

Деректер Өлшемділігін Төмендету: Негізгі Компоненттер Әдісі (PCA)

Бұл презентацияда біз машиналық оқытудағы деректерді өңдеудің маңызды әдісі — Негізгі Компоненттер Әдісін (PCA) қарастырамыз. PCA – деректердің өлшемін азайтуға арналған күшті статистикалық құрал, ол күрделі деректерді оңайлатуға және талдауға ыңғайлы етуге мүмкіндік береді.

Оқытушы: Сарсембаева Т.С.

РСА Ұғымы мен Мақсаты

Ұғымы

Негізгі Компоненттер Әдісі (РСА) – бұл деректердің өлшемін (*dimensionality*) азайтуға арналған статистикалық әдіс. Ол үлкен деректер жинақтарындағы ең маңызды ақпаратты сақтай отырып, оларды қарапайым түрге келтіреді.

Мақсаты

Деректердің негізгі ақпаратын (дисперсиясын) барынша сақтай отырып, бастапқы айнымалыларды жаңа, тәуелсіз айнымалыларға (негізгі компоненттерге) түрлендіру.

Өлшемділіктің Қарғысы (Curse of Dimensionality)

Өлшемділікті төмендетудің қажеттілігі деректер жиынтығындағы айнымалылар санының шектен тыс көбеюінен туындайды. **Өлшемділіктің қарғысы** деп аталатын бұл құбылыс келесі мәселелерге әкеледі:

- **Талдаудың қиындауы:** Айнымалылар саны артқан сайын деректерді түсіну және талдау күрделене түседі.
- **Есептеу уақытының артуы:** Деректер көлемі үлкейген сайын алгоритмдерді орындауға кететін уақыт та едәуір артады.
- **Модельдердің артық оқытылуы (overfitting):** Көптеген айнымалылар модельдің оқу деректеріне тым жақсы бейімделуіне әкеліп, жаңа деректерде нашар нәтиже беруі мүмкін.

PCA осы мәселені шешудің тиімді жолы, деректерді маңызды ақпаратты жоғалтпай ықшамдауға көмектеседі.

РСА-ның Негізгі Артықшылықтары және Қолданылуы



Шуды Азайту

Деректердегі қажетсіз шуды және қайталанатын ақпаратты (redundancy) азайтады, бұл талдаудың дәлдігін арттырады.



Визуализацияны Жеңілдету

Үлкен өлшемді деректерді 2D немесе 3D кеңістікке түрлендіру арқылы оларды оңай визуализациялауға мүмкіндік береді.

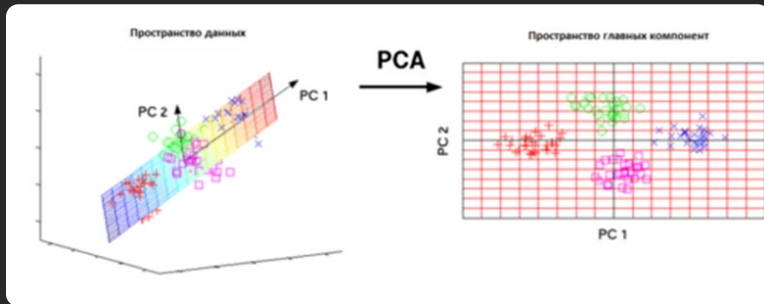


Мультиколлинеарлықты Жою

Айнымалылар арасындағы тәуелділікті азайтып, модельдердің тұрақтылығын арттырады.

РСА кескін өңдеу, бетті тану, қаржылық деректерді талдау және машиналық оқытудың алдындағы алдын ала өңдеу сияқты көптеген салаларда кеңінен қолданылады.

PCA Қадамы: Деректерді Қалыпқа Келтіру



PCA алгоритмінің алғашқы және маңызды қадамы – деректерді стандарттау (қалыпқа келтіру).

- Орташа мән: Әрбір ерекшеліктің (feature) орташа мәнін **0-ге** келтіру.
- Стандартты ауытқу: Әрбір ерекшеліктің стандартты ауытқуын **1-ге** келтіру.

