

Дәріс 7. Деректерді жинау және дайындау

Деректерді алу

Деректер көздерінің негізгі топтары

Файлдарды оқу

Деректермен жұмыс

Тазалау, түрлендіру, масштабтау

Data science — деректерден білім алу ғылымы.

Бұл саладағы деректерді зерттеу немесе модель құру процесі бірден машиналық оқытудан басталмайды. Деректерді жинау және дайындау әрдайым кез келген дерек жұмысындағы бірінші және ең маңызды қадам.

Деректердің сапасы барлық орындалатын модельдің тиімділігін анықтайды. Мысалы, жалған немесе толық емес деректерді пайдаланатын модель қанша күрделі болса да қате тұжырымдамалар береді.

Сондықтан деректер ғалымының уақыты аналитикадан гөрі дайындыққа көбірек жұмсалады. Деректер ғалымының уақыттарының 70-80% деректерді ашу, тазалау және өңдеумен өтеді. - IBM

Бұл — модель таңдаудан, тіпті нейрондық желіні оқытудан да маңызды.

Деректерді алу

Деректер – деректер ғылымы жобасының қаны. Қолдану салалары бойынша деректерді жинау: деректер базалары, әкімшілік жүйелер және жалпы хат алмасу құрылғылары, тиісті қызметтерге интерфейс жасау мүмкіндігімен.

Жоба мақсатына қарай деректер құрылымдалған, жартылай құрылымдалған немесе құрылымдалмаған болуы мүмкін.

Түрі	Мысалы	Сипаттамасы
Құрылымдалған	Excel, SQL кестелері	Бағандар мен жолдардан тұрады, түрлері(типтері) нақты көрсетілген.
Жартылай құрылымдалған	JSON, XML, CSV	Логикалық, бірақ кестелік емес деректерді қамтиды.
Құрылымдалмаған	Мәтін, сурет, видео, аудио	Алдын ала анықталған формат

Дерек көзі — кез келген деректі жинауға, сақтауға немесе өндіруге қабілетті деп есептелетін орта, жүйе немесе құрылым.

Деректер ғылымында дерек көзі — аналитикалық есептеулер, модель құру және болжау үшін шикі деректерді қамтамасыз ететін нәрсе. Бұл дерек көздері арқылы ғалым нақты құбылыстарды санмен сипаттай алады.

Деректер көздерінің негізгі топтары

Деректер ғылымында деректердің бөлімдері: деректер көздерінің негізгі топтары бар. Әрқайсысының өзінің атрибуттары, құрамы және дайындау әдісі бар.

Түрі	Сипаттамасы
Жергілікті файлдар(Local files)	Компьютерде немесе серверде сақталған кестелер немесе құжаттар, әдетте CSV, Excel, JSON, XML форматтарында. Мысалы: сату.csv, бағалар.xlsx, студенттер.json. Қолданылатын құралдар: pandas (read_csv, read_excel), NumPy.
Веб-сайттар және Web Scraping	Ашық вебтен деректер алу. Бұл тәсіл BeautifulSoup, Selenium, Scrapy сияқты кітапханаларды пайдалану арқылы жеңілдетіледі. Мысалы: температура деректерін беретін ауа-райы сайттары, мәтіндері ұсынылатын жаңалық порталдары, бағаларын көрсететін электрондық коммерция сайттары. Артықшылығы: жұмыс істеуге көп дерек бар. Кемшілігі: этикалық және заңды шектеулер.
API интерфейстері (Қолданба бағдарламалау интерфейсі)	Екі веб-қосымшаны немесе қызметті біріктірудің ортақ хаттамасы. Мысалы: Twitter API твиттерді талдауды қамтамасыз етеді, OpenWeather API арқылы ауа-райының деректері болуы мүмкін. Қолданылатын құралдар: requests, json, aiohttp. Артықшылығы: құрылымды, сенімді және ағымдағы деректер.

