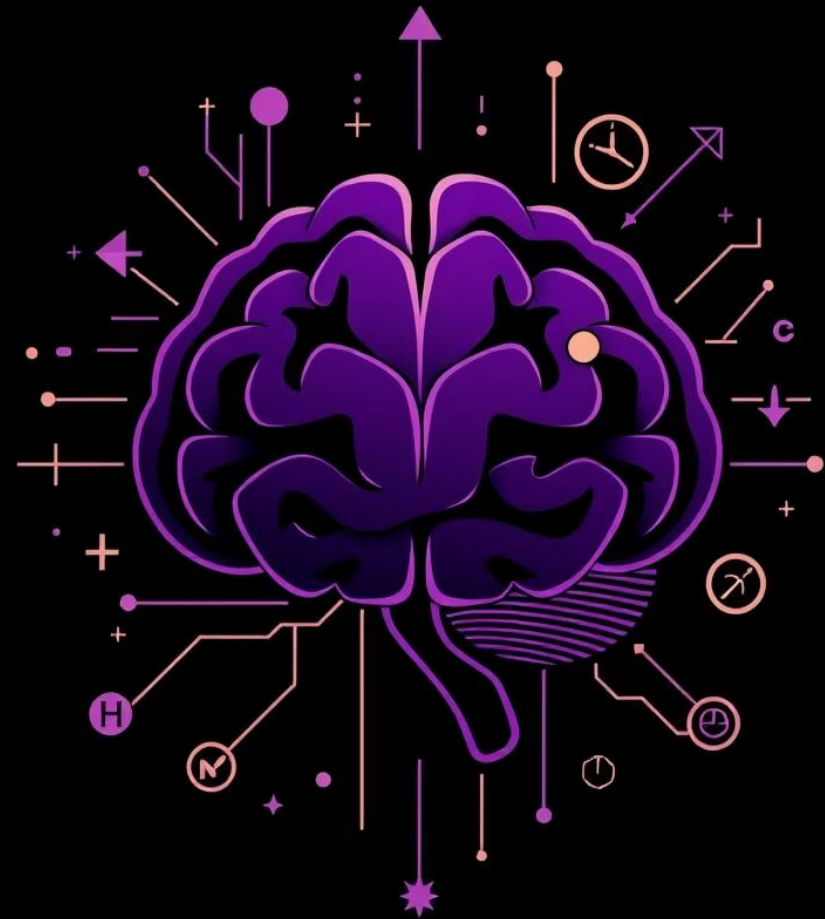


Дәріс 4. Сызықтық алгебра және статистика: Деректер ғылымының негіздері

Бұл дәріс деректер ғылымындағы сызықтық алгебра мен статистиканың маңызын түсіндіруге арналған. Біз векторлар мен матрицалар сияқты негізгі элементтерді, сондай-ақ орталық тенденция мен деректердің өзгергіштігін талдауды қарастырамыз.

Оқытушы: Сарсембаева Т.С.





Дәрістің мақсаты

Сызықтық алгебраның маңызы

Деректерді құрылымдау және сандық бағалау құралы ретінде сызықтық алгебраның рөлін түсіндіру.

Векторлар мен матрицалар

Векторлық операцияларды (қосу, скалярға көбейту, ұзындығын табу) және матрицалардың маңызын меңгеру.

Статистиканың рөлі

Деректерді зерттеу ғылымы ретінде статистиканың маңыздылығын анықтау.

Деректерді талдау

Орталық тенденция көрсеткіштерін (орташа, медиана, мода) және деректердің өзгергіштігін (вариация, стандартты ауытқу) талдауды үйрену.

Негізгі сұрақтарға шолу

1

Сызықтық алгебраның рөлі

Деректер ғылымындағы сызықтық алгебраның рөлі қандай?

2

Векторлар мен матрицалар

Векторлар мен матрицалар дегеніміз не және олар деректерді қалай ұсынады?

3

Орталық тенденция

Орташа, медиана және мода көрсеткіштерінің айырмашылығы неде?

4

Вариация

Вариация (дисперсия) деректердің қандай сипаттамасын көрсетеді?

