

Тақырыбы: t-SNE және Деректердің Визуализациясы: PCA-мен Салыстыру

t-SNE (t-distributed Stochastic Neighbor Embedding) — бұл жоғары өлшемді деректерді төмен өлшемді кеңістікте (әдетте 2D немесе 3D) тиімді визуализациялауға арналған күрделі, сызықтық емес (non-linear) өлшемділікті төмендету алгоритмі.

t-SNE-нің негізгі мақсаты — бастапқы жоғары өлшемді деректер жиынтығындағы ұқсастықты (similarity) төмен өлшемді картада барынша сақтау. Ол әсіресе кластерлерді (топтарды) көрнекі түрде анықтау үшін өте тиімді.

t-SNE-нің жұмыс принципі

t-SNE екі негізгі қадамға негізделген:

1. Ықтималдықтарды есептеу (Жоғары өлшемді кеңістік):

Алдымен, алгоритм жоғары өлшемді кеңістіктегі әрбір нүкте жұбы арасындағы ұқсастықты ықтималдық ретінде өлшейді.

- Жақын көршілер: Жақын орналасқан нүктелердің ұқсастық ықтималдығы жоғары болады.
- Алыс көршілер: Алыс орналасқан нүктелердің ұқсастық ықтималдығы төмен болады.

Бұл үшін Гаусс таралуы (Gaussian distribution) қолданылады.

2. Төмен өлшемді картаны құру (Сызықтық емес трансформация):

Келесі кезекте, алгоритм төмен өлшемді (2D/3D) картадағы нүктелерді орналастырады, сонда бұл картадағы нүктелер арасындағы ұқсастық ықтималдықтары (t-таралуы арқылы есептелген) жоғары өлшемді кеңістіктегі ықтималдықтарға мүмкіндігінше ұқсас болады.

- t-таралу (Student's t-distribution): Төмен өлшемді кеңістіктегі ықтималдықтарды есептеу үшін қолданылады. t-таралуы алыс нүктелерді бір-бірінен "тарқатуға" көмектеседі, бұл кластерлер арасындағы қашықтықты айқынырақ көрсетеді.
- Kullback-Leibler (KL) Дивергенциясы: Екі ықтималдық таралуы арасындағы айырмашылықты (қателікті) өлшеу үшін қолданылады. t-SNE осы KL Дивергенциясын минималдандыру арқылы жұмыс істейді.

t-SNE-нің негізгі параметрі:

- Перплекситілік (Perplexity): t-SNE алгоритмінің әрбір нүкте үшін қанша көршіні ескеретінін (тұрғын саны) анықтайды. Әдетте 5-тен 50-ге дейінгі мәндер қолданылады. Бұл параметр нәтижелік картаның пішініне қатты әсер етеді.

t-SNE-нің нәтижесін интерпретациялау

t-SNE нәтижесіндегі графиктегі негізгі назар:

- Кластерлердің болуы: Бір-біріне жақын орналасқан нүктелер топтары (кластерлер) бастапқы деректердегі ұқсас деректерді білдіреді.

- Кластерлер арасындағы қашықтық: Төмен өлшемді картадағы кластерлер арасындағы қашықтық, әдетте, жоғары өлшемді кеңістіктегі нақты қашықтықты дәл білдірмейді. t-SNE тек жергілікті құрылымды (ішкі топтарды) көрсетуге бағытталған.

PCA мен t-SNE арасындағы айырмашылықтар

PCA және t-SNE екеуі де өлшемділікті төмендететін әдістер болғанымен, олардың жұмыс принциптері, мақсаттары және нәтижелерінің интерпретациясы әртүрлі.

