

t-SNE және Деректерді Визуализациялау: Жоғары Өлшемді Кеңістіктегі Жасырын Құрылымды Ашу

Жоғары өлшемді деректерді талдау және түсіндіру бүгінгі ғылым мен технологияның өзекті мәселесі болып табылады. t-SNE (t-distributed Stochastic Neighbor Embedding) әдісі осы күрделі деректердегі жасырын құрылымдарды төмен өлшемді, көбіне екі өлшемді (2D) кеңістікке түрлендіру арқылы көрнекі етіп көрсетеді.

Оқытушы: Сарсембаева Т.С.

t-SNE Алгоритмінің Жалпы Ұғымы

t-SNE — жоғары өлшемді деректерді **сызықтық емес** түрлендіру арқылы көрнекі бейнелеуге арналған қуатты әдіс. Оның негізгі идеясы өте қарапайым әрі тиімді: жоғары өлшемді кеңістікте **ұқсас (жақын)** орналасқан нүктелерді төмен өлшемді кеңістікте де **жақын** орналастыру. Бұл деректердің ішкі кластерлік құрылымын айқын көрсетуге мүмкіндік береді.

Жоғары Өлшемді Деректер

Күрделі деректер жиыны



Сызықтық емес түрлендіру

Ұқсастықтарды сақтау

Төмен Өлшемді Визуализация

Көрнекі кластерлер

Жоғары Өлшемді Кеңістіктегі Ұқсастықты Бағалау

t-SNE алгоритмінің алғашқы қадамы – жоғары өлшемді кеңістіктегі екі нүкте арасындағы **ұқсастықты** немесе **ықтималдықты** дәл есептеу. Бұл үшін **Гаусс үлестірімі (Gaussian distribution)** қолданылады. Әрбір дерек нүктесі өзінің көршілеріне қатысты ықтималдық таралымын қалыптастырады. Бұл ықтималдық нүктелердің бір-біріне қаншалықты көрші болу мүмкіндігін көрсетеді. Жақын нүктелердің ықтималдығы жоғары болады.

Гаусс Үлестірімі

Гаусс үлестірімі (қоңырау тәрізді қисық) нүктелердің арақашықтығын ықтималдыққа айналдырады. Егер екі нүкте жақын орналасса, олардың бір-біріне тиесілі болу ықтималдығы жоғары болады.

$$P_{j|i} = \frac{\exp(-\|x_i - x_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq i} \exp(-\|x_i - x_k\|^2 / 2\sigma_i^2)}$$

Мұндағы x_i мен x_j – жоғары өлшемді нүктелер, ал σ_i – Гаусс үлестірімінің стандартты ауытқуы.

Төмен Өлшемді Кеңістіктегі Ұқсастықты Бағалау

Жоғары өлшемді кеңістіктегі ұқсастықтарды есептегеннен кейін, t-SNE енді төмен өлшемді кеңістіктегі нүктелер арасындағы ұқсастықты есептейді. Бұл жерде **Студенттің t-үлестірімі (Student's t-distribution)** қолданылады. Бұл үлестірім Гаусс үлестіріміне қарағанда **ауыр құйрықты (heavy-tailed)** болып келеді, яғни алыс орналасқан нүктелерге аз ықтималдық бере отырып, оларды төмен өлшемді кеңістікте **бөліп ұстауға** көмектеседі. Бұл кластерлердің анық көрінуін қамтамасыз етеді.

Студенттің t-Үлестірімі

Студенттің t-үлестірімі Гаусс үлестіріміне қарағанда қашықтықты әлсіз жазалайды, бұл алыс нүктелердің төмен өлшемді кеңістікте бір-біріне өте жақын келмеуін қамтамасыз етеді. Бұл кластерлер арасындағы бос орынды көбейтіп, оларды айқынырақ етеді.

$$Q_{ij} = \frac{(1 + \|y_i - y_j\|^2)^{-1}}{\sum_{k \neq l} (1 + \|y_k - y_l\|^2)^{-1}}$$

Мұндағы y_i мен y_j – төмен өлшемді нүктелер.

Перплекситілік (Perplexity) Ұғымы

Перплекситілік – t-SNE алгоритмінің ең маңызды параметрлерінің бірі. Ол әр нүктенің көршілерінің **тиімді санын** анықтайды. Бұл параметр деректердің **жергілікті (local)** және **жаһандық (global)** құрылымы арасындағы тепе-теңдікті реттейді.

1

Жергілікті Құрылым

Көрші нүктелердің қалай топтасатынын көрсетеді. Төмен перплекситілік жергілікті ұқсастықтарға басымдық береді.

2

Жаһандық Құрылым

Деректер жиынының жалпы пішіні мен үлкен кластерлер арасындағы қатынасты көрсетеді. Жоғары перплекситілік жаһандық құрылымды сақтауға ұмтылады.

3

Дұрыс Таңдау

Перплекситіліктің дұрыс мәнін таңдау нәтиженің сапасы үшін өте маңызды. Әдетте 5-тен 50-ге дейінгі мәндер қолданылады.

