

Дәріс 4. Сызықтық алгебра және статистика

Сызықтық алгебра:

Векторлар мен матрицалар

Статистика

Орташа, медиана, мода

Вариация, корреляция, себеп-салдарлық байланыстар

Сызықтық алгебра - бұл деректер ғылымындағы маңызды математикалық салалардың бірі. Ол деректерді өңдеуге және талдауға қолданылатын негізгі бағдарлама болып табылады. Басқаша айтқанда, сызықтық алгебра ақпаратты құрылымдауға және оны сандық бағалауға мүмкіндік береді.

Деректер жиынтығы – сандармен толтырылған матрицалар мен векторлар арқылы жазылған математикалық түсінік. Сызықтық алгебра осы матрицаларды өңдеуге, олардың қасиеттерін талдауға және өзара қалай байланысты екенін табуға мүмкіндік береді.

Ал статистика – деректерді зерттеу ғылымы. Бұл білім деректерді жинау тәсілдерін таңдау және мұндай талдаудың нәтижелерін пайдаланудағы статистикалық процедураларды анықтауда қажет. Статистика маңызды ғылымдардың бірі болып табылады, ол үлкен деректерді талдауды, шешім қабылдауды және көптеген ғылыми жаңалықтардың іргетасын құрайды. Ол мәселені терең зерттеуді мүмкін етеді.

Статистика деректер талдауының теориясы мен әдістемесінің негізгі тілі болып табылады. Егер машина оқытуы үй ретінде қарастырылса, онда математика мен статистика құрылыс материалдары болар еді. Әртүрлі алгоритмдерді қашан пайдалану керек екенін үйрену үшін, олардың негізінде қандай әдістер бар екенін білу қажет.



Сызықтық алгебра

Сызықтық алгебра - математиканың деректер ғылымы мен машиналық оқытудағы өте пайдалы саласы. Сызықтық алгебра машиналық оқытуда ең маңызды математика дағдылары болып табылады. Күнделікті машиналық оқыту модельдері жиі матрицалармен сипатталады. Өзінің бастапқы түрінде деректер жиі матрица ретінде ұйымдастырылған. Деректерді алдын ала өңдеу, түрлендіру және модельді талдау сызықтық алгебраны қолданады. Сызықтық алгебраны түсіну үшін таныс болуы керек тақырыптар: векторлар және матрицалар.

Векторлар мен матрицалар

Абстракттілі мағынада векторлар - бұл жаңа векторларды қалыптастыру үшін бір-бірін арасында қосуға болатын объектілер және скалярлық шамаларға көбейту арқылы (яғни сандарға), қайтадан жаңа векторларды қалыптастыруға арналған объектілер. Басқаша айтқанда, вектор дегеніміз екі негізгі сипаттамаға ие бағытталған шама: бағыт және ұзындық (модуль).

Математикалық тұрғыда векторды сандар тізбегі ретінде жазуға болады:

$$v=(u_1,u_2,\dots,u_n)\in R^{*n}$$

мұндағы n - өлшемдер саны, ал әрбір u_i — вектордың координатасы.

Геометриялық түсінігі

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

Мысалға, $v=(3,2)$ векторы x осінде 3 бірлікке, y осінде бұл 2 бірліктің ығысуын білдіреді.

Үш өлшемде $v=(x,y,z)$ - нүктенің кеңістіктегі орнын немесе қозғалыс бағытын сипаттайды.

Python тілінде векторларды NumPy кітапханасы арқылы жасауға болады:

```
import numpy as np
v = np.array([2, 4, 6]) # Вектор мысалы
print("Вектор:", v)
```

Векторлардың негізгі қасиеттері мен операциялары:

Векторларды қосу

Бірдей өлшемдегі екі вектордың қосындысы олардың сәйкес координаталарын қосу арқылы табылады:

$$a+b=(a_1+b_1,a_2+b_2,\dots,a_n+b_n)$$

$$\text{Мысал: } (2,3)+(4,-1)=(6,2)$$

Скаляр бойынша көбейту

Векторды санға (скалярға) көбейткенде оның бағыты өзгермейді (егер сан оң болса) және оның ұзындығы осы санға артады:

$$cv=(cv_1,cv_2,\dots,cv_n)$$

$$\text{Мысал: } 2 \times (3,4)=(6,8)$$

Нөлдік вектор

Барлық координаттары нөлге тең вектор:

$$0=(0,0,\dots,0)$$

Ол нөлдік ұзындықтағы векторды ешқандай бағытсыз бейнелейді.

