

Тақырыбы: Негізгі Компоненттер Әдісі (PCA) және Деректердің Өлшемін Төмендету

Негізгі Компоненттер Әдісі (Principal Component Analysis - PCA) — бұл өлшемділікті төмендету (Dimensionality Reduction) үшін қолданылатын танымал статистикалық процедура. Оның мақсаты — бастапқы деректердің негізгі ақпаратын (дисперсиясын) сақтай отырып, ерекшеліктер (атрибуттар) санын азайту.

PCA деректерді жаңа, ортогоналды (өзара тәуелсіз) айнымалыларға – Негізгі Компоненттерге (Principal Components) түрлендіреді. Бұл жаңа айнымалылар бастапқы ерекшеліктердің сызықтық комбинациясы болып табылады және деректердің ең көп дисперсиясын түсіндіреді.

Негізгі қағидасы:

1. Бірінші Негізгі Компонент (PC1): Деректердің ең үлкен дисперсиясын (таралуын) қамтитын бағыт.
2. Екінші Негізгі Компонент (PC2): PC1-ге перпендикуляр (ортогоналды) және қалған ерекшеліктердің ең үлкен дисперсиясын қамтитын бағыт.
3. Бұл процесс барлық бастапқы өлшемдер санына тең болатын компоненттер санына жеткенше жалғасады.

PCA-ның артықшылықтары

- Өлшемділікті төмендету: Деректерді аз айнымалылармен көрсетуге мүмкіндік береді, бұл есептеу жылдамдығын арттырады.
- Шуды азайту: Деректердің аз дисперсиялы бағыттарын (шуды) жою арқылы модельдің жұмысын жақсартады.
- Визуализация: Жоғары өлшемді (3-тен көп) деректерді бірінші екі немесе үш Негізгі Компонентті пайдаланып, екі немесе үш өлшемді кеңістікте көзбен шолуға (визуализациялауға) мүмкіндік береді.
- Мультиколлинеарлықты жою: Бастапқы деректердегі өзара байланысты айнымалыларды (мультиколлинеарлық) өзара тәуелсіз Негізгі Компоненттерге түрлендіру арқылы шешеді.

Қолдану салалары

Сала	PCA-ның қолданылуы
Машиналық оқыту	Модельді оқыту жылдамдығын арттыру және артық оқытуды (overfitting) азайту.
Кескінді өңдеу	Кескіндерді сығу (қысу) және кескіндердегі шуды жою.

Сала	РСА-ның қолданылуы
Биология/Геномика	Ген экспрессиясының деректерін талдау және пациенттердің үлгілерін топтастыру.
Қаржы	Қор нарығындағы айнымалылардың өзара байланысын төмендету және тәуекелді модельдеу.
Деректерді визуализациялау	Күрделі деректерді 2D немесе 3D графиктерінде қарапайым көрсету.

Деректердің өлшемін төмендету

РСА келесі математикалық қадамдар арқылы жүзеге асады:

1. Деректерді орталықтандыру (Standardization):

РСА-ны қолданар алдында деректерді стандарттау немесе қалыпқа келтіру қажет. Бұл барлық ерекшеліктердің бірдей ауқымға ие болуын қамтамасыз етеді. Әдетте, орташа мәнді алып тастау арқылы орталықтандыру қолданылады:

$$V_{\text{centered}} = V - \mu$$

мұндағы V — бастапқы деректер, μ — ерекшеліктің орташа мәні.

2. Ковариация матрицасын есептеу:

Ковариация матрицасы бастапқы айнымалылар арасындағы корреляцияны өлшейді. Бұл матрица РСА-ның негізі болып табылады, себебі ол деректердегі жалпы дисперсияны және ерекшеліктердің өзара байланысын сипаттайды. $D \times D$ өлшемді матрицаны құрайды, мұнда D — бастапқы ерекшеліктер саны.

$$\text{Cov}(\mathbf{X}) = \frac{1}{n-1} \sum_{i=1}^n (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$$

- $\mathbf{X}$, Деректер Жиынтығы, Бірнеше өлшемдерден (ерекшеліктерден) тұратын бастапқы деректер матрицасы.
- $\sum_{i=1}^n$, Қосынды белгісі, "i=1-ден бастап, n-ге дейінгі барлық бақылаулар (жазбалар) бойынша қосуды білдіреді."
- n, Бақылаулар саны, Деректер жиынтығындағы жалпы жазбалар (қатарлар) саны.
- $n-1$, Нормалау коэффициенті, "n-1 еркіндік дәрежесін білдіреді. Бұл жалпы жиынтықтың (популяцияның) ковариациясы емес, үлгінің (sample) ковариациясын есептеу үшін қолданылады."
- $\mathbf{x}_i$, Жеке бақылау, Деректер жиынтығындағы i-ші бақылаудың (қатардың) векторлық мәні.