Ашық деректер базалары (Open Data)	Ғылыми, статистикалық және үкіметтік қоғамдық ақпарат көздері. Мысалы: Kaggle, UCI Machine Learning Repository, data.gov. Артықшылығы: оңай қол жетімді, тегін деректер жиынтықтары. Қолданылуы: зерттеу, білім беру, модельді тестілеу.
Сенсорлар және IoT (Заттар интернеті) құрылғылары	Құрылғылар, сенсорлар, камералар және "ақылды" жүйелермен жиналған ақпарат. Мысалы: температура сенсорлары, GPS координаттары, жүрек соғу жиілігін өлшейтін фитнес білезіктер. Дерек форматы: уақыт ретіндегі деректер. Артықшылығы: нақты уақыттағы деректер.
Кәсіпорын ішкі жүйелері	Ішкі коммерциялық деректер базалары, CRM жүйелері (Тұтынушылармен қарым-қатынасты басқару), ERP платформалары, бухгалтерлік бағдарламалық қамтамасыз ету және логистикалық технологиялар. Мысалы: тұтынушы сұранысы, қорлардың мөлшері, қаржылық көрсеткіштер. Қолданылатын құралдар: SQL, Power BI, Tableau. Артықшылығы: өте жоғары дәлдік және іскерлік контекст.

Файлдарды оқу

Файл – ақпаратты (мәтін, сан, сурет, дыбыс не кесте) ұзақ мерзімді сақтауға арналған контейнер. Файлдар компьютердің қатты дискісіне не бұлтқа сақталады.

Деректер ғылымындағы күнделікті тапсырмалардың бірі ретінде: файлдар жиі пайдаланылады, себебі көптеген деректер жиынтықтары осы форматтарда (CSV, Excel, JSON, XML және т.б.) келеді. Бұл файлдар талдамалық есептер, үлгілер, визуализация және машиналық оқыту үшін бастапқы дереккөз болып табылады.

Файл форматтарының түрлері келесі санаттарға бөлінеді:

Файл түрі	Үйлесімі	Құрылым	Қолдану аймағы
CSV (Ғылым үтірмен ажыратылған мәндер)	.csv	Кестелік (қатар-баған)	Жай деректер, статистикалық есептер
Excel (XLS/XLSX)	.xls, .xlsx	Ұяшықтар және беттер	Бизнес есептілік, бухгалтерлік деректер
JSON (JavaScript Object Notation)	.json	Негізгі мән жұптары	API, веб-қызметтер үшін пішім
XML (eXtensible Markup Language)	.xml	Тегтермен көрсетілген иерархия	Веб-сайттар, үкіметтік жүйелер
TXT (Мәтіндік файл)	.txt	Мәтіндік деректер	Журналдар, таза мәтін файлдары
Parquet / ORC	.parquet, .orc	Бағандық пішім	Үлкен деректер, Hadoop, Spark жүйелері

CSV файлдарымен жұмыс істеу:

Үтірмен ажыратылған мәндер (CSV) файл форматы - кестелер мен мәліметтер базалары үшін ең көп қолданылатын импорт және экспорт форматы болып табылады. Файлда қатарлар мен бағандар бар, мәндер үтір немесе нүктелі үтірмен бөлінген.

Мысалы:

students.csv:

name,age,grade

Nurasyl,19,85

Aiaulym,20,70

Diana,18,95

Python-та pandas арқылы CSV оқу:

```
import pandas as pd
```

```
df = pd.read_csv("students.csv")
print(df.head())
```

Артықшылықтар:

- Барлық жүйелермен үйлесімді;
- Жеңіл және оқуға оңай;
- Жайлы кестелік құрылым.

Excel файлдарын оқу:

Microsoft Excel - электрондық кесте құралы, онда қарапайым және күрделі есептеулерді калькуляторға қарағанда оңай жасауға болады. Бір файлда бір немесе бірнеше бет болуы мүмкін, ал ұяшық формула не мәнді сақтай алады.

