Силабус освітнього компонента
Програма навчальної дисципліни

МЕТОДИ БАГАТОВИМІРНОГО АНАЛІЗУ ТА BIGDATA В СОЦІОЛОГІЇ

Шифр та назва спеціальності
054 – Соціологія

Інститут
ННІ Соціально-гуманітарних технологій

Освітня програма
Соціологічне забезпечення економічної діяльності

Кафедра
Соціології і публічного управління (305)

Рівень освіти
Магістр

Тип дисципліни
Спеціальна (фахова), Обов'язкова

Семестр
2

Мова викладання
Українська, англійська

Викладачі, розробники

Бірюкова Марина Василівна

Maryna.Biriukova@khpi.edu.ua
Доктор соціологічних наук, професор, доцент кафедри соціології і публічного управління
Автор 120 наукових та науково-методичних праць, у тому числі трьох одноосібних монографій та підручників. Лектор з дисциплін: «Математичні методи в соціології», «Практикум з аналізу соціологічних даних», «Комп'ютерні технології організації соціологічних дисциплін», «Технології соціального проектування», «Методи багатомірного аналізу соціологічних даних». Досвід роботи – 33 роки
Детальніше про викладача на сайті кафедри
https://web.kpi.kharkov.ua/sp/profesors-ko-vikladats-kij-sklad/

Загальна інформація

Анотація

Основними завданнями курсу є: вивчення методів наукових досліджень з теорії організації вибіркових спостережень, обробки та аналізу отриманої інформації, застосування багатовимірних методів та bigdata для соціального аналізу, ідентифікації та розпізнавання образів; моделювання і прогнозування соціальних процесів; використання інформаційних технологій для статистичного обгрунтування прийняття рішень при соціологічному забезпеченні економічної діяльності.

Мета та цілідисципліни

Освоєння методологічних і методичних основ використовування методів багатовимірного аналізу та bigdata для дослідження природи соціальних явищ, для побудови багатовимірних моделей існування та функціонування соціальних об'єктів.

Формат занять

Лекції, лабораторні роботи, самостійна робота, консультації. Підсумковий контроль – іспит.

Компетентності

ЗК05. Здатність оцінювати та забезпечувати якість виконуваних робіт.
СК02. Здатність виявляти, діагностувати та інтерпретувати соціальні проблеми українського суспільства та світової спільноти.
СК03. Здатність проектувати і виконувати соціологічні дослідження, розробляти й обґрунтовувати їхню методологію.
СК04. Здатність збирати та аналізувати емпіричні дані з використанням сучасних методів соціологічних досліджень та цифрових технології.
СК07. Здатність розробляти та оцінювати соціальні проекти і програми.

Результати навчання

ПР01. Аналізувати соціальні явища і процеси, використовуючи емпіричні дані та сучасні концепції і теорії соціології.
ПР02. Здійснювати діагностику та інтерпретацію соціальних проблем українського суспільства та світової спільноти, причини їхнього виникнення та наслідки.
ПР03. Розробляти і реалізовувати соціальні та міждисциплінарні проекти з урахуванням соціальних, економічних, правових, екологічних та інших аспектів суспільного життя.
ПР04. Застосовувати наукові знання, соціологічні та статистичні методи, цифрові технології, спеціалізоване програмне забезпечення для розв'язування складних задач соціології та суміжних галузей знань.
ПР05. Здійснювати пошук, аналізувати та оцінювати необхідну інформацію в науковій літературі, банках даних та інших джерелах.
ПР09. Планувати і виконувати наукові дослідження у сфері соціології, аналізувати результати, обґрунтовувати висновки.

Обсяг дисципліни

Загальний обсяг дисципліни 180 год. (6 кредитівECTS): лекції – 32 год., семінарські заняття – 32 год., самостійна робота – 116 год.