- $\bar{x}$, Орташа вектор, Әрбір айнымалының (бағанның) жеке-жеке орташа мәндерінен тұратын вектор.
- $(x_i - \bar{x})$, Ауытқу векторы, i -ші бақылау мен орташа мән арасындағы айырмашылық. Бұл деректердің орталықтан қалай ауытқығанын көрсетеді.
- $(x_i - \bar{x})^T$, Ауытқу векторының транспозициясы, Баған векторын қатар векторына айналдыру.

3. Меншікті мәндерді және Меншікті векторларды есептеу (Eigen Decomposition):

Ковариация матрицасы үшін Меншікті Векторлар (Eigenvectors) және олардың сәйкес Меншікті Мәндері (Eigenvalues) есептеледі.

- Меншікті Векторлар: Жаңа Негізгі Компоненттердің бағыттарын анықтайды. Бұл векторлар PCA-ның жаңа базисі болып табылады.
- Меншікті Мәндер: Әрбір Негізгі Компоненттің қамтитын дисперсия мөлшерін көрсетеді.

4. Компоненттерді таңдау және Түрлендіру:

Меншікті векторлар олардың сәйкес меншікті мәндерінің кему ретімен сұрыпталады. Ең үлкен меншікті мәнге сәйкес келетін вектор (PC1) ең көп дисперсияны қамтиды.

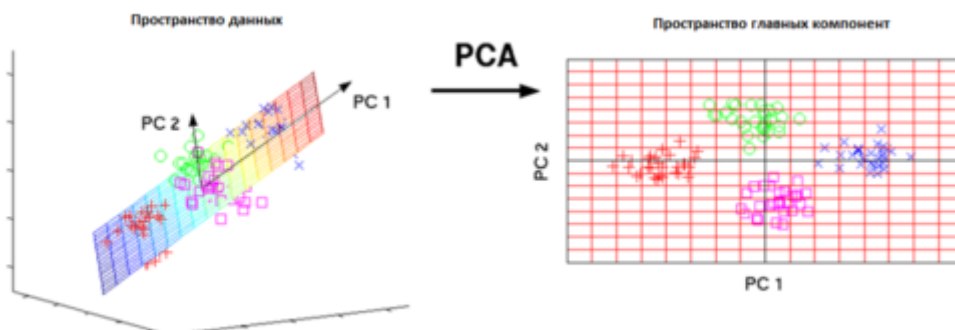
Біз k санын (жаңа, төмендетілген өлшем) таңдаймыз, мұнда $k \leq D$. Тек алғашқы k Меншікті Векторлар (Негізгі Компоненттер) ғана алынады.

Деректерді жаңа k өлшемді кеңістікке түрлендіру:

$$Z = X \cdot W_k$$

мұндағы X — бастапқы деректер (орталықтандырылған), W_k — таңдалған k Меншікті Векторлардан тұратын матрица, Z — түрлендірілген, өлшемі төмендетілген деректер.

Сурет 8.1: PCA-ның геометриялық түсіндірмесі



Оңтайлы өлшемділікті таңдау

Деректердің өлшемін D -дан k -ға дейін төмендеткенде, k мәнін дұрыс таңдау — ең маңызды қадам. Тым аз k ақпаратты жоғалтуға, ал тым көп k өлшемділікті төмендетудің мақсатына жетуге мүмкіндік бермейді.

Оңтайлы k мәнін таңдау әдістері:

1. Жинақталған Дисперсияны Түсіндіру (Cumulative Explained Variance):

Бұл әдіс жалпы дисперсияның белгілі бір пайызын (әдетте 85%-95%) қамтитын компоненттер санын таңдауға негізделген.

$$\text{Жинақталған дисперсия} = \frac{\sum_{i=1}^k \lambda_i}{\sum_{i=1}^D \lambda_i}$$

мұндағы λ_i — i -ші Негізгі Компоненттің меншікті мәні.