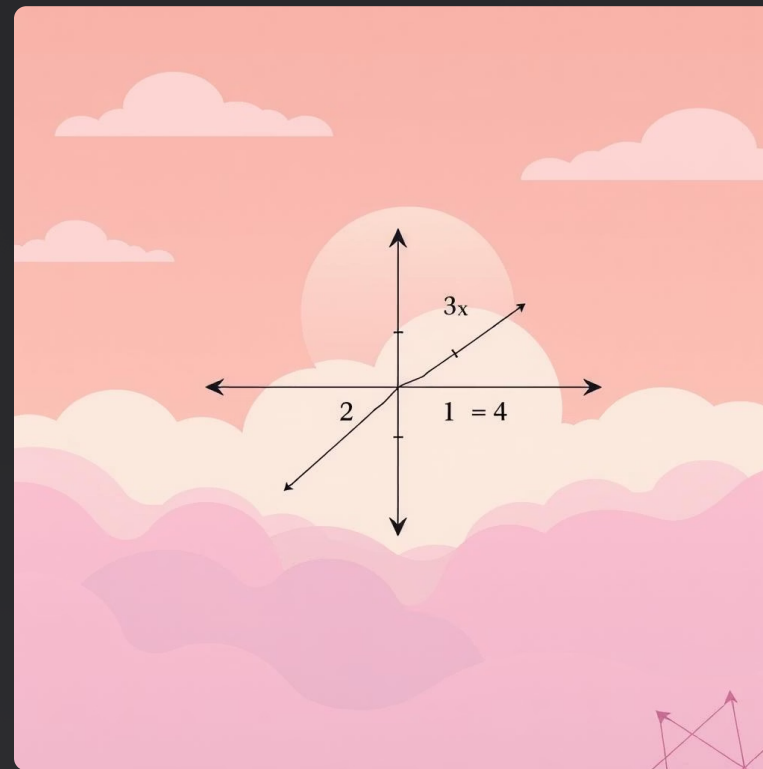
Неліктен қажет? Себебі PCA дисперсияға өте сезімтал. Егер деректер стандартталмаса, дисперсиясы үлкен ерекшеліктер PCA нәтижесіне шамадан тыс әсер етуі мүмкін. Стандарттау барлық белгілердің тең әсер етуін қамтамасыз етеді, бұл әділ және дәл нәтиже алуға көмектеседі.

Дисперсия Матрицасын (Covariance Matrix) Құру

Деректерді стандарттағаннан кейін екінші қадам – деректердің дисперсия матрицасын есептеу.

- **Диагональдағы элементтер:** Әрбір жеке айнымалының дисперсиясын, яғни оның өзгеру дәрежесін көрсетеді.
- **Диагональдан тыс элементтер:** Екі айнымалының бір-бірімен қаншалықты бірге өзгередінін (ковариациясын) көрсетеді. Оң ковариация айнымалылардың бір бағытта, теріс ковариация қарама-қарсы бағытта өзгередінін білдіреді.

Бұл матрица деректердің көп өлшемді кеңістіктегі бағытын және таралуын түсіну үшін өте маңызды, себебі ол айнымалылар арасындағы өзара байланысты сипаттайды.



Меншікті Мәндер (Eigenvalues) және Векторлар (Eigenvectors)

PCA-ның математикалық негізі: дисперсия матрицасының меншікті мәндері (Eigenvalues) мен меншікті векторларын (Eigenvectors) табу.



Меншікті Векторлар

Бұл векторлар жаңа негізгі компоненттердің бағыттарын анықтайды. Олар деректердің ең көп дисперсиясы шоғырланған бағыттарды көрсетеді.

