Міністерство освіти і науки України

Харківський національний університет імені В.Н. Каразіна

Кафедра теоретичної та прикладної системотехніки

“ЗАТВЕРДЖУЮ”

Проректор з науково-педагогічної роботи

_____

“_____”_____2018 р.

Робоча програма навчальної дисципліни

Аналіз даних

рівень вищої освіти __перший (бакалаврський)_____

галузь знань _0502 «Автоматика і управління»_____

напрям підготовки 6.050201 «Системна інженерія»_____

вид дисципліни ___за вибором__

факультет комп’ютерних наук

2018 / 2019 навчальний рік

Програму обговорено та рекомендовано до затвердження вченою радою факультету комп’ютерних наук
“ 29 ” серпня 2018 року, протокол № 9

РОЗРОБНИК ПРОГРАМИ:
доктор технічних наук, професор, професор кафедри теоретичної та прикладної системотехніки **Угрюмов Михайло Леонідович,**
кандидат технічних наук, доцент, доцент кафедри теоретичної та прикладної системотехніки **Бакуменко Ніна Станіславівна.**

Програму схвалено на засіданні кафедри теоретичної та прикладної системотехніки
Протокол від “ 19 ” червня 2018 року № 12

Завідувач кафедри теоретичної та прикладної системотехніки
_____ Шматков С. І.

Програму погоджено методичною комісією факультету комп’ютерних наук
Протокол від “ 27 ” червня 2018 року № 7

Голова методичної комісії факультету комп’ютерних наук
_____ Васильєва Л. В.

ВСТУП

Програма навчальної дисципліни «Аналіз даних» розроблена відповідно до освітньо-професійної програми підготовки першого (бакалаврського) рівня напряму підготовки 6.050201 «Системна інженерія».

1. Опис навчальної дисципліни

1.1. Мета викладання навчальної дисципліни.

Засвоєння студентами основ моделювання чисельних даних та методології статистичного аналізу, статистичного оцінювання взаємозв'язку величин, зниження розмірності даних та ін., вироблення навичок по адаптації стандартних алгоритмів до нових – чисельних рішень складних прикладних задач, а також придбання знань про пакети прикладних програм спеціального призначення.

Об'єктом вивчення дисципліни «Аналіз даних» є сучасні математичні моделі та методи статистичного аналізу даних при управлінні складними комп'ютеризованими системами, у якій розробляються математичні моделі, методи й алгоритми аналізу даних, а також шляхи використання для цієї мети сучасних комп'ютерних систем, спеціалізованих пакетів прикладних програм.

Предметом вивчення є методи й алгоритми статистичного аналізу даних при управлінні складними комп'ютеризованими системами, оцінки їх ефективності та ін., для рішення яких розробляється математичне забезпечення комп'ютерних систем, а також використовуються спеціалізовані пакети прикладних програм.

1.2. Основні завдання вивчення дисципліни.

Основними завданнями вивчення навчальної дисципліни є:
– діагностування систем на основі даних моніторингу;
– статистичне оцінювання параметрів розподілів;
– дисперсійний аналіз даних;
– кореляційний аналіз даних;
– множинний регресійний аналіз;
– оцінювання інформативності (значущості) змінних при невизначеності даних;
– прогнозування часових рядів.

1.3. Кількість кредитів – 3.

1.4. Загальна кількість годин – 90.

1.5. Характеристика навчальної дисципліни	
За вибором	
Денна форма навчання	Заочна (дистанційна) форма навчання
Рік підготовки	
4-й	-й
Семестр	
7-й	-й
Лекції	
24 год.	год.
Практичні, семінарські заняття	
24 год.	год.
Лабораторні заняття	
0 год.	год.
Самостійна робота	
42 год.	год.
Індивідуальні завдання	
0 год.	