Мысалы:

Python-та pandas арқылы Excel оқу:

```
import pandas as pd
```

```
grades = pd.read_excel("grades.xlsx", sheet_name="Sheet1")
print(grades.head())
```

Артықшылығы: формулалар мен пішімдер өзгеріссіз сақталады.

Кемшілігі: үлкен деректермен жұмыс істеу баяу.

JSON файлдары:

Веб-қосымшалар мен API арқылы алынған деректердің ең көп тараған пішімі JSON (JavaScript Object Notation) болып табылады. Ол негізгі мән жұптары ретінде ағаш түрінде сақталады. JSON артықшылықтары: ол адамдарға да, машиналарға да оқуға оңай, веб-қызметтердің жалпы пішімі.

Мысалы:

students.json:

```
[
  {"name": "Nurdaulet", "age": 20, "grade": 85},
```

```
{ "name": "Aqtoty", "age": 20, "grade": 80 }  
]
```

Python-та оқу:

```
import pandas as pd
```

```
data = pd.read_json("students.json")  
print(data.head())
```

XML файлдарын оқу

XML (eXtensible Markup Language) – деректер алмасу тілі. Ол тегтермен және иерархиялық қатынастарды көрсетумен ерекшеленеді.

Мысалы:

students.xml:

```
<students>  
  <student>  
    <name>Alikhan</name>  
    <age>19</age>  
    <grade>80</grade>  
  </student>  
</students>
```

Python-та оқу:

```
from bs4 import BeautifulSoup
```

```
with open("students.xml") as f:  
    soup = BeautifulSoup(f, "xml")
```

```
names = [n.text for n in soup.find_all("name")]  
print(names)
```

Деректермен жұмыс

Деректерді жұмыс / Деректерді өңдеу — бұл деректерді алу кезеңінен кейінгі маңызды қадам болып табылады, ол деректерді тазарту, түрлендіру, масштабтау, кодтау және айнымалылардың санын азайту операцияларын қамтиды.

Бұл қадамның мақсаты - деректерді талдау және дәл, тұрақты форматта модельдеу үшін түрлендіру.

Деректер ғылымында деректерді басқару «шикізат деректерінен білім алу» жұмыс процесінің маңызды құрамдас бөлігі болып табылады. Демек, алынған деректер шикі және оны тікелей талдау мүмкін емес; сондықтан оны өңдеу арқылы:

- қателіктер мен қарама-қайшылықтар жойылады;
- жоғалған жазбалар немесе қайталанулар реттеледі;
- екі тиісті деректер сәйкестігін бірдей диапазонға сәйкестендіреді;
- категориялық және сандық айнымалылар реттеледі;
- көптеген айнымалылардың өлшемдерін азайту жүзеге асырылады;
- және оларды талдауға және машина үйренуге дайын форматқа айналдырады.

Деректерді өңдеудің негізгі мақсаттары:

- Сапаны арттыру - қателіктер мен қарама-қайшылықтарды түзету арқылы сенімді деректерге қол жеткізу.
- Түсініктілік - деректерді адам мен машинаға оқуға ыңғайлы ету.
- Стандарттау - барлық деректер көздерінің біркелкі форматқа келтірілуі.
- Жылдамдық пен тиімділік - үлкен көлемдегі деректерді оңай өңдеу.
- Модельге дайындық - ерекшеліктер мен оқыту алдында бәрін алдын ала ламинаттау.

Тазалау, түрлендіру, масштабтау

Деректерді тазарту:

Деректерді тазарту дегеніміз деректер кестесіндегі қателерді, сәйкессіздіктерді және жетіспейтін мәндерді анықтау және өңдеу дегенді білдіреді.