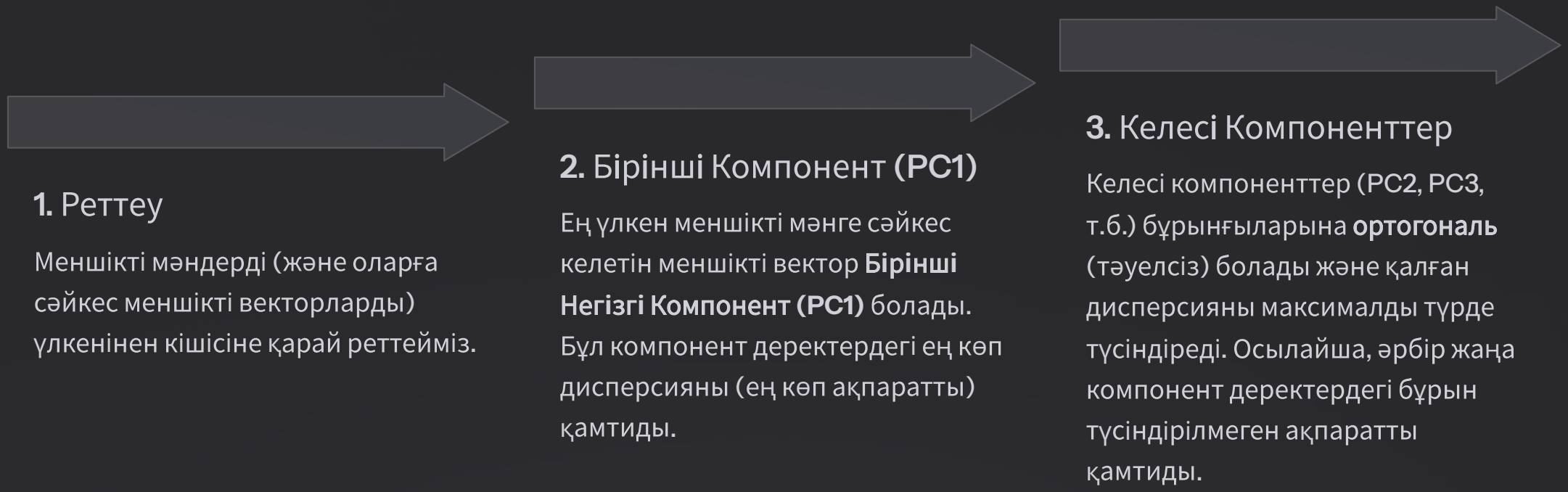


Меншікті Мәндер

Бұл мәндер әрбір меншікті вектор бағытындағы дисперсия мөлшерін, яғни сол бағыттағы ақпараттың қаншалықты көп екенін көрсетеді.

Негізгі Компоненттерді Құру

Меншікті мәндерді үлкенінен кішісіне қарай реттеу арқылы ең маңызды компоненттерді анықтаймыз:



1. Реттеу

Меншікті мәндерді (және оларға сәйкес меншікті векторларды) үлкенінен кішісіне қарай реттейміз.

2. Бірінші Компонент (PC1)

Ең үлкен меншікті мәнге сәйкес келетін меншікті вектор **Бірінші Негізгі Компонент (PC1)** болады. Бұл компонент деректердегі ең көп дисперсияны (ең көп ақпаратты) қамтиды.

3. Келесі Компоненттер

Келесі компоненттер (PC2, PC3, т.б.) бұрынғыларына **ортогональ** (тәуелсіз) болады және қалған дисперсияны максималды түрде түсіндіреді. Осылайша, әрбір жаңа компонент деректердегі бұрын түсіндірілмеген ақпаратты қамтиды.

Оңтайлы Компоненттер Санын Таңдау

Өлшемділікті төмендету үшін қанша негізгі компонентті сақтау керектігін анықтау өте маңызды. Ол үшін бірнеше әдіс бар:

1

1. Дисперсия Пайызы

Ең көп қолданылатын әдіс – **жинақталған дисперсия пайызын** сақтау. Әдетте, **80%-дан 95%-ға** дейінгі дисперсияны сақтайтын компоненттер санын таңдау жеткілікті болып есептеледі. Бұл деректердің көп бөлігін сақтай отырып, шуды азайтуға мүмкіндік береді.

2

2. Тас-құлама Графигі (Scree Plot)

Бұл график әр компоненттің меншікті мәнін көрсетеді. Графиктегі «**шынтақ**» (**elbow**) нүктесі, яғни төмендеу қарқыны күрт бәсеңдейтін нүкте оңтайлы компоненттер саны болып саналады.

3

3. Кайзер Критерийі

Кайзер Критерийі қарапайым ереже ұсынады: тек **меншікті мәні 1-ден үлкен** компоненттерді ғана таңдау. Себебі меншікті мәні 1-ден үлкен компоненттер бастапқы ерекшеліктердің **орташа дисперсиясынан** артық ақпарат береді.

Қорытынды: PCA-ның Деректерді Талдаудағы Рөлі

Негізгі Компоненттер Әдісі (PCA) – күрделі деректерді оңайлатудағы, талдауға және визуализациялауға ыңғайлы етудегі ең қуатты құралдардың бірі. Ол деректерді өңдеудегі "қара жәшік" емес, керісінше, деректердің ішкі құрылымын түсінуге көмектесетін ашық және түсінікті әдіс.

- PCA арқылы өлшемділікті тиімді төмендету машиналық оқыту модельдерінің **жұмыс жылдамдығын** арттырып қана қоймай, олардың **жалпылау қабілетін** (generalization) де жақсартады.
- PCA-ның табысты қолданылуы оның математикалық қадамдарын дұрыс түсінуге және оңтайлы компоненттер санын дұрыс таңдауға байланысты.