Ерекшелік	PCA (Негізгі Компоненттер Әдісі)	t-SNE (t-distributed Stochastic Neighbor Embedding)
Сипаты	Сызықтық (Linear)	Сызықтық емес (Non-linear)
Мақсаты	Глобалды құрылымды (жалпы дисперсияны) сақтау. Деректерді барынша сығу.	Жергілікті құрылымды (жақын көршілердің ұқсастығын) сақтау. Визуализациялау.
Қашықтық	Графиктегі қашықтық дисперсияны (нақты қашықтықты) білдіреді.	Графиктегі қашықтық ұқсастықты білдіреді. Кластерлер арасындағы қашықтық интерпретацияланбайды.
Есептеу жылдамдығы	Өте жылдам, үлкен деректермен жұмыс істей алады.	Баяу, үлкен деректерде есептеу шығыны жоғары.
Қолдануы	Алдын ала өңдеу, Модельді оқыту жылдамдығын арттыру, Шуды жою.	Визуализация, Кластерлеу нәтижелерін көрсету, Деректердегі жасырын топтарды ашу.
Тұрақтылық	Тұрақты, бастапқы деректердің кішкене өзгерістеріне сезімтал емес.	Кездейсоқ бастауға және перплекситілікке сезімтал, нәтиже тұрақсыз болуы мүмкін.

Деректерді визуализациялау және түсіндіру

Визуализацияның маңыздылығы

Жоғары өлшемді деректерді 2D-ға проекциялау арқылы аналитиктер:

1. Деректердің таралуын көру: Деректердің қалай шоғырланғанын немесе тарқағанын бақылау.
2. Кластерлерді растау: Кластерлеу алгоритмдерінің нәтижесін (мысалы, k -means) көзбен шолып тексеру.
3. Аномалияларды (Outliers) табу: Басқа топтардан алыс орналасқан жеке нүктелерді анықтау.
4. Ерекшеліктерді таңдауға көмектесу: Кейбір ерекшеліктердің кластерлерді бөлуде қаншалықты маңызды екенін түсіну.

t-SNE нәтижелерін түсіндірудегі шектеулер

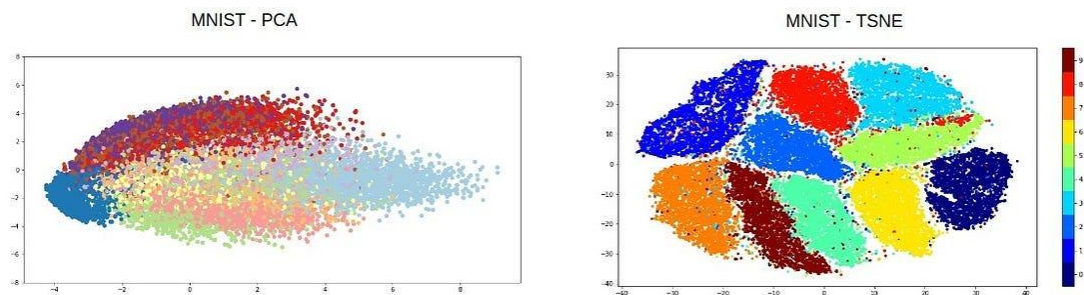
t-SNE қуатты құрал болғанымен, оның нәтижелерін түсіндіргенде сақ болу қажет:

1. Қашықтықтың мәні: t-SNE нәтижесіндегі кластерлер арасындағы қашықтық нақты жоғары өлшемді қашықтықты білдірмейді. Яғни, екі кластер бір-бірінен алыс болса, олардың ұқсастығы төмен деген сөз, бірақ олардың нақты қашықтығын салыстыруға болмайды.
2. Кластер мөлшері: Графиктегі кластердің өлшемі (тығыздығы) бастапқы деректердегі кластердің нақты мөлшерін білдірмейді.
3. Кездейсоқтық: t-SNE кездейсоқ бастауға сезімтал болғандықтан, әр орындалған сайын кластерлердің орналасуы (ротациясы, аударылуы) өзгеруі мүмкін. Кластерлердің орналасуын емес, тек олардың ішкі тұтастығын (қандай нүктелер бір топқа кіретінін) интерпретациялау керек.