Оптимизация: KL-Дивергенция

t-SNE алгоритмінің негізгі мақсаты – жоғары өлшемді кеңістіктегі деректер нүктелерінің арасындағы ықтималдық таралымы (P) мен төмен өлшемді кеңістіктегі нүктелердің арасындағы ықтималдық таралымы (Q) арасындағы айырмашылықты барынша азайту. Бұл айырмашылықты бағалау үшін **Куллбек-Лейблер Дивергенциясы (KL-Divergence)** атты метрика қолданылады.

KL-Дивергенция дегеніміз не?

KL-дивергенция екі ықтималдық таралымының қаншалықты әртүрлі екенін өлшейді. t-SNE контекстінде ол жоғары өлшемді және төмен өлшемді кеңістіктердегі көршілік ұқсастықтарының қаншалықты сәйкес келетінін көрсетеді.

$$C = \sum_i \sum_j P_{ij} \log \frac{P_{ij}}{Q_{ij}}$$

Мақсат – осы C мәнін минималдандыру, яғни екі ұқсастық матрицасының бір-біріне мүмкіндігінше жақын болуын қамтамасыз ету.

t-SNE Жұмыс Қағидасы: Градиенттік Түсу

t-SNE осы KL-дивергенцияны **градиенттік түсу (Gradient Descent)** әдісі арқылы итеративті түрде азайтады. Бұл процесте төмен өлшемді кеңістіктегі деректер нүктелерінің орналасуы әрбір қадамда реттеліп отырады.



Бастапқы Орналасу

Нүктелер төмен өлшемді кеңістікте кездейсоқ орналасады.



KL-Дивергенция Есептеу

Ағымдағы орналасуға негізделген қате есептеледі.



Градиенттік Түсу Қадамы

Нүктелер қатені азайту үшін жаңа позицияларға жылжиды.



Итерациялар

Процесс конвергенцияға жеткенше немесе максималды итерация санына жеткенше қайталанады.

PCA-мен Салыстыру: Сызықтық және Сызықтық Емес

Деректерді өлшемділігін төмендетудің екі танымал әдісі – PCA және t-SNE – әртүрлі қағидаларға негізделген және әртүрлі мақсаттарға қызмет етеді.

PCA (Негізгі Компоненттер Әдісі)

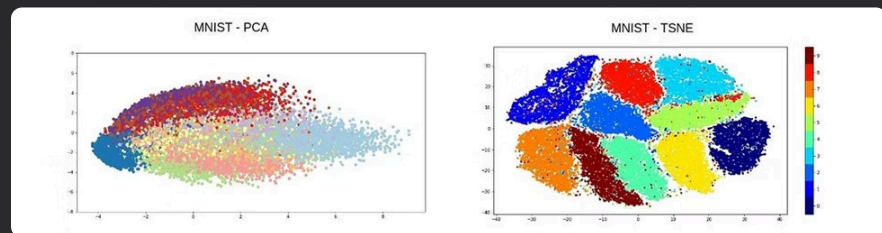
Сызықтық трансформация. Ол деректердің ең үлкен **дисперсиясын** сақтауға бағытталған. PCA үлкен масштабтағы деректер құрылымын, мысалы, деректердің жалпы бағыттарын анықтауда тиімді.

- Мақсаты: Дисперсияны максималдандыру
- Түрі: Сызықтық
- Нәтиже: Глобалды құрылым

t-SNE (t-distributed Stochastic Neighbor Embedding)

Сызықтық емес трансформация. Ол негізінен **жергілікті ұқсастықты** сақтауға бағытталған. t-SNE кластерлердің ішкі мәнін, яғни топтардың ішіндегі және арасындағы қатынастарды ашуда ерекшеленеді.

- Мақсаты: Жергілікті ұқсастықты сақтау
- Түрі: Сызықтық емес
- Нәтиже: Жергілікті кластерлік құрылым



Визуализация Нәтижелерін Интерпретациялау Ережелері

t-SNE карталарын дұрыс түсіну үшін кейбір маңызды ережелерді есте сақтау қажет. Әдістің табиғатынан туындайтын қате түсініктерден аулақ болу маңызды.



Арақашықтық Маңызды

Тек екі нүктенің төмен өлшемді кеңістікте жақын болуы олардың жоғары өлшемді кеңістікте ұқсас екенін білдіреді. Алыс орналасқан кластерлер әртүрлі болып саналады.



Өлшем Маңызды Емес

Кластерлердің төмен өлшемді кеңістіктегі **өлшеміне** және **пішініне** сенуге болмайды. Олар жоғары өлшемді кеңістіктегі шындықты дәл көрсетпеуі мүмкін.



Сақтық Таныту

t-SNE нәтижелерін басқа әдістермен (мысалы, PCA) бірге қолданып, жан-жақты талдау жасау ұсынылады. Бұл қате интерпретацияны болдырмауға көмектеседі.

Қорытынды: Екі Әдісті Біріктіріп Қолдану

t-SNE — жоғары өлшемді деректердегі күрделі құрылымдарды және кластерлерді ашуға мүмкіндік беретін таптырмас құрал. PCA жалпы дисперсияны сақтауға бағытталған сызықтық әдіс болса, t-SNE сызықтық емес трансформация арқылы деректердің ішкі мәнін ашады. Деректерді толық түсіну үшін аналитиктер осы екі әдістің артықшылықтарын біліп, оларды біріктіріп қолдануы керек.

"Деректерді тек көру жеткіліксіз, оны түсіну қажет."