Вектордың ұзындығы (модулі)

Вектордың ұзындығы немесе нормасы Пифагор теоремасы бойынша есептеледі:

$$\|v\|=\sqrt{v_1^2+v_2^2+\dots+v_n^2}$$

$$\text{Мысал: } \|(3,4)\|=\sqrt{3^2+4^2}=5$$

Python мысалы (NumPy арқылы)

```
import numpy as np
# Векторды анықтау
v = np.array([3, 4])
w = np.array([1, -2])
sum_vw = v + w # Векторларды қосу
scaled_v = 2 * v # Скалярға көбейту
norm_v = np.linalg.norm(v) # Норма (ұзындығы)

print("Қосынды:", sum_vw)
print("Масштабталған вектор:", scaled_v)
print("Ұзындығы:", norm_v)
```

Нәтиже:

Қосынды: [4 2]

Масштабталған вектор: [6 8]

Ұзындығы: 5.0

Матрица - екі өлшемді сандар жиынтығы. Біз матрицаларды тізімдер тізімі (list of lists) ретінде енгіземіз, мұнда әрбір ішкі тізім бір жолды білдіреді және өлшемі бірдей.

Егер A матрица болса, онда $A[i][j]$ i - жолдағы және j - бағандағы элемент болып табылады.

Математикалық жазу ережелеріне сәйкес матрицаларды белгілеу үшін әдетте бас әріптер қолданылады.

$$A = \begin{bmatrix} 1, & 2, & 3 \\ 4, & 5, & 6 \end{bmatrix}$$

Бұл 2 жолдан және 3 бағаннан тұратын A матрицасы. (2×3)

$$B = \begin{bmatrix} 1, & 2 \\ 3, & 4 \\ 5, & 6 \end{bmatrix}$$

Бұл 3 жолдан және 2 бағаннан тұратын B матрицасы. (3×2)

Жолдар (rows): көлденең бағыттағы сандар тізбегі.

Бағандар (бағандар): тік бағыттағы сандар тізбегі.

Матрицалардың маңызына тоқталатын болсақ, матрицалар деректер ғылымы мен математикалық модельдеудің негізгі құралдарының бірі болып табылады. Олар біз үшін бірнеше себептерге байланысты өте маңызды:

- Біріншіден деректерді ұсыну үшін маңызды.

Матрицаны көптеген векторлардан тұратын мәліметтер жиынтығын ұсыну тәсілі ретінде қарастыруға болады.

Әр вектор матрицаның бір жолын білдіреді.

Мысалы, егер бізде 1000 адамның бойы, салмағы және жасы туралы мәліметтер болса, біз оны келесідей білдіре аламыз: $data = \begin{bmatrix} 170, & 70, & 40, \\ 165, & 60, & 26, \\ 177, & 80, & 19, \\ \dots \end{bmatrix}$

Бұл жерде әр жол жеке адамның мәліметін берсе, әр баған белгілі бір сипаттаманы (бой, салмақ, жасы) көрсетеді.

- Екіншіден сызықтық түрлендіру ретінде маңызды.

$(m \times k)$ – бұл сызықтық функция, ол k өлшемді векторларды m өлшемді векторларға түрлендіретін

Басқаша айтқанда, егер сізде вектор болса

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

және матрица

$$A = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

[456]]

болса, онда көбейту нәтижесі $A \times x$ жаңа вектор береді:

$$A \times x = \begin{bmatrix} 1 \cdot x_1 + 2 \cdot x_2 + 3 \cdot x_3 \\ 4 \cdot x_1 + 5 \cdot x_2 + 6 \cdot x_3 \end{bmatrix}$$

Бұл мысалда 3 өлшемді вектор $(x) \rightarrow 2$ өлшемді векторға түрлендірілген.

Мұндай түрлендірулер сызықтық алгебра мен машиналық оқытудың негізгі бөлігі болып табылады, өйткені олардың көмегімен модельдер мен есептеулер жасалады.

- Үшіншіден байланысты көрсету үшін (графтар мен желілер).

Матрицалар тек сандар жиынтығы ғана емес, сонымен қатар олар объектілер арасындағы қатынастарды да көрсете алады.

Мысалы, егер бізде желі (график) болса, онда оның түйіндері (nodes) және беттері (edges) бар. Әр жіп екі түйінді байланыстырады: (i, j)

Мұндай байланысты көршілес матрица (adjacency matrix) ретінде көрсетуге болады.

Мысалы:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{bmatrix}$$

Мұнда:

$A[i][j]=1$, Егер i және j түйіндері бір бірімен байланысты болса;

$A[i][j]=0$, егер байланыс болмаса.

Осылайша, матрицалар әлеуметтік медианы, интернет байланысын, логистикалық маршруттарды және көптеген күрделі жүйелерді модельдеу үшін қолданылады.