5

Корреляция және себеп-салдарлық

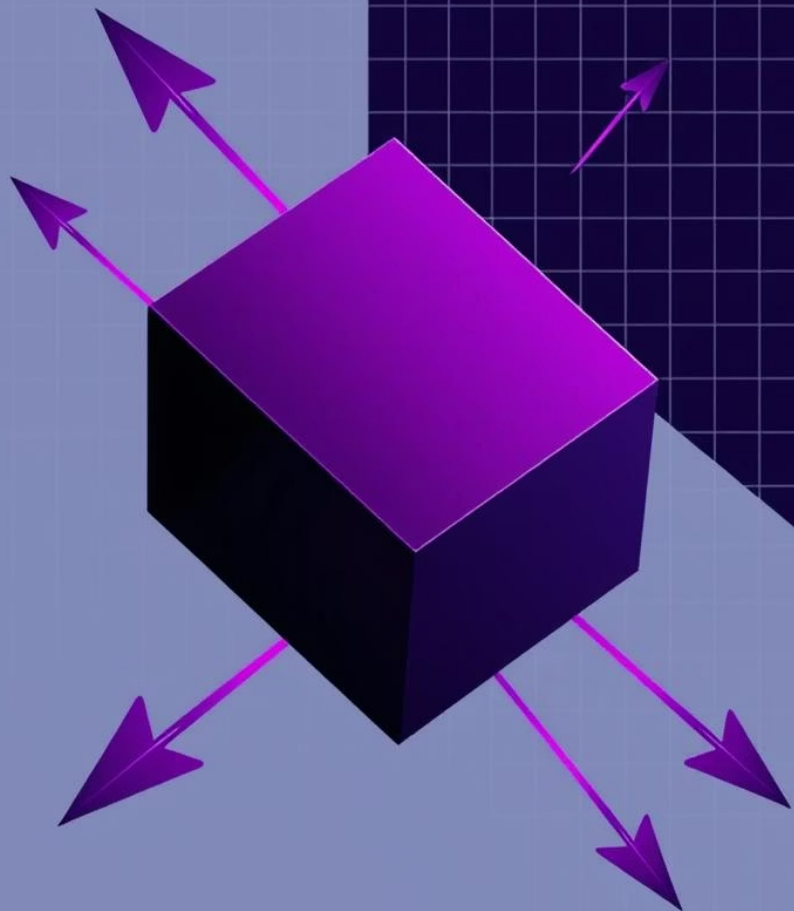
Корреляция мен себеп-салдарлық байланыстың айырмашылығы неде?

Сызықтық алгебра: Деректерді құрылымдау

Сызықтық алгебра – деректерді өңдеуге және талдауға қолданылатын негізгі бағдарлама. Ол ақпаратты құрылымдауға мүмкіндік береді. Деректер жиынтығын матрицалар мен векторлар арқылы жазуға болады.

Машиналық оқытуды үй деп есептесек, математика мен статистика оның құрылыс материалдары болып табылады. Сызықтық алгебра машиналық оқытудағы ең маңызды математикалық дағдылардың бірі.

Күнделікті машиналық оқыту модельдері жиі матрицалармен сипатталады. Деректерді алдын ала өңдеу, түрлендіру және модельді талдау сызықтық алгебраны қолданады.



Векторлар: Бағытталған шамалар

1

Анықтамасы

Векторлар - бағыты мен ұзындығы (модулі) бар бағытталған шамалар. Оларды қосуға және скалярға көбейтуге болады.

2

Математикалық жазу

Векторды сандар тізбегі ретінде жазуға болады: $v=(u_1,u_2,...,u_n) \in R^n$, мұндағы n - өлшемдер саны.

3

Геометриялық түсінік

Екі өлшемді кеңістікте вектор - координаталардың басынан (0,0) белгілі бір нүктеге бағытталған көрсеткі.

4

Ұзындығы (нормасы)

