

Тақырыбы: Кластерлеу әдістері: k -means, DBSCAN және Бағалау Метрикалары

Кластерлеу (Clustering) — бұл бақылаусыз машиналық оқытудың (Unsupervised Machine Learning) негізгі міндеті. Оның мақсаты — деректер жиынтығын ұқсас объектілерден тұратын топтарға (кластерлерге) бөлу.

Негізгі қағидасы:

- Кластер ішіндегі ұқсастық: Бір кластерге жататын деректер элементтері бір-біріне барынша ұқсас болуы керек (жоғары когезия).
- Кластерлер арасындағы айырмашылық: Әр түрлі кластерлерге жататын деректер элементтері бір-бірінен барынша өзгеше болуы керек (төмен байланыс).

Кластерлеу — мәліметтердегі жасырын құрылымды, заңдылықтарды және бұрын белгісіз топтарды (сегменттерді) анықтауға мүмкіндік беретін барлаушы деректерді талдау (Exploratory Data Analysis) құралы.

. Кластерлеуді қолдану салалары

Сала	Қолданылуы	Кластерлеудің міндеті
Маркетинг	Клиенттерді сегменттеу	Мінез-құлқы, сатып алу әдеттері немесе демографиялық сипаттамалары бойынша біртекті тұтынушы топтарын анықтау.
Биология	Геномдарды және ақуыздарды талдау	Ұқсас функциялары бар гендерді немесе ақуыздарды топтастыру.
Қаржы	Алаяқтықты анықтау	Қалыпты транзакциялардан ерекшеленетін әдеттен тыс транзакциялар топтарын (аномалияларды) анықтау.
География	Кескін өңдеу және картография	Спутниктік суреттердегі ұқсас жер беті ерекшеліктерін (мысалы, орман, су, қала) топтастыру.

Кластерлеу әдістерінің түрлері

Кластерлеу алгоритмдері әдетте келесі негізгі түрлерге бөлінеді:

1. Бөлуші әдістер (Partitioning Methods): Деректерді белгілі бір кластер санына (K) бөледі (мысалы, k -means, k -medoids).
2. Тығыздыққа негізделген әдістер (Density-based Methods): Кластерлерді жоғары тығыздық аймақтары ретінде анықтайды және шуды (аномалияларды) кластерлерден бөледі (мысалы, DBSCAN, OPTICS).

3. Иерархиялық әдістер (Hierarchical Methods): Кластерлерді біріктіру (агломеративті) немесе бөлу (дивизивті) арқылы ағаш тәрізді құрылымды құрады.

. Негізгі Кластерлеу Алгоритмдері k-means (K-Орташа) Алгоритмі

k-means — ең қарапайым және жиі қолданылатын бөлуші кластерлеу алгоритмі. Ол деректерді алдын ала белгіленген K кластерге бөлуге бағытталған.

k-means алгоритмінің қадамдары:

1. K мәнін таңдау: Талдаушы деректерді бөлу керек кластерлер санын K анықтайды.
2. Бастапқы центрлерді таңдау: Деректер жиынтығынан кездейсоқ K нүкте таңдалады, олар бастапқы центроидтар (кластер орталықтары) болады.
3. Кластерлерді тағайындау: Деректер жиынтығындағы әрбір нүкте өзіне ең жақын орналасқан центроидқа тағайындалады. Әдетте қашықтық өлшемі ретінде Евклидтік қашықтық қолданылады.

$$d(x_i, c_j) = \sqrt{\sum_{l=1}^m (x_{il} - c_{jl})^2}$$

мұнда x_i — дерек нүктесі, c_j — центроид, m — өлшемдер саны.

4. Центроидтарды жаңарту: Әрбір кластер үшін, кластерге тағайындалған барлық нүктелердің орташа мәні есептеліп, жаңа центроид ретінде орнатылады.
5. Қайталау: 3-ші және 4-ші қадамдар центроидтардың орналасуы айтарлықтай өзгермегенше немесе максималды итерация санына жеткенше қайталанады.

k-means-тің кемшіліктері:

- K мәнін алдын ала анықтауды талап етеді.
- Центроидтарды бастапқы таңдауға сезімтал.
- Тек дөңгелек (сфералық) пішінді кластерлерді таба алады.
- Шуға және аномалияларға өте сезімтал.

DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

DBSCAN — тығыздыққа негізделген алгоритм. Ол пішіні кез келген кластерлерді анықтай алады және шуды (аномалияларды) тиімді түрде өңдейді. DBSCAN центроидтарды пайдаланудың орнына, нүктелердің тығыздығына сүйенеді.

DBSCAN-ның негізгі параметрлері:

1. Epsilon (ϵ немесе Eps): Бір нүктенің көрші аймағын анықтайтын максималды қашықтық радиусы.
2. Minimum Points (MinPts): Кластер құру үшін ϵ радиусындағы көрші аймақта болуы тиіс нүктелердің минималды саны.