1.6. Заплановані результати навчання

Відповідно до вимог освітньо-кваліфікаційного рівня підготовки за результатами вивчення дисципліни студенти повинні –

знати:

- основні цілі та вихідні передумови застосування статистичних методів при управлінні складними комп'ютеризованими системами;
- методи попередньої обробки даних;
- основні поняття вибіркового методу;
- методи перевірки статистичних гіпотез;
- характеристики статистичного зв'язку кількісних даних;
- моделі дисперсійного аналізу статистичних даних;
- методи кореляційного аналізу статистичних даних;
- методи регресійного аналізу;
- методи оцінювання інформативності (значущості) змінних при невизначеності даних;
- методи прогнозування часових рядів;

уміти:

- вибирати адекватні методи статистичного аналізу даних у відповідності з метою дослідження та характером статистичних даних;
- знаходити чисельні характеристики статистичних розподілів вибірок даних;
- перевіряти основні гіпотези щодо параметрів розподілення даних;
- видаляти аномальні спостереження у скалярних та векторних даних;
- робити одно факторний та двофакторний дисперсійний аналіз даних;
- знаходити рівняння лінійної регресії;
- перевіряти значущість коефіцієнтів лінійної регресії;
- знаходити довірчі інтервали для коефіцієнтів лінійної регресії;
- будувати лінійну множинну регресію;
- знаходити рівняння нелінійної регресії;
- визначати статистичні оцінки для параметрів нелінійної регресії;
- оцінювати інформативність змінних з врахуванням точності вимірювання змінних стану і наявності парної кореляції між ними;
- оцінювати довірчі інтервали для математичного очікування нелінійних залежностей методом Монте-Карло;
- представляти змістовну інтерпретацію результатів статистичного аналізу;
- працювати з сучасними програмними системами статистичного аналізу даних (Statistica, MATLAB Statistic Toolbox, SPSS);

придбати навички:

- формулювання змістовної та математичної постановок задач, здійснювання формалізації представлення даних, структуризації поставлених задач;
- засвоєння основних методів і прийомів аналізу та обробки різних видів інформації; розробки моделей та методів статистичного аналізу даних;
- проведення верифікації математичних методів, оцінки їх якості на основі перевірки існуючих статистичних гіпотез ;
- вирішення задач чисельного характеру з застосуванням спеціалізованих пакетів;

мати уявлення:

- про класичні і сучасні методи статистичного аналізу даних;
- про межу можливих застосувань методів статистичного аналізу даних.

2. Тематичний план навчальної дисципліни

Розділ 1. Оцінювання невизначеностей при вимірюванні в статистичному дослідженні.

Тема 1. Діагностування систем на основі даних моніторингу.

Діагностування як процес розпізнавання стану систем. Етапи процесу діагностування. Показники якості діагностування. Інформаційно-аналітичне забезпечення процесів діагностування систем на основі даних моніторингу.

Типи невизначеності: епістеміческа, алеаторна (параметрична). Форми обліку результатів спостереження.

Тема 2. Статистичне оцінювання параметрів розподілив.

Статистичні оцінки математичного очікування, факторної дисперсії по вибірки. Типи та види робастного оцінювання. Метод максимальної правдоподібності (М-оцінювання). Формулювання та перевірка гіпотез про рівність центрів розподілів, рівність дисперсій.

Поняття довірчого інтервалу. Довірча ймовірність. Побудова довірчого інтервалу для математичного очікування при відомій, невідомої дисперсії. Побудова довірчого інтервалу для дисперсії.

Тема 3. Дисперсійний аналіз даних.

Постановка задачі однофакторного дисперсійного аналізу. Основні посилання дисперсійного аналізу. Постановка задачі двофакторного дисперсійного аналізу. Факторній аналіз. Метод головних компонент.

Тема 4. Кореляційний аналіз даних.

Коефіцієнт детермінації як універсальна характеристика ступеню тісноти статистичного зв'язку. Кореляційне відношення. Дослідження лінійної залежності за допомогою парного коефіцієнта кореляції. Множинні та часткові коефіцієнти кореляції.

Розділ 2. Формальні математичні моделі систем і процесів.

Тема 5. Множинний регресійний аналіз.

Регресійний аналіз. Основні методи побудови регресійних моделей. Модель лінійної регресії. Оцінка повноти, адекватності моделі, інтерпретація та оцінки коефіцієнтів рівнянь регресії, рівень значущості коефіцієнтів. Обмеження регресійної моделі - мультіколінеарность, гомоскедастічность.

