Зерттеу деректерін талдау

Статистика пән ретінде негізінен өткен ғасырда дамыды. Статистиканың математикалық негізі болған ықтималдықтар теориясы 17-19 ғасырлар аралығында дамыды. Томас Байес, Пьер-Симон Лаплас және Карл Гаусс еңбектеріне негізделген. Ықтималдықтың таза теориялық табиғатынан айырмашылығы, статистика деректерді талдау және модельдеумен айналысатын қолданбалы ғылым болып табылады. Заманауи статистика қатаң ғылыми пән ретінде 1800 жылдардың аяғынан басталады. — Фрэнсис Гальтон және Карл Пирсон. 20 ғасырдың басындағы Р.А.Фишер тәжірибелік дизайн және максималды ықтималдықты бағалау сияқты негізгі ұғымдарды енгізген заманауи статистиканың жетекші жаңашылы болды. Осы және басқа да көптеген статистикалық тұжырымдамалар негізінен деректер ғылымының шалғай бұрыштарында кездеседі. Бұл кітаптың негізгі мақсаты – деректер ғылымы мен үлкен деректер контекстінде осы ұғымдарды бөліп көрсетуге және олардың маңыздылығын немесе олардың жетіспеушілігін түсіндіруге көмектесу. Бұл тарауда кез келген деректер ғылымы жобасындағы бірінші қадамға назар аударылады: деректер интеллектісі. Барлау деректерін талдау (EDA) – статистиканың салыстырмалы түрде жаңа саласы. Классикалық статистика дерлік статистикалық қорытындыға, яғни шағын үлгілерден популяцияларға (немесе популяцияларға) қорытынды жасауға арналған кейде күрделі процедуралар жиынтығына бағытталған. 1962 жылы Джон В.Туки (1.1-сурет) өзінің «Деректерді талдаудың болашағы» атты концептуалды мақаласында [Тукей-1962] статистиканы өзгертуге шақырды. Ол статистикалық қорытындыны оның құрамдас бөліктерінің бірі ретінде қамтитын деректерді талдау деп аталатын жаңа ғылыми пәнді ұсынды. Тукей инженерлік және есептеуіш қауымдастықтармен жалған байланыс орнатты (ол «бит» терминдерін - ағылшын екілік цифрынан және «бағдарламалық қамтамасыз ету» деп атады) және оның бастапқы принциптері таңқаларлық күшті болып шықты және деректер ғылымының негізін құрайды. Барлау деректерін талдау саласы Туридің «Бақылау нәтижелерін талдау» [Тукей-1977] кітабының арқасында пайда болды, ол қазір классикаға айналды. Есептеу қуатының және деректерді экспрессивті талдау бағдарламалық қамтамасыз етуінің болуымен деректерді барлау талдауы бастапқы доменнен әлдеқайда жоғары дамыды. Бұл пәннің негізгі драйверлері жаңа технологияның жылдам дамуы, әртүрлі және үлкенірек деректерге қол жеткізу және сандық талдауды көптеген пәндер бойынша көбірек қолдану болды. Дэвид Донохо, мұғалім

Стэнфорд университетінің статисті және бұрынғы Туки түлегі Принстонда (Нью-Йорк штаты) Тукейдің 100 жылдығына арналған семинарда жасаған баяндамасының негізінде «Деректерді ғылымға 50 жыл» тамаша мақаласын жазды. Нью-Джерси [Донохо-2015]. Онда Донохо деректер ғылымының өсуін Тукейдің деректерді талдауға қосқан ізашар үлесіне қарайды.