Нүктелердің түрлері:

- Ядролық нүкте (Core Point): Оның ϵ радиусы ішінде MinPts -тен көп көршісі бар нүкте.
- Шекаралық нүкте (Border Point): Ядролық нүкте емес, бірақ қандай да бір ядролық нүктенің ϵ радиусы ішінде орналасқан нүкте.
- Шу нүктесі (Noise Point): Ядролық нүкте де, шекаралық нүкте де емес. Бұл нүктелер ешбір кластерге жатпайды (аномалия).

Кластер құру қағидасы:

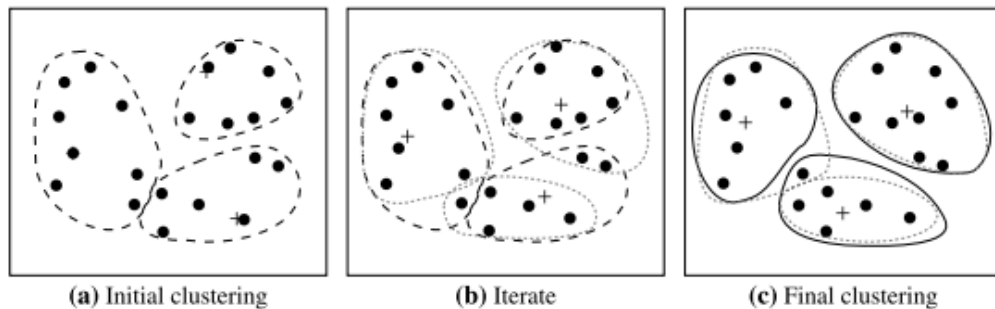
Алгоритм кездейсоқ ядролық нүктеден бастап, оған тікелей қолжетімді барлық нүктелерді және олардың барлық ядролық көршілерін біріктіру арқылы кластерді кеңейтеді.

Артықшылықтары:

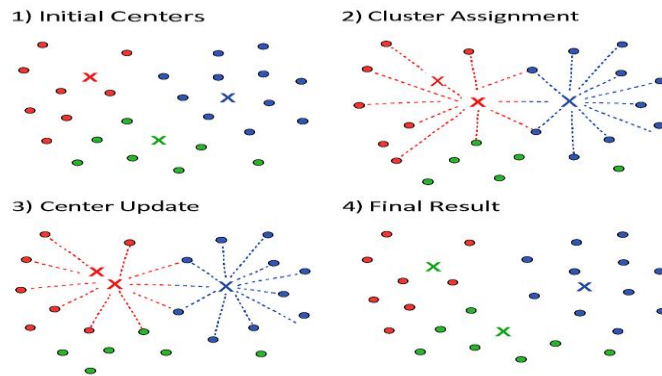
- Пішіні кез келген кластерлерді анықтай алады.
- Шуды (аномалияларды) тиімді түрде өңдейді.
- Алдын ала кластер санын (K) анықтауды қажет етпейді.

Сурет 6.1: k -means алгоритмінің итеративті процесі

The k -means partitioning algorithm.



k -means-тің үш немесе төрт кезеңін көрсететін схема : 1) Бастапқы центрлер; 2) Кластерге тағайындау; 3) Центрді жаңарту; 4) Соңғы нәтиже.)



Кластерлеудің сапасын бағалау

Кластерлеу — бақылаусыз оқыту болғандықтан, оның сапасын бағалау классификацияға қарағанда күрделірек. Бағалау әдістері екіге бөлінеді:

Ішкі бағалау метрикалары

Ішкі метрикалар оқыту деректерінің ішкі құрылымына сүйенеді және кластерлердің біртектілігін (compactness) және бөлінуін (separation) өлшейді.

1. Силуэт Коэффициенті (Silhouette Coefficient): Бұл метрика кластер ішіндегі ұқсастық пен кластерлер арасындағы айырмашылықты біріктіреді. Мән $[-1, 1]$ аралығында болады.
 - 1 -ге жақын мән — Кластерлер тығыз және жақсы бөлінген.
 - 0 -ге жақын мән — Кластерлердің арасындағы шекара анық емес.
 - -1 -ге жақын мән — Деректер дұрыс емес кластерге тағайындалған.
2. Инерция (Inertia - k-means үшін): Бұл метрика кластердегі нүктелердің өз центроидына дейінгі қашықтықтарының квадраттарының қосындысы. k-means алгоритмі осы мәнді минималдандыруға тырысады.

$$\text{Inertia} = \sum_{j=1}^K \sum_{x \in C_j} d(x, c_j)^2$$

Инерция K кластер саны артқан сайын әрдайым төмендейді. Сондықтан, K мәнін таңдау үшін Шынтақ Әдісі (Elbow Method) қолданылады. Бұл әдіс инерция графигіндегі ең үлкен бұрылыс нүктесін (шынтақты) таңдайды.