Статистика

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. Статистика деректер жиынтығынан мағына құрастыру үшін математика мен процестерді қолданады.

Орташа, медиана, мода

Деректермен жұмыс жасаған уақытта біз орташа, медиана және мода деген көрсеткіштермен жұмыс жасаймыз. Бұл деректердің орталық тенденциясын сипаттайтын үш негізгі көрсеткіш.

Орташа (Mean)

Орташа мән барлық элементтердің қосындысын олардың санына бөлу арқылы табылады:

$$\bar{x} = (x_1 + x_2 + \dots + x_n) / n$$

Мысал:

Баға тізімі: 70, 80, 90

$$\text{Орташа} = (70 + 80 + 90) / 3 = 80$$

Орташа мән (mean) жалпы тенденцияны сипаттайды, бірақ экстремалды (қалыптан тыс) мәндерге өте сезімтал болып келеді.

Медиана (Median)

Медиана - бұл деректерді өсу ретімен орналастырған кезде ортасында болатын мән.

Егер элементтер саны жұп болса, медиана центрдегі екі мәnniң орташа саны болып табылады.

Мысал:

$$\text{Тізім: } [2, 3, 4, 9, 10] \rightarrow \text{Медиана} = 4$$

$$\text{Тізім: } [2, 3, 4, 6, 9, 10] \rightarrow \text{Медиана} = (4 + 6) / 2 = 5$$

Медиана экстремалды мәндерге төзімді және "орташа" деректерді дәлірек сипаттайды.

Мода (Mode)

Мода - деректер жиынтығындағы ең көп таралған мән.

Мысал:

$$[2, 4, 4, 5, 6, 6, 6, 7] \rightarrow \text{Сән} = 6$$

Мода көбінесе сапалы немесе категориялық деректерге қолданылады (мысалы, ең танымал бренд, көбінесе таңдалған жауап).

Вариация, корреляция, себеп-салдарлық байланыстар

Вариация - бұл деректердің өзгергіштігін сипаттайтын статистикалық көрсеткіш, яғни әрбір мәннің орташа мәннен ауытқу дәрежесі.

Басқаша айтқанда, вариациялар деректердің қаншалықты "біркелкі" немесе "әртүрлі" екенін көрсетеді.

Егер барлық мәндер бір-біріне өте жақын болса - вариация аз, ал егер мәндер кең ауқымда бөлінсе - вариация жоғары болады.

Ал оны есептеу үшін Стандартты ауытқу (Standard deviation) табу керек “ σ ”. Стандартты ауытқуды шығармастан бұрын біз дисперсияны есептеп алуымыз керек “ σ^2 ”.

Дисперсия (Variance) - орташа квадраттық ауытқудың өлшемі:

$$\sigma^2 = \frac{\sum (x - \bar{x})^2}{n}$$

Стандартты ауытқу (Standard deviation) - дисперсияның квадрат түбірі:

$$\delta = \sqrt{\frac{\sum (x - \bar{x})^2}{n}};$$

Корреляция - екі айнымалы арасындағы байланыс деңгейін көрсетеді.

Корреляция коэффициенті (r) -1-ден +1-ге дейін

$r = +1 \rightarrow$ күшті оң байланыс (біреуі ұлғайған кезде, екіншісі де артады)

$r = -1 \rightarrow$ күшті теріс байланыс (біреуі өскенде, екіншісі азаяды)

$r \approx 0 \rightarrow$ жоқ немесе әлсіз байланыс

Мысал:

Балмұздақтың температурасы мен сатылымы арасындағы байланыс $\rightarrow R \approx +0.9$

Күнделікті ұйқы уақыты және шаршау деңгейі $\rightarrow R \approx -0.7$

Маңызды: корреляция-тек байланыс бар екенін көрсетеді, бірақ себеп-салдарлық байланысты дәлелдемейді.

Себеп-салдарлық байланыс - бір айнымалының екіншісіне тікелей әсерін білдіреді.

Мысалы:

Күн сәулесінің жоғарылауы → температураның жоғарылауы.

Су ішу → денені ылғалдандыру.

Корреляция болғанымен, ол міндетті түрде себеп-салдар байланысы бола алмайды.

Мысалы:

Егер балмұздақ сатылымы бір уақытта өссе, "балалар балмұздақ сатылымын көбейтеді" деген тұжырым жасау қате, өйткені нақты себеп — жаз.

Әдебиеттер тізімі:

1. 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>
2. 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/>
3. Грас, Дж. (2020). Data Science. Наука о данных с нуля. Қолжетімді: https://www.ursmu.ru/userfiles/media/default/sveden/education/1.2.1_iiimo-21_mm_kaf.inf.pdf