Деректер әртүрлі дереккөздерден (файлдар, API, веб-тіркестіру, сенсорлар және т.б.) алынғанда, олардың сапасы көбіне нашар болады. Кездесетін жалпы мәселелерге мыналар жатады:

1. Жетпейтін мәндер - егер нақты ұяшықта енгізілген мәлімет болмаса, сол ұяшық жетпейтін мәнге ие деп аталады (мысалы, NaN, NULL, None).

Шешімдер: оны орташа мәнмен, медианамен, толтыру режимімен алмастыру немесе жою.

2. Қате терілген деректер - қате түрленген деректер көптеген жолдармен болуы мүмкін, мысалы: сандардың орнына мәтіндік мәндер, мәтіннің орнына жазылған сандар, күн мен санның орындары өзгертілген және т.б. Егер түрі дұрыс емес болса, бұл арифметикалық немесе уақыттық талдауға кедергі келтіруі мүмкін.

Шешім:

- `df["date"] = pd.to_datetime(df["date"])` #Оны керекті күн форматына өзгерту.
- `df["age"] = df["age"].astype(int)` #Оны бүтін сан түріне өзгерту.

3. Қайталаулар – бұл деректер жиынтығындағы бірдей бақылаулардың көшірмелері болып табылады. Олар бағалауларды бұрмалап, статистикалық қорытындылардың сенімділігін төмендетеді.

Шешімдер:

- `df = df.drop_duplicates()`
- немесе белгілі бағандарға: `df = df.drop_duplicates(subset=["student_id"])`

4. Ерекше мәндер - басқа мәндерден айтарлықтай ерекшеленетін мән. Мысалы, егер студенттердің жасы 18-25 болса және бір жазбада «120» деп айтса, бұл ерекше мән. Ерекше мәндер статистикалық және машинамен оқыту модельдерінің дәлдігін азайтады.

Жауаптар:

- Z-балл әдісі,
- IQR (аралық төрттік жоғарыдан ауытқу) әдісі
- визуализация (қораптық диаграмма)

5. Санаттық деректерді тазарту - көп жағдайда санаттық айнымалылар орфографиялық айырмашылықпен жазылады:

Қате: male, Male, M

Дұрыс:Male

Мұндай сәйкессіздіктерді `str.lower()` немесе `replace()` әдістерімен бірдей етуге болады.

6. Аномалиялар және логикалық сәйкессіздіктер - кейде деректер мәндері логикалық түрде сәйкессіз болуы мүмкін.

Мысалы:

- `age = -5`
- `salary = "NaN USD"`
- `height = 300 см`

Мұндай жағдайларда деректерді тазалау және сүзу қажет:

- `df = (df[df["age"] > 0] [df["age"] < 120])`

7. Деректерді стандарттау және нормалау - деректер әртүрлі бірліктерде/масштабтарда болуы мүмкін. Мысалы, "salary" доллармен, "баға" теңгемен, "жас" жылдармен. Бұл модельдеуге белгілі бір қолайсыздық туғызады, сондықтан деректердің бір-біріне сәйкес шкалаға келтірілген болуы керек.

- Min–Max нормалануы немесе z-балдық стандарттау осындай әдіс болуы мүмкін.

Деректерді түрлендіру:

Деректерді түрлендіру – бұл бастапқы деректерді талдауға, визуализацияға және машина оқыту үлгілеріне қолайлы форматқа өзгерту әдісі.

Мұндағы негізгі мақсат – әртүрлі дереккөздерден алынған деректерді бірдей, сандық және стандартталған форматқа түрлендіру.

- Алгоритмдермен оңай интерпретацияланатын формаға деректерді түрлендіру;
- Әртүрлі бірліктегі айнымалыларды (мәтін, сан, уақыт) үйлестіру;
- Олардың дерек ауқымдары мен шкалаларын теңдестіру үшін;
- Модельді тұрақты және дәл ету үшін.