Передумови вивчення дисципліни (пререквізити)

Для успішного проходження курсу необхідно мати знання та практичні навички з наступних дисциплін: «Математичні методи в соціології","Практикум з комп'ютерної обробки соціологічних даних","Соціологічний супровід економічної діяльності". "Інтернет-дослідження економічної діяльності".

Особливості дисципліни, методи та технології навчання

Під час проведення практичних занять з навчальної дисципліни передбачено пояснення алгоритму виконання практичних завдань та їх відпрацювання. Застосовуються наступні методи навчання: пояснювально-ілюстративний; репродуктивний(відпрацювання певних алгоритмів аналізу даних); частково-пошуковий або евристичний метод(під час виконання індивідуальних завдань). На практичних заняттях використовується проєктний підхід до навчання, гейміфікація, акцентується увага на застосуванні інформаційних технологій в організації соціологічних досліджень: проектна і командна робота, peer-to-peer, , кейси.

Програма навчальної дисципліни

Теми лекційних занять

Тема 1. Основні елементи формалізму
Аналіз соціологічної інформації, зібраної в ході емпіричних соціологічних досліджень, є не просто сукупністю технічних прийомів і методів. Неодновимірність багатьох досліджуваних соціологом понять. Непрямий її прояв – порушення транзитивності відношення порядку.Метричне та неметричне БШ. Відповідні функції стресу. Неявне порівняння відстаней між близькістю, закладене у формулі функції стресу для метричного шкалірування. Поняття монотонної регресії,

що використовується при розрахунку функції стресу для неметричного шкалірування.Важливість для соціології неметричного шкалірування. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат.

Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування

Постановка завдання; важливість врахування специфіки метрик окремих респондентів. Спосіб обліку таких метрик в індивідуальному БШ. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ. Одномірне розгортання. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування. Модель ідеальної точки в багатовимірному випадку. Неметричне багатовимірне розгортання. Вид вихідних даних. Функція стресу. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно). Особливості інтерпретації результатів.

Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування.

Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування, та інтерпретації його результатів. Можливі способи одержання вихідних даних. Безпосереднє отримання близькості від респондентів, класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.Робота з БШ статистичними програмами - процедура БШ доступна в більшості статистичних програм. Існує вибір між метричним БШ (який дозволяє працювати з інтервалами чи даними про співвідношення рівня), і неметричним БШ (який працює з порядковими даними).Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей.

Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот

Загальне уявлення про моделювання частот таблиці спряженості. Змістовне розуміння таких моделей, їх роль для соціолога. Мультиплікативні та адитивні моделі частот. Роль логарифмування мультиплікативної моделі. Можливість різного розуміння як сенсу розглянутих вкладів, так і того "середнього" рівня, з яким порівнюються спостерігаються частоти в процесі їх моделювання.Канонічний кореляційний аналіз – один із мето- дів багатовимірного аналізу даних. Необхідність сполучення моделі, закладеної в конкретному методі оцифровки, з вмістом розглянутої задачі. Приклад моделі такого роду – модель, використовувана в методі шкалірування, званому методом послідовних розбивок. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами". Моделі частот, що відповідають канонічному аналізу. Побудова соціологічних індексів за допомогою техніки канонічного аналізу. Вирішення проблеми зважування складових індекс ознак.

Тема 5. Логлінейний аналіз

Логлінейний аналіз - метод багатовимірного статистичного аналізу для вивчення таблиць спряженості. Логлінейний аналіз дозволяє статистично перевіряти гіпотезу про систему одночасно мають місце парних і множинних взаємозв'язків в групі ознак, виміряних за номінальними шкалами. Багатовимірний статистичний аналіз.Моделі частот, що відповідають логлінейному аналізу. Насичена модель. Мета переходу до логарифмів частот. Сенс вкладів різної розмірності.Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі. Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.

Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак

Поняття причини в соціології. Принципова неможливість повністю його формалізувати. Роль статистичних методів при вивченні причинних відносин.Граф причинних зв'язків. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні. Правила редукції причинних схем та формування рівнянь.Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків. Різниця між

статистичним та причинним зв'язком. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи.Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології. Поняття структури багатовимірної випадкової величини. Формування узагальнених показників на базі аналізу структури зв'язків ознак. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів).

Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів

Класифікація як один із фундаментальних процесів у науці. Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі.
Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).
Класифікація як один із фундаментальних процесів у науці. Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковомупросторі.Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).

Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації

Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації. Сенс протиставлення термінів "класифікація" і "типологія". Підстава типології. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування. Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.

Тема 9. Функції відстані між об'єктами

Аксіоматичне визначення функції відстані і ролі цієї функції в соціології. Приклади непридатності евклідової відстані з точки зору апріорного змістовного розуміння типів об'єктів.
Можливість використання евклідової відстані в розглянутих прикладах за рахунок зміни ознакового простору.Сучасний аналіз даних обумовлюється способами отримання величин, методами їх обробки й залежить від розвитку математичних методів і моделювання. Функції відстані, відмінні від евклідова: зважене евклидово, сіті-блок, Махаланобіса, Хеммінгово.

Тема 10. Основні види процедур класифікації. Відстані між класами

Актуальність дослідження сутності та методів багатовимірного аналізу соціологічної інформації обумовлена специфікою соціальної реальності, що завжди уявляється як складний, багатогранний та багатозначний феномен, який інтегрує багатовимірність суспільства з багатовимірністю внутрішнього світу окремої людини. Виділення ієрархічних і неієрархічні алгоритмів класифікації. Багатовимірний статистичний аналіз (у широкому значенні) - розділ математичної статистики, що поєднує методи вивчення даних, які характеризують багатовимірні об'єкти. Агломеративні та дівізімні алгоритми. Причини необхідності розгляду відстаней між класами в ієрархічних процедурах. Алгоритм найближчого сусіда як приклад способу класифікації, що використовує такі відстані.

Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Обумовленість цих гіпотез апріорними уявленнями дослідника про типи об'єктів. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.Факторний аналіз найбільш яскраво відображує риси багатомірного аналізу в частині дослідження зв'язку між ознаками. Кластерний аналіз ці риси відображує з боку класифікації об'єктів. Загальне уявлення про розмиті класифікації. Роль функції належності у відповідних алгоритмах. Доцільність комплексного використання декількох алгоритмів класифікації в соціологічних завданнях побудови типології.Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.

Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології

Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики. Основні фактори, що визначають інтерпретацію вихідних даних: апріорні уявлення дослідника про спосіб породження цих даних (у тому числі – про моделі сприйняття респондентами пропонованих ним питань, об'єктів, про ймовірнісну природу даних і т. д.); мета дослідження; концептуальні уявлення соціолога про досліджуване явище; характер моделі явища, "закладеної" в математичному методі, використання якого планується; розгляд спостережуваних змінних як непрямих показників латентних факторів, насправді цікавлять дослідника і т. п.
Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. д.

Тема 13. Дані. Метадані

Згідно з ГОСТ, дані – подання інформаціїу формалізованому вигляді, придатному для передачі, інтерпретації та обробки.
Вихідне поняття даних - філософське, воно виникає в епістемологіїпід час розгляду основою проблеми гносеології – пізнаваності світу, пошукута осмислення істини. Процедури верифікації чи фальсифікації данихстворюють інформацію, осмислення істини створює знання.
Життєвий цикл даних – це послідовність етапів, якуконкретна порція даних проходить від початкового етапу створення чи отриманнядо моменту архівації чи видалення.
При зборі даних виникають метадані, що містять будь-якуінформацію про зібрані дані.
Огляд основних аналітичних інструментів роботи з Bigdata соціальних науках (Python, R, SAS, та ін). Читання та запис даних, формати файлів. Завантаження даних із різних джерел. Взаємодія з базами даних. Читання даних із Excel. Робота з CSV файлами та даними у форматі JSON. Парсинг простих даних XML. Читання даних із таблиць HTNL. Читання даних із файлу SAS. Взаємодія з HTML та Web API.