Сыртқы бағалау метрикалары

Егер деректерде "шындық" сыныптық жапсырмалар (Ground Truth) болса, бұл метрикалар кластерлеу нәтижесін осы шын жапсырмалармен салыстырады.

1. Анықталған реттелген индекс (Adjusted Rand Index - ARI): Кластерлердің бөлінуі мен шын сыныптардың бөлінуінің қаншалықты ұқсас екенін өлшейді. Мән $[-1, 1]$ аралығында, мұнда 1 — толық сәйкестік.
2. Бір-біріне ұқсас балл (Homogeneity, Completeness): Кластердің бір ғана сыныптың мүшелерінен тұру дәрежесін (Homogeneity) және бір сыныптың барлық мүшелерінің бір кластерге жату дәрежесін (Completeness) өлшейді.

Тапсырмаларда кластерлеу

Кластерлеуді шешім қабылдаудың бастапқы кезеңі ретінде қолдануға болады.

К мәнін оңтайландыру (Elbow Method)

k -means үшін К мәнін дұрыс таңдау өте маңызды. Elbow Method:

1. К мәнін 1-ден бастап белгілі бір максималды мәнге дейін арттырып, әрбір К үшін инерция есептеледі.
2. Инерция мәнінің К санына тәуелділік графигі сызылады.
3. Графикте инерцияның төмендеу қарқыны күрт бәсеңдейтін нүкте (шынтақ) оңтайлы К мәні ретінде таңдалады.

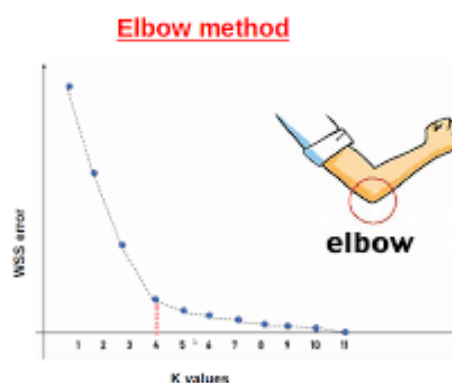
Алдын ала өңдеудің әсері

k -means сияқты қашықтыққа негізделген алгоритмдер деректердің ауқымына өте сезімтал. Сондықтан, кластерлеу алдында деректерді стандарттау (Z -score Standardization) немесе қалыпқа келтіру (Min-Max Normalization) міндетті болып табылады.

DBSCAN-ның икемділігі

DBSCAN алгоритмі әсіресе географиялық деректерде, мысалы, тұрғындардың тығыз орналасуын немесе жол қозғалысының тығыз аймақтарын (бір-бірімен тікелей байланысқан кластерлер) анықтау үшін тиімді, себебі ол кластерлердің кез келген пішінін таба алады.

Сурет 6.2: К мәнін таңдау үшін Elbow Method



Инерция (Y осі) мен K саны (X осі) арасындағы байланысты көрсететін график қойылады, онда инерцияның төмендеу қарқыны өзгертін "шынтақ" нүктесі (оңтайлы K) көрсетіледі.)

Қорытынды

Кластерлеу әдістері, атап айтқанда k-means және DBSCAN, деректерді талдаушылар үшін деректердегі жасырын құрылымды ашуға мүмкіндік беретін күшті құралдар болып табылады.

k-means — есептеуі жылдам, іске асыруы оңай, бірақ K саны мен бастапқы центроидтарға сезімтал.

DBSCAN — пішіні кез келген кластерлерді таба алады және шуға төзімді, бірақ оның параметрлерін (ϵ , MinPts) таңдау біліктілікті қажет етеді.

Кластерлеудің сапасын бағалау (Силуэт, Шынтақ әдісі) модельдің объективті түрде тиімді жұмыс істеуін қамтамасыз етеді, бұл оны іскери және ғылыми тапсырмаларда тиімді қолдануға жол ашады.

Қолданылған әдебиеттер тізімі

1. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques (3rd ed.)*. Elsevier. (k-means және кластерлеудің негізгі түрлері)
2. Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*. In KDD. (DBSCAN алгоритмін ұсынған бастапқы мақала)
3. Rousseeuw, P. J. (1987). *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*. Journal of Computational and Applied Mathematics, 20, 53–65. (Силуэт коэффициентінің теориялық негізі)
4. Tan, P. N., Steinbach, M., & Kumar, V. (2023). *Introduction to Data Mining*. Pearson. (Кластерлеуді бағалау және K мәнін таңдау әдістері)
5. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer. (Өртүрлі кластерлеу әдістерін салыстыру)