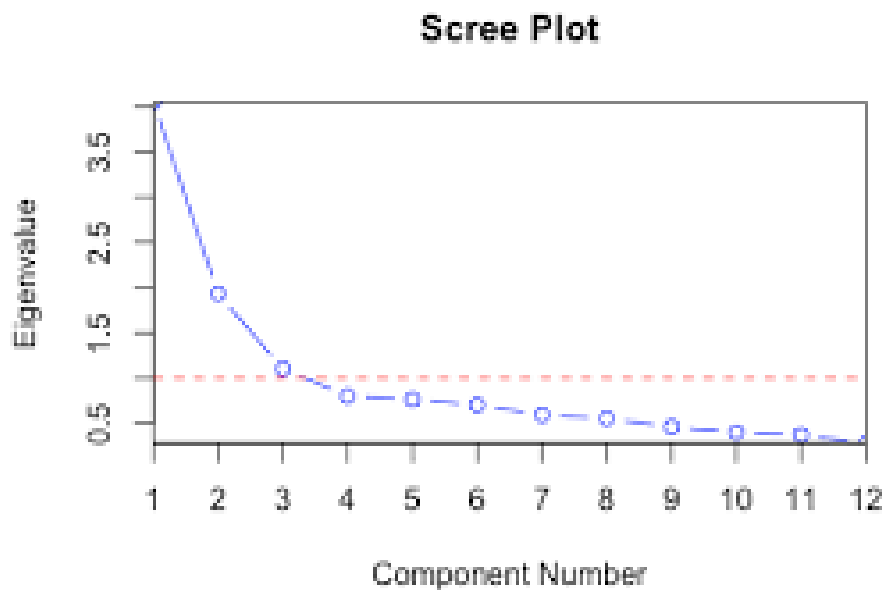
2. Тас-құлама Графигі (Scree Plot):

Бұл график i -ші Негізгі Компоненттің (X осі) түсіндірген дисперсия пайызын (Y осі) көрсетеді. Графиктегі ең үлкен "шынтақ" немесе "құлау" нүктесі оңтайлы k мәні ретінде қарастырылуы мүмкін. Бұл нүктеден кейін компоненттердің қосымша дисперсиясы айтарлықтай төмендейді.

3. Кайзер Критерийі:

Тек меншікті мәні 1-ден үлкен компоненттерді ғана таңдау. Бұл әдіс бастапқы ерекшеліктердің орташа дисперсиясынан (1) артық ақпарат беретін компоненттерді ғана сақтайды.

Сурет 8.2: Тас-құлама Графигі (Scree Plot)



Қорытынды

Негізгі Компоненттер Әдісі (PCA) — деректерді алдын ала өңдеудегі ең қуатты және кеңінен қолданылатын әдістердің бірі. Ол деректердің негізгі ақпаратын сақтай отырып, күрделі деректерді оңайлатады және талдауға ыңғайлы етеді. PCA арқылы өлшемділікті тиімді төмендету машиналық оқыту модельдерінің жұмыс жылдамдығын арттырып қана қоймай, сонымен қатар олардың жалпылау қабілетін (generalization) жақсартады. PCA-ның табысты қолданылуы оның математикалық қадамдарын дұрыс түсінуге және оңтайлы компоненттер санын таңдауға байланысты.

Қолданылған әдебиеттер тізімі

1. Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.)*. Springer. (PCA-ның терең теориялық және практикалық сипаттамасы)
2. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer. (Өлшемділікті төмендету және PCA-ның статистикалық негіздері)
3. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (3rd ed.)*. Elsevier. (Деректерді алдын ала өңдеу және PCA-ның қолданылуы)
4. Shlens, J. (2014). *A Tutorial on Principal Component Analysis*. arXiv preprint arXiv:1404.1100. (PCA-ны қарапайым және қолжетімді түсіндіретін оқу құралы)
5. Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.)*. Springer. (PCA-ның меншікті мәндер мен векторларға негізделген математикалық сипаттамасы)