Тема 14. Великі дані. Системи керування великими даними

Великі дані можуть бути різних типів. Інформацію, отриманув результаті обліку або вимірювання будь-яких об'єктів або параметрів, називаютьмайстер-даними (MasterData). Наприклад, облік кількості, виміри координаті швидкостей конкретних молекул - це майстер-дані.
Транзакційні дані (в англомовній літературі застосовуються терміниTransactionalData, ApplicationSpecificData, OperationalData) – це дані,що відображають результат виконання будь-яких операцій. Транзакційні дані описуютьвзаємодію об'єктів один з одним або з навколишнім світом, які можнаотримати за допомогою обробки майстер-даних.
Ретроспективні дані (Historicaldata) – це дані, забезпеченіпозначки часу.
Посилальні дані (довідники, НСІ, нормативно-посилальна інформація, ReferenceData, LookupData, Dictionaries) – це базові незмінні дані,заздалегідь відомі із зовнішніх джерел, такі як нормативи, скорочення,акроніми, словники, стандарти.
Формат даних. Структуровані дані мають заздалегідь визначенийформат. Напівструктуровані або слабоструктуровані дані -це дані, які часто зібрані з різних джерел.

Тема 15. Програмні платформи та системи для Великих даних

В даний час використовується значна кількість платформта систем Великих даних. Системи обробки великих даних єфреймворками, тобто каркасами, для використання яких необхідноз'єднати їх з іншими фреймворками, прикладним програмним забезпеченнямкористувача та системою зберігання даних.
В аналітичному звіті BigDataAnalyticsMarketStudy, 2017 Edition наводиться така діаграма інфраструктур Великих даних, впровадженихна підприємствах, представлена у розрізі розмірів підприємств
Розподілена обробка даних тісно пов’язана з паралельною обробкою даних.Однак така обробка завжди виконується за допомогою окремих машин у кластері,підключеному до мережі.

Розподілена обробка даних - це метод виконанняприкладних програм групою систем. Користувач може працювати з мережевимислужбами та прикладними процесами, розташованими в кількох взаємопов'язанихабонентських системах. Розподілена обробка даних підвищує ефективністьінформаційних потреб користувачів і забезпечує ефективність та результативністьрішень.

Тема 16. Машинне навчання за допомогою бібліотекиScikit-learn.
Види машинного навчання. Основні бібліотеки машинного навчанняPython (Scikit-learn, Keras, TensorFlow). Створення тренувальних наборів -передобробка даних. Точність та достовірність моделі. Вибір найкращоїмоделі.
Кроки типового практичного сценарію машинного навчання. Завантаженнянабору даних. Дослідження даних за допомогою Pandas. Візуалізаціяознак за допомогою Matplotlib. Розбиття даних для навчання татестування. Створення моделі. Вивчення моделі. Тестування моделі. Налаштування параметрів моделі та оцінка її точності. Формування прогнозів на підставі «живих» даних, які ще невідомі моделі.
Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.
Лінійні моделі для регресії та класифікації (модель лінійної регресії,логістична регресія, та ін).
Наївні байєсівські класифікатори. Дереварішень та випадковий ліс. Спосіб опорних векторів.
Основи нейронних мереж.
Метод основних компонентів. Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін).