PCA-ны t-SNE алдында қолдану

Үлкен деректер жиындарымен жұмыс істегенде, есептеу уақытын қысқарту және шуды азайту үшін PCA-ны t-SNE-ден бұрын алдын ала өңдеу қадамы ретінде қолданған жөн. Әдетте, PCA арқылы өлшемділікті 50-ге дейін төмендетіп, содан кейін t-SNE-ні қолдану ұсынылады.

Сурет 9.1: PCA және t-SNE нәтижелерінің салыстырмасы



PCA Нәтижесі (Глобалды Құрылым)

- Графиктің сипаты: Нүктелер әдетте кеңістікке біркелкі таралып, бір үлкен бұлтқа ұқсайды.
- Кластерлер: Кластерлер, егер олар өте айқын бөлінсе ғана, көрінуі мүмкін. Көбінесе, кластерлер бір-бірімен араласып жатады.
- Түсіндіру: Егер екі нүкте PCA графигінде алыс орналасса, бұл олардың жоғары өлшемді кеңістікте де нақты физикалық тұрғыдан (дисперсия бойынша) алыс екенін білдіреді. PCA жалпы таралған қалыпты (trend) және деректердің үлкен айырмашылықтарын көрсетеді.

t-SNE Нәтижесі (Жергілікті Құрылым)

- Графиктің сипаты: Нүктелер тығыз, айқын, әдемі "бұлттарға" немесе "шеңберлерге" бірігеді. Бұл визуалды түрде әлдеқайда тартымды және "кластерлі" болып көрінеді.
- Кластерлер: t-SNE-нің негізгі күші — бұрын белгісіз болған кішкентай, күрделі топтарды (кластерлерді) айқын бөліп көрсетуі.
- Түсіндіру:
 - Бірге топтасу: Графикте бірге топтасқан нүктелер жоғары өлшемді кеңістікте ұқсас екенін білдіреді.
 - Арақашықтыққа мән бермеу: Егер екі кластер бір-бірінен алыс орналасса, бұл олардың әртүрлі екенін көрсетеді, бірақ олардың арасындағы нақты қашықтық (мысалы, 5 бірлік немесе 10 бірлік) жоғары өлшемді кеңістіктегі нақты қашықтыққа сәйкес келмеуі мүмкін. Кластерлер арасындағы аумақтық кеңістікке (boiling/bubble size) мәлімдеме жасамау маңызды.

Қорытынды

t-SNE — жоғары өлшемді деректердегі күрделі жергілікті құрылымдарды және кластерлерді ашуға мүмкіндік беретін таптырмас визуализация құралы. PCA сызықтық және жылдам әдіс ретінде жалпы дисперсияны сақтауға бағытталған болса, t-SNE сызықтық емес трансформация арқылы ұқсас нүктелерді бірге топтастырып, деректердің ішкі мәнін ашады. Аналитиктер екі әдістің артықшылықтарын біліп, оларды деректерді толық түсіну үшін біріктіріп қолдануы керек.

Қолданылған әдебиеттер тізімі

1. van der Maaten, L. J. P., & Hinton, G. (2008). *Visualizing data using t-SNE*. Journal of Machine Learning Research, 9(86), 2579–2605. (t-SNE алгоритмін ұсынған бастапқы мақала)
2. Jolliffe, I. T. (2002). *Principal Component Analysis (2nd ed.)*. Springer. (PCA-ның негізгі қағидалары)
3. Wattenberg, M., Viégas, F., & Johnson, I. (2016). *How to Use t-SNE Effectively*. Distill. (t-SNE-ні тиімді қолдануға арналған практикалық нұсқаулық және интерпретациялау ережелері)
4. Dougherty, G. (2020). *Principal Component Analysis (PCA) and t-SNE in data dimensionality reduction*. In Data Mining. Springer.
5. Géron, A. (2022). *Hands-On Machine Learning with Scikit-Learn, Keras & TensorFlow (3rd ed.)*. O'Reilly Media. (PCA және t-SNE-нің Python-дағы практикалық мысалдары)