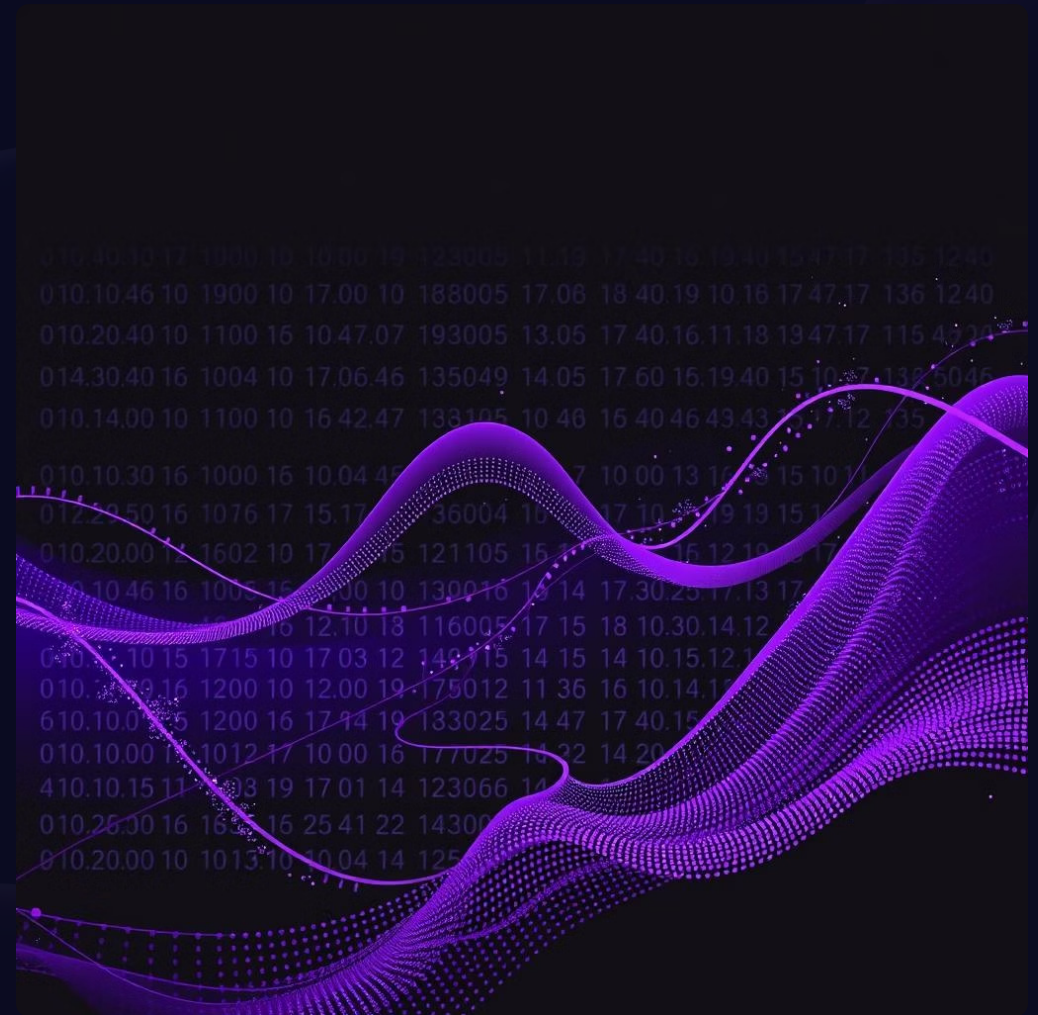
Вектордың ұзындығы Пифагор теоремасы бойынша есептеледі: $\|v\| = \sqrt{v_1^2+v_2^2+\dots+v_n^2}$.

Матрицалар: Деректерді ұсыну және түрлендіру

Матрицалар - екі өлшемді сандар жиынтығы (жолдар мен бағандар). Олар деректер ғылымы мен математикалық модельдеудің негізгі құралдарының бірі.

:Матрицалардың маңызы

- **Деректерді ұсыну:** Көптеген векторлардан тұратын мәліметтер жиынтығын ұсыну. Мысалы, 1000 адамның бойы, салмағы және жасы туралы мәліметтерді (1000x3) матрицасы ретінде сақтау.
- **Сызықтық түрлендіру:** k-өлшемді векторды m-өлшемді векторға түрлендіретін функция ретінде қызмет етеді. Бұл машиналық оқытудың негізгі бөлігі.
- **Байланысты көрсету:** Объектілер арасындағы қатынастарды көрсету үшін (мысалы, графтардағы көршілес матрица).



Статистика: Деректерді талдау ғылымы

Статистика – деректерді жинау, талдау, интерпретациялау және ұсыну ғылымы. Ол түсініктерді табады және құбылыстар арасындағы байланысты табуға үлес қосады. Статистикалық талдау арқылы біз айырмашылықты, өзгерісті және айнымалылар арасындағы қатынасты табамыз.

Деректерді жинау

Деректерді жүйелі түрде жинау және ұйымдастыру.

Талдау және интерпретация

Деректерден мағыналы ақпарат алу және оны түсіндіру.

Ұсыну

Талдау нәтижелерін түсінікті және тиімді түрде жеткізу.

Орталық тенденция көрсеткіштері

Деректермен жұмыс жасағанда, біз орталық тенденцияны сипаттайтын үш негізгі көрсеткішті қолданамыз: орташа, медиана және мода.



Орташа (Mean)

Барлық элементтердің қосындысын олардың санына бөлу. Экстремалды мәндерге өте сезімтал.

Мысал: [70, 80, 90] → Орташа = 80



Медиана (Median)

Деректерді өсу ретімен орналастырған кезде ортасында болатын мән. Экстремалды мәндерге төзімді.

Мысал: [2, 3, 4, 9, 10] → Медиана = 4



Мода (Mode)

Деректер жиынтығындағы ең көп таралған мән. Көбінесе категориялық деректерге қолданылады.

Мысал: [2, 4, 4, 5, 6, 6, 6, 7] → Мода = 6

Вариация және байланыстар

Вариация - деректердің өзгергіштігін, яғни әрбір мәннің орташа мәннен ауытқу дәрежесін сипаттайды. Егер мәндер бір-біріне жақын болса, вариация аз, ал кең ауқымда болса, жоғары болады.