Модель логістичної регресії. Інтерпретація коефіцієнтів логістичної регресії.

Некоректно поставлені завдання. Алгоритми, що регулярізують (робастні алгоритми): адаптивні, інваріантні. Методи регулярізації в задачах ідентифікації, апроксимації даних та прогнозування часових рядів.

Тема 6. Методи оцінювання інформативності (значущості) змінних при невизначеності даних.

Методи оцінюванні диференціальної інформативності з врахування точності вимірювання змінніх стану и наявність парної кореляції между ними: кореляційного аналізу, дисперсійного аналізу і методи розпізнавання образів. Методи розпізнавання образів: детерміністські (дискримінантного аналізу, багатовимірного шкалювання і логічні), ймовірносно-статистичні (методи Байеса, послідовного аналізу і оцінювання на основі теорії інформації). Стохастичний аналіз інформативності: індекси Соболя. Taguchi S / N Ratio.

Оцінювання інформативності на основі методів структурно-параметричного аналізу і синтезу регресійних моделей: факторного аналізу (головних компонент (МГК), нелінійні МГК, Грамма-Шмідта, аналізу компонентів на основі теорії інформації) і спрямованого перебору (ітеративні - на основі різних типів аппроксіматоров, в тому числі штучних нейронних мереж, що навчається), послідовного аналізу варіантів, вагові з адаптацією, локально-стохастичні на основі самоорганізації.

Статистичні оцінки довірчих інтервалів для математичного очікування нелінійних залежностей методом Монте-Карло.

Тема 7. Прогнозування часових рядів.

Стаціонарність, автоковаріації і автокореляції. Основні описові статистики для часових рядів. Використання трендовій лінійної регресії з детермінованими чинниками для моделювання часового ряду. Прогнози по регресії з детермінованими чинниками. Лаговий оператор. Стабільне та ефективне оцінювання параметрів трендових регресійних моделей.

Багатовимірні часові ряди. Багатовимірні трендові регресійні моделі. Ранг коінтеграції (розмірність простору коінтегрованих часових рядів).

Згладжування часового ряду. Авторегресійні трендові моделі.

Розладнання часових рядів. Критерії тренду.

3. Структура навчальної дисципліни

Назви розділів і тем	Кількість годин											
	денна форма						заочна форма					
	усього	у тому числі					усього	у тому числі				
		л	п	лаб.	інд.	с. р.		л	п	лаб.	інд.	с. р.
1	2	3	4	5	6	7	8	9	10	11	12	13
Розділ 1. Оцінювання невизначеностей при вимірюванні в статистичному дослідженні.												
Тема 1. Діагностування систем на основі даних моніторингу	12	2	2			8						
Тема 2. Статистичне оцінювання параметрів розподілив.	14	4	4			6						
Тема 3. Дисперсійний аналіз даних.	14	4	4			6						
Тема 4. Кореляційний аналіз даних.	14	4	4			6						
Разом за розділом 1	54	14	14			26						
Розділ 2. Формальні математичні моделі систем і процесів.												
Тема 5. Множинний регресійний аналіз.	14	4	4			6						
Тема 6. Методи оцінювання інформативності (значущості) змінних при невизначеності даних.	8	2	2			4						
Тема 7. Прогнозування часових рядів.	14	4	4			6						
Разом за розділом 2	36	10	10			16						
Усього годин	90	24	24			42						

4. Теми практичних, лабораторних занять

№ п/п	Назва теми	Кількість годин
1	Перевірка гіпотези про вид розподілу ознаки	2
2	Статистичне оцінювання параметрів розподілів	4
3	Дисперсійний аналіз даних	4
4	Кореляційний аналіз даних	4
5	Множинний регресійний аналіз.	4
6	Оцінювання інформативності (значущості) змінних при невизначеності даних.	2
7	Прогнозування часових рядів	4
	Разом	24

5. Завдання для самостійної роботи

№ п/п	Зміст	Кількість годин
1	Дослідити перевірку значущості коефіцієнтів рівняння множинної регресії	6
2	Дослідити побудову довірчого та предикативного інтервалів для значення функції регресії	6
3	Провести розрахунки для двофакторного дисперсійного аналізу	6
4	Дослідити методи оцінки рангової кореляції	6
5	Дослідити метод головних компонент	6
6	Дослідити експоненційне згладжування для прогнозування часових рядів	6
7	Дослідити методи ковзного середнього для прогнозування часових рядів	6
	Разом	42

6. Індивідуальні завдання

Індивідуальне завдання пов'язане із застосуванням математичних моделей та методів статистичного аналізу даних в конкретному завданні, розробкою програми для його реалізації і обґрунтуванням корисності і ефективності прийнятого рішення.