Теми практичних занять

Тема 1. Основні елементи формалізму
Проблеми неодновимірності багатьох досліджуваних соціологом понять. Особливості вивчення простору сприйняття соціологічних явищ та процесів – основне завдання БШ. Ідеї Кумбса щодо урахування можливості упорядкування відстаней між об'єктами.Векторна модель або модель ідеальної крапки як основа БШ.Функція відстані (аксіоматичне визначення). Відповідні функції стресу.Простір сприйняття респондентами запропонованих їм об'єктів. Формальне визначення близькості. Вихідні дані для БШ – матриця близькості між об'єктами. Метричне та неметричне БШ. Формальні аспекти проблем розмірності шуканого евклідового простору і обертання, що визначають його осей координат.Розв'язання практичних завдань.
Тема 2. Багатовимірне розгортання та індивідуальне багатовимірне шкалірування
Постановка завдання важливість врахування специфіки метрик окремих респондентів. Вид вхідних і вихідних даних, функції стресу в індивідуальному БШ. Одномірне розгортання. Обґрунтування необхідності переходу до простору довільної розмірності для успішного виконання завдання шкалірування. Неметричне багатовимірне розгортання. Особливості інтерпретації результатів.Спосіб обліку таких метрик в індивідуальному БШ. Модель ідеальної точки в багатовимірному випадку.Функція стресу. Специфіка вихідних даних (наявність двох видів точок, що відповідають об'єктам і респондентам відповідно).Розв'язання практичних завдань.
Тема 3. Проблеми формування вихідних даних і інтерпретації результатів у багатовимірному шкалірування
Роль соціолога при отриманні даних, вихідних для багатовимірного шкалірування та інтерпретації його результатів. Класифікація відповідних способів опитування; проблеми, що постають при такому способі збору даних. Приклади розрахунку матриці близькості на основі аналізу достатньо надійних даних іншого роду.Використання формальних та неформальних методів при інтерпретації результатів багатовимірного шкалірування. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей. Можливі способи одержання вихідних даних. Проблеми застосування статистичних методів в соціології.Основні функції та процедури аналізу даних. Значення змістовних концепцій дослідника при вирішенні проблем вибору розмірності евклідова простору і повороту його осей. Створення багатовимірних таблиць за допомогою вторинних змінних.Загальна характеристика сучасних програмних засобів аналізу соціологічних даних.Розв'язання практичних завдань.

Тема 4. Канонічний аналіз. Загальне уявлення про методи, які засновані на моделях частот

Загальне уявлення про моделювання частот таблиці спряженості. Мультиплікативні та адитивні моделі частот. Роль логарифмування мультиплікативної моделі. Основне завдання канонічного аналізу. Принципи їх отримання на основі аналізу таблиці спряженості.Моделі частот, що відповідають канонічному аналізу. Зв'язок канонічних коефіцієнтів кореляції з критерієм «хі-квадрат". Загальне уявлення про оцифрування значень номінальних ознак. Канонічний аналіз як метод оцифровки і метод вимірювання зв'язку між двома номінальними ознаками зі "спільними альтернативами". Поняття зв'язку між двома групами ознак.Послідовність канонічних коефіцієнтів кореляції.
Принципи отримання канонічних коефіцієнтів кореляції на основі аналізу таблиці спряженості.
Використання канонічної кореляції в аналізі таблиць спряженості.
Необхідність сполучення моделі, закладеної в конкретному методі оцифровки.
Побудова соціологічних індексів за допомогою техніки канонічного аналізу.
Вирішення проблеми зважування складових індекс ознак.
Розв’язання практичних завдань.

Тема 5. Логлінейний аналіз

Причини відмінності реального розподілу від рівномірного. Моделі частот, що відповідають логлінейному аналізу. Насичена модель. Мета переходу до логарифмів частот. Гіпотези про взаємозв'язок ознак. Їх роль при побудові моделей частот. Розрахунок коефіцієнтів логлінейної моделі для двовимірного випадку. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності).Порівняння логлінейного аналізу з номінальним регресійним і дисперсійним аналізом, а також з методом послідовних розбивок. Порівняння здійснюється на змістовному рівні. Різне розуміння залежної ознаки: кількісна ознака в дисперсійному аналізі, кількісна або номінальна – в номінальному регресійному і частота, що стоїть в клітці багатовимірної таблиці спряженості, – в логлінейном аналізі.Неможливість отримання нового знання на основі аналізу рівномірного розподілу (суть аналізу даних – вивчення змін, порівняння показників різного роду).Сенс вкладів різної розмірності.Роль критерію "хі-квадрат" при використанні логлінейного аналізу. Відносини переважання. Інтерпретація коефіцієнтів через відносини переважання (для моделі довільної розмірності). Різні можливості пошуку поєднань значень предикторів: перевірка гіпотез про наявність багатовимірних зв'язків у логлінейном аналізі і можливість пошуку найбільш дієвих поєднань в методі послідовних розбивок і регресійному аналізі, заздалегідь заданий набір поєднань значень предикторів в дисперсійному аналізі.Розв’язання практичних завдань.

