



Дәріс 7. Деректерді жинау және дайындау

Оқытушы: Сарсембаева Т.С.

Негізгі сұрақтар

Data Science жобасындағы
маңыздылығы

Неліктен деректерді жинау және
дайындау Data Science
жобасының ең маңызды бөлігі
болып саналады?

Деректердің түрлері

Деректердің негізгі түрлері
қандай?

Файлдарды оқу

Pandas көмегімен CSV және Excel
файлдарын оқудың негізгі
әдістері қандай?

Деректерді тазалау

Деректерді тазалау не үшін
қажет және оның негізгі
қадамдары қандай?

Деректерді масштабтау

Деректерді масштабтау
(Standardization) дегеніміз не және
ол не үшін қолданылады?

Деректерді жинау және дайындауға шолу

Data science — деректерден білім алу ғылымы. Деректерді жинау және дайындау кез келген дерек жұмысындағы бірінші және ең маңызды қадам болып табылады. Деректердің сапасы модельдің тиімділігін анықтайды. Деректер ғалымының уақытының 70-80% деректерді ашу, тазалау және өңдеумен өтеді.

Деректер түрлері

Деректер құрылымдалған (Excel, SQL), жартылай құрылымдалған (JSON, CSV) немесе құрылымдалмаған (мәтін, сурет) болуы мүмкін.

Негізгі дерек көздері

Жергілікті файлдар (CSV, Excel), Web Scraping, API интерфейстері, Ашық деректер базалары (Kaggle), IoT және Кәсіпорын ішкі жүйелері (CRM, ERP).

Файлдарды оқу үшін Pandas кітапханасы қолданылады: `pd.read_csv()`, `pd.read_excel()`, `pd.read_json()`.

Деректерді тазалау — деректер кестесіндегі қателерді, сәйкессіздіктерді және жетіспейтін мәндерді анықтау және өңдеу процесі. Бұған жетіспейтін мәндерді (NaN) өңдеу, қайталануларды жою (`df.drop_duplicates()`), ерекше мәндерді (Outliers) анықтау және санаттық деректерді біріздендіру кіреді.

Деректерді түрлендіру және масштабтау — деректерді талдауға қолайлы форматқа өзгерту (мәтінді санға кодтау) және әртүрлі қасиеттерді бірдей масштабқа келтіру. Бұл KNN, K-Means, SVM және Нейрон желілері сияқты алгоритмдер үшін маңызды. Стандарттау (Z-есеп) ең жиі қолданылатын әдіс.

Кіріспе: Деректер ғылымындағы деректерді дайындаудың маңыздылығы

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 және оның түрлері

Деректер – деректер ғылымы жобасының қаны. Жоба мақсатына қарай деректер құрылымдалған, жартылай құрылымдалған немесе құрылымдалмаған болуы мүмкін.

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

Дерек көзі — кез келген деректі жинауға, сақтауға немесе өндіруге қабілетті орта, жүйе немесе құрылым. Деректер ғылымында дерек көзі — аналитикалық есептеулер, модель құру және болжау үшін шикі деректерді қамтамасыз ететін нәрсе.

Деректер көздерінің негізгі топтары

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

1

Жергілікті файлдар

Компьютерде немесе серверде сақталған кестелер немесе құжаттар (CSV, Excel, JSON). Құралдар: pandas, NumPy.

2

Веб-сайттар және **Web Scraping**

Ашық вебтен деректер алу (BeautifulSoup, Selenium). Мысалы: ауа-райы, жаңалықтар. Артықшылығы: көп дерек. Кемшілігі: этикалық шектеулер.

3

API интерфейстері

Екі веб-қосымшаны біріктіру хаттамасы (Twitter API, OpenWeather API). Құралдар: requests, json. Артықшылығы: құрылымды, сенімді деректер.

4

Ашық деректер базалары

Ғылыми, статистикалық және үкіметтік ақпарат көздері (Kaggle, UCI). Артықшылығы: оңай қол жетімді, тегін деректер.

5

Сенсорлар және **IoT**

Құрылғылар, сенсорлар, камералармен жиналған ақпарат (температура сенсорлары, GPS). Дерек форматы: уақыт ретіндегі деректер. Артықшылығы: нақты уақыттағы деректер.

6

Кәсіпорын ішкі жүйелері

Ішкі коммерциялық деректер базалары (CRM, ERP). Құралдар: SQL, Power BI. Артықшылығы: өте жоғары дәлдік және іскерлік контекст.

Деректермен жұмыс: Тазалау, Түрлендіру, Масштабтау

Деректерді өңдеу — деректерді алу кезеңінен кейінгі маңызды қадам. Ол деректерді тазарту, түрлендіру, масштабтау, кодтау және айнымалылардың санын азайту операцияларын қамтиды. 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. Шикі деректерді 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.
- Оларды талдауға және машина үйренуге дайын форматқа айналдырады.

Сапаны арттыру

Қателіктерді түзету арқылы сенімді деректерге қол жеткізу.

Түсініктілік

Деректерді адам мен машинаға оқуға ыңғайлы ету.

Стандарттау

Барлық деректер көздерінің біркелкі форматқа келтірілуі.

Модельге дайындық

Ерекшеліктер мен оқыту алдында бәрін алдын ала ламинаттау.

Деректерді тазарту — деректер кестесіндегі қателерді, сәйкессіздіктерді және жетіспейтін мәндерді анықтау және өңдеу. Деректер әртүрлі дереккөздерден алынғанда, олардың сапасы көбіне нашар болады.

Деректерді масштабтау және стандарттау

Масштабтау — әртүрлі қасиеттерді бірдей масштабқа немесе ауқымға өзгерту. KNN, K-Means, SVM және Нейрон желілері сияқты алгоритмдер кіріс деректерінің масштабына сезімтал. Масштабтау әрбір қасиеттің әсерін теңгеруге, олардың шамасын салыстыруға және модельдің әр қасиетке бірдей салмақ беруіне мүмкіндік береді.

Мысал: "жас" (0–100) және "жалақы" (0–100000) сияқты айнымалылар болғанда, алгоритм "жалақы" айнымалысын бірінші орынға қоюы мүмкін, себебі оның ауқымы үлкен.

Минимум–Максимум Масштабтау	$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$	Белгілі бір ауқымға нормализациялау үшін, SVD сияқты алгоритмдерде қолданылады.
Стандарттау (Z-есеп масштабтау)	$z = \frac{x - \mu}{\sigma}$	Қасиеттің таралуын норма таралуға жақындату, әртүрлі өлшемдерді салыстыру үшін қажет.
Төзімді масштабтау	$x' = \frac{x - \text{median}}{\text{IQR}}$	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.

$$z = \frac{x - \mu}{\sigma}$$

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

Дәрісті меңгеруге арналған бақылау сұрақтары

→ Деректерді дайындау уақыты

Неліктен деректерді дайындау процесі (70-80%) модельді таңдаудан да көп уақыт алады?

→ Деректер түрлерінің айырмашылығы

Құрылымдалған (SQL) және жартылай құрылымдалған (JSON) деректердің айырмашылығы неде?

→ **API** қолданысы

API дегеніміз не және ол деректер алу үшін қалай қолданылады?

→ Жетіспейтін мәндерді өңдеу

Деректерді тазалау кезінде "жетіспейтін мәндерді" (NaN) өңдеудің екі әдісін атаңыз.

→ Масштабтау қажеттілігі

"Жас" (0-100) және "жалақы" (0-100000) сияқты екі түрлі айнымалыны неліктен масштабтау (Standardization) қажет?



Әдебиеттер тізімі

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