Рис. 1.1. Джон Тьюки, выдающийся статистик, чьи идеи, разработанные более чем 50 лет назад, формируют фундамент науки о данных

Құрылымдық деректер элементтері

Деректер көптеген көздерден алынады: сенсорлық көрсеткіштер, оқиғалар, мәтін, суреттер және бейнелер. Заттар интернеті (IoT) ақпарат ағындарын таратады. Бұл деректердің көп бөлігі құрылымдалмаған: кескіндер әр пиксельде RGB (қызыл, жасыл, көк) пішіміндегі түсті ақпараты бар пикселдер жиынтығы болып табылады. Мәтіндер көбінесе бөлімдерге, бөлімшелерге және т.б. бөлінген сөздік пен сөздік емес таңбалар тізбегінен тұрады. Пернелерді басу ағындары қолданбамен немесе веб-бетпен әрекеттесетін пайдаланушы әрекеттерінің тізбегі болып табылады. Шын мәнінде, деректер ғылымының негізгі міндеті - бұл бастапқы деректер ағынын практикалық қызметте пайдалы ақпаратқа айналдыру. Осы кітапта талқыланған статистикалық тұжырымдамаларды қолдану үшін құрылымдалмаған бастапқы деректер өңделуі және құрылымдық пішінге (мысалы, реляциялық дерекқордан келуі мүмкін) немесе статистикалық зерттеу үшін жиналуы керек.

Негізгі терминдер

Үздіксіз деректер (үздіксіз) аралықта кез келген мәнді қабылдай алатын деректер. Синонимдер: интервал, өзгермелі нүкте саны, сандық мән.

Дискретті деректер (дискретті) Масштаб мәндері сияқты бүтін мәндерді ғана қабылдай алатын деректер. Синонимдер: бүтін, сан.

Категориялық деректер Белгілі бір мәндер жиынын, атап айтқанда мүмкін санаттар жиынын ғана қабылдай алатын деректер. Синонимдер: санаулар, санамаланған деректер, факторлар, атаулы деректер, полихотомиялық деректер.

Екілік деректер Мәндердің тек екі санаты бар категориялық деректердің ерекше жағдайы (0/1, шын/жалған). Синонимдер: дихотомиялық, логикалық, жалауша, көрсеткіш, логикалық.

Реттік деректер (реттік) Анық реттелген категориялық деректер. Синонимдер: реттік фактор.

Құрылымдық деректердің екі негізгі түрі бар: сандық және категориялық. Сандық деректер екі түрде келеді: жел жылдамдығы немесе уақыт ұзақтығы сияқты үздіксіз және оқиғаның орын алу саны сияқты дискретті. Категориялық деректер теледидар экранының түрі (плазма, СКД, жарық диоды, т.б.) немесе штат атауы (Алабама, Аляска, т.б.) сияқты бекітілген мәндер жинағын ғана қабылдайды. Екілік деректер категориялық деректердің маңызды ерекше жағдайы болып табылады. Бұл деректер 0/1, иә/жоқ немесе шын/жалған сияқты екі мәннің біреуін ғана қабылдайды. Категориялық деректердің тағы бір пайдалы түрі – категориялар реттелген реттік деректер; олардың мысалы - сандық сипаттама (1, 2, 3, 4 немесе 5). Деректер түрлерінің таксономиясымен неге алаңдау керек? Деректерді талдау және болжамды модельдеу мақсаттары үшін деректер түрі визуализация түрін, деректерді талдау немесе статистикалық модельді анықтауда маңызды рөл атқарады. Шын мәнінде, R және Python сияқты деректер ғылымының бағдарламалық жүйелері есептеу өнімділігін жақсарту үшін осы деректер түрлерін пайдаланады. Ең бастысы, айнымалының деректер түрі бағдарламалық жүйенің осы айнымалы үшін есептеулерді қалай өңдейтінін анықтайды.

Бағдарламалық жасақтама әзірлеушілері (бағдарламалық қамтамасыз ету) және мәліметтер базасы (ДҚ) бағдарламашыларында сұрақ туындауы мүмкін: аналитикаға «категориялық» және «реттік» деректер ұғымдары не үшін қажет? Өйткені, санаттар тек мәтіндік (немесе сандық) мәндер жиынтығы болып табылады және негізгі деректер базасы олардың ішкі көрінісімен автоматты түрде жұмыс істейді. Дегенмен, деректерді мәтіндікке қарағанда категориялық деп нақты анықтау кейбір артықшылықтарды ұсынады.