Тема 6. Причинний аналіз. Стратегія аналізу структури взаємозв'язків ознак

Граф причинних зв'язків. Повторення принципів побудови часткових коефіцієнтів кореляції і регресії. Важливість для соціолога вивчення відповідних зв'язків. Поняття "помилкової" кореляції. Основні причинні схеми, що призводять до їх появи.Обчислення коваріацій (кореляцій) між будь-якими двома ознаками на основі графа зв'язків. Структурні рівняння. Обчислення структурних коефіцієнтів. Їх зв'язок з частковими коефіцієнтами регресії. Основна теорема причинного аналізу. Її роль у вивченні статистичних залежностей.Поняття структури багатовимірної випадкової величини. Формування узагальнених показників на базі аналізу структури зв'язків ознак. Роль статистичних методів при вивченні причинних відносин. Структурні коефіцієнти. Вхідні (зовнішні, незалежні) і вихідні (внутрішні, залежні) змінні. Правила редукції причинних схем та формування рівнянь. Різниця між статистичним та причинним зв'язком. Вивчення статистичних зв'язків на основі причинних схем як основне завдання причинного аналізу. Поняття допоміжної теорії вимірювань Блейлока. Причинний аналіз як концептуальний підхід до вивчення соціальних явищ. Проблема формалізації завдання вивчення причинно-наслідкових відносин в соціології. Комплексне використання декількох методів вивчення зв'язків між ознаками для вирішення соціологічних задач (аналіз структури випадкової величини; факторний і дисперсійний аналіз; пошук детермінуючих поєднань значень предикторів).Розв’язання практичних завдань.

Тема 7. Завдання розпізнавання образів. Поняття автоматичної класифікації об'єктів

Класифікація як один із фундаментальних процесів у науці. Загальне уявлення про завдання розпізнавання образів (синоніми: образ, клас, кластер, таксон; неоднозначність трактування термінів в літературі). Виділення завдань: пошук класів, опис класів, визначення найбільш

ефективної системи ознак. Виділення задачі автоматичної класифікації об'єктів (синоніми: багатовимірна класифікація, розпізнавання образів без вчителя, кластерний аналіз, таксономія).Ознаковий простір. Задача класифікації як пошук згущення точок – моделей об'єктів в ознаковому просторі. Роль наявності або відсутності навчальної вибірки.Розв’язання практичних завдань.

Тема 8. Проблема "стикування" змісту і формалізму при використанні алгоритмів класифікації

Специфіка рішення соціологічних завдань побудови типології за допомогою методів автоматичної класифікації. Сенс протиставлення термінів "класифікація" і "типологія". Виділення основних формальних елементів алгоритмів автоматичної класифікації, що вимагають стикування зі змістовними концепціями соціолога.Підстава типології. Роль апріорних уявлень дослідника про шуканих типах у виборі і реалізації алгоритму, інтерпретації результатів його застосування. Розв’язання практичних завдань.

Тема 9. Функції відстані між об'єктами

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Основні види гіпотез: компактності, зв'язності (безперервності), унімодального розподілу.Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза. Приклади алгоритмів, що шукають закономірності розташування точок у ознаковому просторі, що відповідають кожній з гіпотез: алгоритм Форель (гіпотеза компактності), алгоритм найближчого сусіда (гіпотеза зв'язності), алгоритм, заснований на виділенні локальних максимумів функції приналежності (гіпотеза унімодального розподілу).Роль функції належності у відповідних алгоритмах. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Коригування результатів класифікації з метою забезпечення відповідності класифікації і типології.Розв’язання практичних завдань.