Түрлендіру түрі	Мақсаты	Python үлгісі
Кодтау	мәтін айнымалы мәндерін қандай да бір санға өзгерту.	LabelEncoder, OneHotEncoder
Шкалалау	Ерекшеліктерді нормализациялау.	StandardScaler, MinMaxScaler, Normalizer
Логарифмдік және оненциалды түрлендіру операциялары	Ерекшеліктерді белгілі бір түрде түрлендіру.	log(), np.sqrt()
Екілік ету	Ерекшеліктерді 0 немесе 1 екілік мәндерге өзгерту.	Binarizer
Ерекшелік жасауы	Қосымша айнымалыларды жасап, үлгілердің сапасын арттыру.	Бағанды есептеу (мысалы, BMI)
Өлшемдерді қысқарту	Ең маңызды атрибуттарды сақтай отырып, деректер жинағындағы ерекшеліктер санын азайту.	PCA, LDA, t-SNE

Масштабтау:

Масштабтау - дегеніміз әртүрлі қасиеттерді бірдей масштабқа немесе ауқымға өзгерту.

Үлкен санды машина үйрену алгоритмдері (KNN, K-Means, градиенттік түсу, SVM және Нейрон желілері секілді) кіріс деректеріңіздің масштабына сезімтал. Яғни, үлкен масштабтағы қасиеттер модельдің шешіміне үлкен рөл ойнайды.

Масштабтау негізінен әрбір қасиеттің әсерін теңгеруге жеткізу үшін пайдаланылады, олардың шамасын салыстыруға жасап, модель әр қасиетке бірдей салмақ беретіндей етеді.

Мысал: Үздіксіз айнымалылардың басқа мүмкін мәндері "жас" (0–100) және "жалақы" (0–100000) болып табылады, сол кезде алгоритм автоматты түрде "жалақы" айнымалысын бірінші орынға қою үшін болады, модельде қолданылатын екі қасиеттің де қолданылуы кезінде. Оның шығару ауқымы әлдеқайда үлкен.

Қасиеттерді масштабтаудың теңестіру де орындалуда.

Әдіс атауы	Формула	Қолданылатын шарттар
Минимум–Максимум Масштабтау	$\left(x' = \frac{\{x - x_{\min}\}}{\{x_{\max} - x_{\min}\}} \right)$	Белгілі бір ауқымға нормализациялану үшін, о секілді кейбір алгоритмдерде қолданылады, мәндер осы алгоритмдерге енгізілуден бұрын нормализациялануы керек
Стандарттау (Z-есеп масштабтау)	$\left(z = \frac{\{x - \mu\}}{\{\sigma\}} \right)$	Қасиеттің таралуын норма таралу (Гаусс таралуы) жақындату үшін нормализациялау, өлшемдері әртүрлі өлшемдерді салыстыру кезінде қолданылады. Ол сондай-ақ алгоритмді тиімдірек етеді, өйткені барлық қасиеттер диапазоны эквивалентті болып табылады
Төзімді масштабтау	$\left(x' = \frac{\{x - median\}}{\{IQR\}} \right)$	Сыртқы шығарылатын мәндер көп болғанда

Стандарттау (Z-есеп Нормализациясы)

Стандарттау деректермен жұмыс жасағанда жиі және оңай қолданылады;

Әрбір қасиеттің мәндері орташа мәнге орталықталып, оның стандартты ауытқуына масштабталынады.

Формула төмендегідей:

$$z = x - \mu / \sigma$$

мұнда:

- x қасиеттің шынайы мәні
- μ қасиеттің орташа мәні
- σ мұнда қасиеттің стандартты ауытқуы
- z стандартталған (масштабталған) мәні

Әдебиеттер тізімі

1. Джоэл Грус. Python-мен алғышарттардан деректер ғылымы. O'Reilly Media, 2020.
2. Уэс МакКинни. Деректер талдау үшін Python: Pandas, NumPy және IPython арқылы деректерді өңдеу. O'Reilly, 2022.
3. Джон Хопкинс Университеті. Деректер ғылымында деректерді өңдеу. Coursera курсы. <https://www.coursera.org/learn/data-cleaning>