Індивідуальне завдання виконується у вигляді контрольної роботи.

7. Методи контролю

Контроль роботи студентів при вивченні дисципліни і засвоєння ними навчального матеріалу здійснюється на практичному зайнятті шляхом проведення «летючок», контрольних опитувань і захисту звітів по індивідуальних завданнях. Підсумковий контроль здійснюється при виконанні контрольної роботи і на заліку.

Студенти, що не захистили впродовж семестру звіти з практичних завдань, до заліку не допускаються.

Заліковий квиток містить два теоретичних і одне практичне питання. Максимальна кількість балів за відповіді на кожне теоретичне питання складає по 12 балів, на практичне питання - 16 балів. Проведення поточного контролю, письмового модульного контролю, фінальний контроль у вигляді заліку.

8. Схема нарахування балів

Поточний контроль, самостійна робота, індивідуальні завдання										Залікова робота	Сума
Розділ 1				Розділ 2			Контрольна робота, передбачена навчальним планом	Індивідуальне завдання	Разом		
Т1	Т2	Т3	Т4	Т5	Т6	Т7					
10	5	5	5	10	5	5	15		60	40	100

Т1, Т2 ... – теми розділів.

Критерії оцінювання

Сума балів за всі види навчальної діяльності протягом семестру	Оцінка	
	для чотирирівневої шкали оцінювання	для дворівневої шкали оцінювання
90 – 100	відмінно	зараховано
70-89	добре	
50-69	задовільно	
1-49	незадовільно	не зараховано

9. Рекомендована література

Основна література

1. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии. – М.: ЮНИТИ, 1998. – 1000 с.

2. Тюрин Ю.Н., Макаров А.А. Статистический аналіз данніх на компьютере. – М.: ИНФРА, 1998. - 528 с.

3. Лапач С.Н., Чубенко А.В., Бабич П.Н. Статистика в науке и бизнесе. – К.: МОРИОН, 2002. – 640 с.

4. Львовский Е.Н. Статистические методы построения эмпирических формул: Учеб. пособие для втузов. – М.: Высш. шк., 1988. – 239 с.

Допоміжна література

1. Адлер Ю.П. Маркова Е.В. Планирование эксперимента при поиске оптимальных условий. – М. : Наука, 1976. – 280 с.

2. Бродский В.З. Введение в факторное планирование эксперимента.–М.: Наука,1976. – 223 с.

3. Статистические методы в инженерных исследованиях (лабораторный практикум): Учеб. пособие/ Под ред. Г.К. Круга. – М.: Высш. шк., 1983. – 216 с.

4. Андерсон Т. Статистический анализ временных рядов. – М.: Мир, 1976. – 755 с.

5. Анализ данных/ Н.С. Бакуменко, О.С. Радивоненко. – Учеб. пособие по лабораторному практикуму.–Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т.», 2007.– 87 с.

6. Решение экономических задач с использованием статистических пакетов обработки данных: учеб. Пособие по лаб. практикуму / М.С. Мазорчук, Н.С. Бакуменко. – Харьков: Нац. аэрокосм. ун-т «Харьк. авиац. ин-т», 2008. – 95 с.

10. Посилання на інформаційні ресурси в Інтернеті, відео-лекції, інше методичне забезпечення

1. http://www.machinelearning.ru/wiki/index.php?title=Статистический_анализ_данных_%28курс_лекций%2C_К.В.Воронцов%29/2016%2C_ФУПМ
2. https://courses.prometheus.org.ua/courses/IRF/Stat101/2016_T3/about
3. https://ru.coursera.org/learn/data-analytics-business