Тема 10. Основні види процедур класифікації. Відстані між класами

Виділення ієрархічних і неієрархічні алгоритмів класифікації.Агломеративні та дівізімні алгоритми. Оптимізація розбиття в сенсі максимізації заздалегідь обраного функціоналу якості як один з основних елементів формалізму в неієрархічні алгоритмах класифікації. Основний змістовний сенс оптимізації. Сенс вимірювання близькості між класами в таких випадках. Способи вимірювання сумарних оцінок близькості один до одного об'єктів усередині класів.Розв’язання практичних завдань.

Тема 11. Гіпотези про розташування об'єктів у ознаковому просторі

Роль гіпотез про характер розташування об'єктів у виборі алгоритму класифікації. Приклади соціологічних завдань побудови типології, для яких була б розумна кожна гіпотеза. Загальне уявлення про розмиті класифікації. Роль функції належності у відповідних алгоритмах. Змістовні уявлення соціолога про типи та умови вибору кроку розбиття при інтерпретації результатів. Розв’язання практичних завдань.

Тема 12. Поняття інтерпретації вихідних даних і основні методологічні принципи використання методів аналізу даних в соціології

Інтерпретація вихідних даних як одне з основних ланок "стикування" соціології і математики. Виділення методологічних принципів, дотримання яких є необхідним для того, щоб аналіз соціологічних даних був ефективний, не відводив соціолога в сторону від реальності: забезпечення певної однорідності вихідних даних; облік моделі, "закладеної" в кожному методі аналізу даних, при виборі алгоритму аналізу, два основні принципи інтерпретації результатів аналізу: необхідність її узгодження з інтерпретацією вихідних даних і заповнення при її здійсненні тих втрат, які мали місце при переході до формалізму; необхідність комплексного використання декількох методів для вирішення одного завдання і т. д.Розв’язання практичних завдань.

Тема 13.Дані. Метадані

Створення даних (DataGeneration/DataCapture). Обслуговування даних (DataMaintenance). Синтез даних (DataSynthesis). Використання даних (DataUsage). Публікація даних (DataPublication). Архівація даних (DataArchival). Знищення даних (DataPurging) Розв’язання практичних завдань.

Тема 14. Великі дані. Системи керування великими даними
Розподілені файлові системи. Розподілені фреймворки. Бенчмаркінг. Серверне програмування. Планування. Системи розгортання. Розв'язання практичних завдань.

Тема 15. Програмні платформи та системи для Великих даних
Системи керування потоками даних. Системи зберігання Великих даних. Платформи Великих даних. Обробка даних у реальному часі. Системи керування Великими даними. Аналітичні платформи. Розв'язання практичних завдань.

Тема 16.Машинне навчання за допомогою бібліотеки Scikit-learn.
Кроки типового практичного сценарію машинного навчання. Завантаження набору даних. Дослідження даних за допомогою Pandas. Візуалізація ознак за допомогою Matplotlib. Налаштування параметрів моделі та оцінка її точності. Функціонал бібліотеки Scikit-Learn. Класифікація за допомогою К-сусідів.Лінійні моделі для регресії та класифікації (модель лінійної регресії, логістична регресія, та ін). Дерева рішень та випадковий ліс. Основи нейронних мереж.Алгоритми кластеризації (кластеризація методом К-середніх, ієрархічна кластеризація, та ін). Розв'язання практичних завдань.

Теми лабораторних робіт

Лабораторних занять не передбачено.

Самостійна робота

Самостійна робота за курсом складається із самостійного вивчення студентами тем та питань, які не викладаються на заняттях, виконання індивідуальних завдань. Студентам також рекомендуються додаткові матеріали (відео, статті) для самостійного вивчення та аналізу.

