курса

4.2.1 Задачи из учебника: G.James, D.Witten, T.Hastie, R.Tibshirani, J.Taylor, An Introduction to Statistical Learning With Applications in Python, Springer International Publishing (2023), https://www.statlearning.com

4.2.2 Jupyter ноутбуки на Python из ресурсов к задачнику: https://github.com/intro-stat-learning/ISLP_labs/blob/stable/Ch02-statlearn-lab.ipynb

10.4. Вопросы и задания для самостоятельной работы студентов

1. Что такое распределение Стьюдента и t-критерий Стьюдента?
2. Что такое F-распределение и F-критерий?
3. Что такое проверка гипотез и как она работает (нулевая гипотеза, альтернативная гипотеза, p-значения и т. д.)?
4. Упражнение 8 на странице 129 из учебника.
5. Упражнение 9 на странице 129 из учебника.
6. Что такое классификатор Байеса и граница принятия решений Байеса?
7. Доказать, что представление логистической функции и логит-представление для модели логистической регрессии эквивалентны.
8. Доказать, что при предположении, что наблюдения в k-м классе взяты из распределения $N(\mu_k, \sigma^2)$, байесовский классификатор относит наблюдение к классу, для которого дискриминантная функция максимальна.
9. Упражнение 13 на странице 196 из учебника.

10. Упражнение 14 на странице 197 из учебника.
11. Что такое k-кратная кросс-валидация?
12. Упражнение 5 на странице 225 из учебника.
13. Упражнение 6 на странице 226 из учебника.
14. Упражнение 8 и 9 на странице 286 из учебника.
15. Упражнение 6, 7 и 8 на странице 327 из учебника.

10.5. Перечень вопросов для зачета

1. Точность прогнозирования и интерпретируемость статистической модели. Оценка точности модели, измерение качества соответствия.
2. Простая и множественная линейные регрессии. Оценка коэффициентов регрессии, точности модели. Качественные предикторы. Выбросы (outliers).
3. Логистическая регрессия. Множественная и мультиномиальная логистические регрессии.
4. Линейный и квадратичный дискриминантные анализы.
5. Наивный байесовский алгоритм. Сравнение методов классификации.
6. Перекрестная проверка. Подход к набору проверки.
7. Перекрестная проверка с исключением одного. Перекрестная проверка k-кратности.
8. Перекрестная проверка в задачах классификации. Бутстрап.
9. Выбор подмножества. Пошаговый выбор.
10. Выбор оптимальной модели. Методы сжатия.
11. Регрессия гребня. Лассо. Выбор параметра настройки.
12. Методы сокращения размерности. Регрессия главных компонентов. Частичные наименьшие квадраты.
13. Полиномиальная регрессия. Ступенчатые функции.
14. Базисные функции. Сплайны регрессии. Кусочные полиномы.
15. Локальная регрессия. Обобщенные аддитивные модели. GAM для задач регрессии и классификации
16. Основы деревьев решений. Деревья регрессии. Деревья классификации. Деревья против линейных моделей. Бэггинг. Случайные леса.
17. Бустинг. Деревья аддитивной регрессии Байеса.

18. Прогнозирование временных рядов. Краткое описание RNN.
19. Анализ главных компонентов. Что такое главные компоненты?
20. PCA и другие применения главных компонентов. Пропущенные значения и заполнение матриц.
21. Методы кластеризации. Кластеризация методом k-средних. Иерархическая кластеризация.
22. Проверка гипотезы. Ошибки типа I и типа II. Проблема множественного тестирования. Частота ошибок по семействам.

10.6. Образцы зачетных билетов

МОО ВО РОССИЙСКО-АРМЯНСКИЙ УНИВЕРСИТЕТ
ИНСТИТУТ МАТЕМАТИКИ, ФИЗИКИ И ВЫСОКИХ ТЕХНОЛОГИЙ
Кафедра общей физики и квантовых наноструктур

Направление: Электроника и наноэлектроника
Дисциплина: Статистический анализ данных
(магистратура II курс, I семестр)

Экзаменационный билет № **

1. Регрессия главных компонентов. Частичные наименьшие квадраты.
2. Методы кластеризации. Кластеризация методом k-средних.
3. Бустинг. Деревья аддитивной регрессии Байеса.

Зав. кафедрой ОФКН _____Д.Б. Айрапетян
20 г.

МОО ВО РОССИЙСКО-АРМЯНСКИЙ УНИВЕРСИТЕТ
ИНСТИТУТ МАТЕМАТИКИ, ФИЗИКИ И ВЫСОКИХ ТЕХНОЛОГИЙ
Кафедра общей физики и квантовых наноструктур

Направление: Электроника и наноэлектроника
Дисциплина: Статистический анализ данных
(магистратура II курс, I семестр)

Экзаменационный билет № ***

1. Логистическая регрессия. Множественная и мультиномиальная логистические регрессии.
2. Деревья классификации. Бэггинг.

3. PCA и другие применения главных компонентов. Пропущенные значения и заполнение матриц.

Зав. кафедрой ОФКН _____Д.Б. Айрапетян
20 г.

---

11. Методический блок

11.4. Методика преподавания

Лекционные занятия споследующими практическими заданияни, выполняемые как на выцхислительных средствах в аудитории, так и дома на персональных компьютерах