Деректер категориялық екенін білу бағдарламалық жасақтама жүйесіне график құру немесе модельді сәйкестендіру сияқты статистикалық процедуралардың қалай әрекет ету керектігін көрсете алады. Атап айтқанда, R және Python тілдерінде реттік деректер графиктерде, кестелерде және үлгілерде пайдаланушы анықтаған ретті сақтай отырып, реттелген.фактор ретінде ұсынылуы мүмкін.

Деректерді сақтау және индекстеуді оңтайландыруға болады (реляциялық дерекқордағы сияқты).

Белгілі бір категориялық айнымалы қабылдайтын мүмкін мәндер бағдарламалық жасақтамада жүзеге асырылады (мысалы, enum). Үшінші «артықшылық» күтпеген немесе күтпеген әрекетке әкелуі мүмкін: әдепкі бойынша, R ішіндегі деректерді импорттау функциялары (мысалы, read.csv) мәтіндік бағанды факторға автоматты түрде түрлендіретіндей әрекет етеді. Осы бағандағы келесі әрекеттер осы баған үшін жалғыз жарамды мәндер бастапқыда импортталған мәндер болып табылады және жаңа мәтіндік мән тағайындау ескерту мен NA (қолжетімсіз) мән хабарын береді деп болжайды.

Құрылымдық деректерге арналған негізгі идеялар •

Бағдарламалық жүйеде деректер әдетте түрі бойынша жіктеледі. •

Деректер түрі үздіксіз, дискретті, категориялық (оның ішінде екілік түрі бар) және реттік болуы мүмкін. •

Бағдарламалық жүйеде деректерді теру бағдарламалық жүйеге деректерді өңдеу жолын көрсетеді.

Төртбұрышты деректер

Деректер ғылымында талдауға арналған типтік қолдау құрылымы электрондық кесте немесе дерекқор кестесі сияқты төртбұрышты деректер нысаны болып табылады.

Негізгі терминдер

Деректер кадры (data frame) Тікбұрышты деректер (электрондық кестедегі сияқты) статистикалық және машиналық оқыту үлгілеріне арналған типтік деректер құрылымы болып табылады.

Мүмкіндік Кестедегі баған мүмкіндік деп аталады. Синонимдер: атрибут, енгізу, болжау, айнымалы.

Нәтиже Көптеген деректертану жобалары нәтижені болжауды қамтиды, көбінесе иә/жоқ пішімінде (мысалы, 1.1-кестеде бұл «Аукцион бәсекеге қабілетті болды ма, жоқ па?» деген сұраққа жауап). Белгілер кейде экспериментте немесе статистикалық зерттеуде нәтижені болжау үшін қолданылады. Синонимдер: тәуелді айнымалы, жауап, мақсат, шығыс.

Жазбалар Кестедегі жол жазба деп аталады. Синонимдер: жағдай, үлгі, прецедент, инстанция, бақылау, шаблон, үлгі, үлгі.

Тікбұрышты деректер негізінен жазбаларды (жағдайларды) көрсететін жолдар мен мүмкіндіктерді (айнымалыларды) көрсететін бағандары бар екі өлшемді матрица болып табылады. Бастапқыда деректер әрқашан бұл пішінде бола бермейді: құрылымдалмаған деректерді (мысалы, мәтінді) тікбұрышты деректердегі мүмкіндіктер жиынтығы ретінде көрсетуге болатындай етіп өңдеу және түрлендіру қажет («Құрылымдық деректер элементтері» тарауын қараңыз. «осы тараудың басында). Реляциялық дерекқорлардағы деректер көптеген аналитикалық және модельдеу тапсырмалары үшін шығарылып, бір кестеге орналастырылуы керек.