Дисперсия (Variance)

Орташа квадраттық ауытқудың өлшемі.

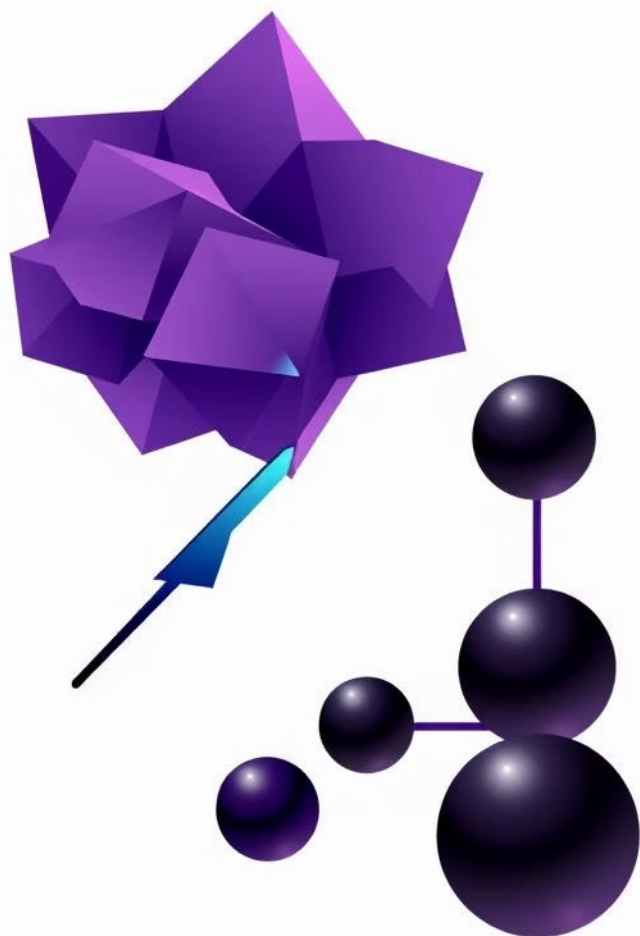
$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

Стандартты ауытқу (Standard deviation)

Дисперсияның квадрат түбірі.

$$\sigma = \sqrt{\frac{\sum (x_i - \mu)^2}{N}}$$





Корреляция және себеп-салдарлық байланыс

Корреляция

Екі айнымалы арасындағы байланыс деңгейін көрсетеді. Коэффициент (r) -1-ден +1-ге дейін болады.



- $r = +1$: Күшті оң байланыс
- $r = -1$: Күшті теріс байланыс
- $r \approx 0$: Жоқ немесе әлсіз байланыс

Маңызды ескерту

Корреляция тек байланыс бар екенін көрсетеді, бірақ себеп-салдарлық байланысты дәлелдемейді.



Мысал: Балмұздақ сатылымы мен балалардың көбеюі арасында корреляция болуы мүмкін, бірақ екеуінің де себебі – жаз мезгілі.

Себеп-салдарлық байланыс

Бір айнымалының екіншісіне тікелей әсерін білдіреді.



- **Мысал:** Күн сәулесінің жоғарылауы → температураның жоғарылауы.
- **Мысал:** Су ішу → денені ылғалдандыру.

Дәрісті меңгеруге арналған бақылау сұрақтары:

Вектор дегеніміз не және оның матрицадан айырмашылығы қандай?

Матрицалар деректер ғылымында қандай үш негізгі мақсатта қолданылады?

Деректер жиынтығында қалыптан тыс (экстремалды) мәндер болса, орталық тенденцияны сипаттау үшін "орташа" (mean) немесе "медиана" (median) көрсеткіштерінің қайсысын қолданған дұрыс және неліктен?

Стандартты ауытқу (Standard deviation) дегеніміз не және ол деректер туралы қандай ақпарат береді?

"Корреляция себеп-салдарлық байланысты білдірмейді" дегенді мысалмен түсіндіріңіз.

Әдебиеттер тізімі:

ScienceHunter. (2023, 18 сәуір). Статистические типы данных, используемые в машинном обучении. ScienceHunter. Қолжетімді: <https://www.sciencehunter.net/Blog/story/statisticheskie-tipyi-dannyih-ispolzuemyie-v-mashinnom-obuchenii>

SkillFactory Blog. (2024, 27 наурыз). Линейная алгебра для Data Science — 10 примеров практического использования. Қолжетімді: <https://blog.skillfactory.ru/linejnaya-algebra-dlya-data-science-10-primerov-prakticheskogo-ispolzovaniya/>

Грас, Дж. (2020). Data Science. Наука о данных с нуля. Қолжетімді: https://www.ursmu.ru/userfiles/media/default/sveden/education/1.2.1_iimo-21_mm_kaf.inf.pdf