Література та навчальні матеріали

Основна література
1. Горбачик А.П., Сальнікова С.А. Аналіз даних соціологічних дослідженьзасобами SPSS: Навч. посіб.- Луцьк, 2008. – 164 с.IBM SPSS 20 інструкція користувача// https://www.xn--80aaexjatkpdggghih8b1a2yhv.com.ua/ibm/spss-20/%D1%96%D0%BD%D1%81%D1%82%D1%80%D1%83%D0%BA%D1%86%D1%96%D1%8F-%D0%BA%D0%BE%D1%80%D0%B8%D1%81%D1%82%D1%83%D0%B2%D0%B0%D1%87%D0%B0.
2. Паніотто В.І., Максименко В. С., Харченко Н.М. Статистичний аналізсоціологічних даних. - Київ, 2004. – 270 с.Литвин В.В. Аналіз даних та знань: підручник/ В.В. Литвин, В.В. Пасічник, Ю.В. Нікольський.- Л.: Магнолія, 2020.- 276с. (базовий підручник).

Допоміжна література
3. Лупан І.В., Авраменко О.В., АкбашК.С.Комп'ютерні статистичні пакети: навчально-методичнийпосібник. - 2-е вид. - Кіровоград: 'КОД". 2015. - 230 с. - http://dspace.cuspu.edu.ua/jspui/bitstream/123456789.
4. MakingSenseofMultivariateData Analysis//https://us.sagepub.com/en-us/nam/book/making-sense-multivariate-data-analysis
5. Бахрушин В.Є.Методи аналізу даних: навчальний посібник для студентівВ.Є. Бахрушин. - Запоріжжя : КПУ, 2011. - 26В с. - http://web.kpi.kharkov.ua/auts/wp-content/uploads/sites/67/2017/02/DAMAP_Ivashko_posobie2.pdf
6. Інтелектуальний аналіз даних: практикум/ М.Т. Фісун, І.О. Кравець, П.П. Казмірчук.- Л.: Новий Світ-2000, 2020.- 162с.Гладун А.Я., Рогушина Ю. В. DataMining: пошук знань в даних. Київ. ТОВ «ВД «АДЕФ- Україна», 2016. — 452 с..

Система оцінювання

Критерії оцінювання успішності студента та розподіл балів

100% підсумкової оцінки складаються з результатів оцінювання у вигляді іспиту (20%) та поточного оцінювання (80%). Іспит: виконання разрахункового завдання та усна доповідь. Поточне оцінювання: 16 онлайн тестів за темами (48%), два індивідуальні завдання (22%)та два разрахункових завдання (10%)

Шкала оцінювання

Сума балів	Національнаоцінка	ECTS
90–100	Відмінно	A
82–89	Добре	B
75–81	Добре	C
64–74	Задовільно	D
60–63	Задовільно	E
35–59	Незадовільно (потрібне додаткове вивчення)	FX
1–34	Незадовільно (потрібне повторне вивчення)	F

Норми академічної етики і політика курсу

Студент повинен дотримуватися «Кодексу етики академічних взаємовідносин та доброчесності НТУ «ХПІ»: виявляти дисциплінованість, вихованість, доброзичливість, чесність, відповідальність. Конфліктні ситуації повинні відкрито обговорюватися в навчальних групах з викладачем, а при неможливості вирішення конфлікту – доводитися до відома співробітників дирекції інституту. Нормативно-правове забезпечення впровадження принципів академічної доброчесності НТУ «ХПІ» розміщено на сайті: http://blogs.kpi.kharkov.ua/v2/nv/akademichna-dobrochesnist/

Погодження

Силабус погоджено

Дата погодження, підпис

30.06.2023

Завідувач кафедри
Володимир МОРОЗ

Дата погодження, підпис

30.06.2023

Гарант ОП
ЮрійКАЛАГІН