Таблица 1.1. Типичный формат данных

Категория	Валюта	Рейтинг продавца	Длитель-ность	Дань закры-тия	Цена закры-тия	Цена откры-тия	Конкурентно-способность?
Music/Movie/ Game	US	3249	5	Mon	0,01	0,01	0
Music/Movie/ Game	US	3249	5	Mon	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	0
Automotive	US	3115	7	Tue	0,01	0,01	1
Automotive	US	3115	7	Tue	0,01	0,01	1

Деректер жақтаулары және индекстер

Дәстүрлі дерекқор кестелерінде индекс деп аталатын бір немесе бірнеше бағандар бар. Ол белгілі бір SQL сұрауларының өнімділігін айтарлықтай жақсарта алады. Python-да pandas кітапханасын пайдаланған кезде, негізгі төртбұрышты деректер құрылымы деректер кестесін қамтитын DataFrame нысаны болып табылады. Әдепкі бойынша, жолдардың реті негізінде DataFrame үшін автоматты бүтін индекс жасалады. Pandas бағдарламалық құрал кітапханасында белгілі бір әрекеттердің тиімділігін арттыру үшін көп деңгейлі/иерархиялық индекстерді де көрсетуге болады. R тілінде негізгі төртбұрышты деректер құрылымы data.frame объектісі, деректер кадры болып табылады. data.frame нысанында жол ретіне негізделген жасырын бүтін индексі де бар. row.names1 төлсипаты арқылы R үшін түпнұсқалық теңшелетін кілт жасау мүмкін болса да, data.frame нысаны пайдаланушы анықтайтын немесе көп деңгейлі индекстерді қолдамайды. Бұл кемшілікті жою үшін екі жаңа бағдарламалық пакет кең тарады: data.table және dplyr. Екі бума да көп деңгейлі индекстерді қолдайды және data.frame нысанымен жұмыс істеуде айтарлықтай жылдамдықты қамтамасыз етеді.

Тікбұрышты емес деректер құрылымдары

Тікбұрышты мәліметтерден басқа, басқа да деректер құрылымдары бар. Уақыт сериясында бірдей белбеу өлшеулерінің дәйекті деректері бар. Бұл деректер статистикалық болжау әдістері үшін шикізат болып табылады және олар сонымен қатар құрылғылар шығаратын мәліметтердің негізгі құрамдас бөлігі болып табылады - Интернет заттары. Картографиялық және геокеңістіктік аналитикада қолданылатын кеңістіктік деректер құрылымдары тікбұрышты деректер құрылымдарына қарағанда күрделі және өзгермелі. Олардың Объектілік көрінісінде деректердің орталық бөлігі объект (мысалы, үй) және оның кеңістіктік координаттары болып табылады. Өріс проекциясында, керісінше, кеңістіктің кішкене бірліктеріне және тиісті метрикалық көрсеткіштің мәніне назар аударылады (мысалы, пиксель жарықтығы). Графикалық (немесе желілік) деректер құрылымдары физикалық, Әлеуметтік және дерексіз байланыстарды көрсету үшін

қолданылады. Мысалы, Facebook немесе LinkedIn сияқты әлеуметтік желінің графигі желідегі адамдар арасындағы байланысты білдіруі мүмкін. Жолдармен біріктірілген тарату орталықтары физикалық желінің мысалы болып табылады. Графикалық құрылымдар желіні оңтайландыру және ұсыныс жүйелері сияқты міндеттердің белгілі бір түрлерінде кеңінен қолданылады. Деректер ғылымында осы деректер түрлерінің әрқайсысының мамандандырылған әдістемесі бар. Бұл кітаптың басты бағыты — тікбұрышты мәліметтер-болжамды модельдеудегі негізгі құрылымдық элемент.

Орталық ережені бағалау

Өлшеу немесе сандық деректері бар айнымалылар мыңдаған түрлі мәндерге ие болуы мүмкін. Деректерді зерттеудің негізгі кезеңі әр белгі (айнымалы) үшін "типтік мәнді" алу болып табылады: деректердің көпшілігі қай жерде орналасқанын бағалау (яғни олардың орталық тенденциясы).