

№13 Дәріс. Кластерлеу әдістері (K-means)

Дәрістің мақсаты:

Студенттерге кластерлеудің мәнін түсіндіру, K-means алгоритмінің жұмыс принципін, артықшылықтары мен шектеулерін және оны нақты деректерге қолдану тәсілдерін үйрету.

Кілттік сөздер:

Кластерлеу, K-means, қадағаланбайтын оқыту, центроид, Евклид қашықтығы, итерация, кластер орталығы, сегментация, машиналық оқыту.

Дәріс жоспары:

1. Кіріспе
2. Кластерлеу түсінігі
3. K-means алгоритмі
 - Негізгі идеясы
 - Алгоритм қадамдары
4. Қашықтық өлшемдері
5. K мәнін таңдау (Elbow method)
6. K-means алгоритмінің артықшылықтары және кемшіліктері
7. Қолдану салалары
8. Қорытынды
9. Бақылау сұрақтары
10. Пайдаланылған әдебиеттер

1. Кіріспе

Қазіргі таңда деректердің қарқынды өсуі және олардың күрделенуі жасанды интеллект пен машиналық оқытудың маңызын арттыра түсті. Бұл технологиялар тек болжау мен классификациялау үшін ғана емес, сондай-ақ үлкен көлемдегі деректерді құрылымдау, түсіну және басқару үшін де қолданылады. Осындай маңызды бағыттардың бірі — **кластерлеу**.

Кластерлеу — бұл бақылаусыз оқыту әдістерінің бірі, мұнда модель алдын ала белгіленген категорияларсыз деректер арасынан табиғи топтарды анықтайды. Басқаша айтқанда, кластерлеу — мәліметтер жиынын ішкі ұқсастықтарына қарай бірнеше топқа бөлу тәсілі.

Кластерлеу көптеген қолданбалы салаларда кеңінен қолданылады, мысалы:

- маркетингте тұтынушыларды сегменттеу
- медицинада науқастарды диагностикалық топтарға бөлу
- қаржы саласында транзакцияларды талдап, аномалияларды анықтау
- биологияда ген экспрессияларын топтау
- ақпараттарды автоматты түрде ұйымдастыру және іздеу жүйелерінде

K-means — кластерлеудің ең кең таралған және тиімді әдістерінің бірі. Оның қарапайымдылығы, есептеу тиімділігі және жоғары түсініктілігі оны көптеген практикалық міндеттерде қолдануға мүмкіндік береді. Бұл алгоритм

деректерді k топқа бөледі және әрбір кластердің орталығын (центроид) есептейді.

Осы дәрісте біз K-means алгоритмінің жұмыс принципі, оның артықшылықтары, кемшіліктері және практикалық қолдану мысалдарымен танысамыз.

2. Кластерлеу түсінігі

Кластерлеу — деректерді олардың өзара ұқсастығы бойынша бірнеше топқа (кластерлерге) бөлу процесі.

Мақсаты: бір кластер ішіндегі объектілердің ұқсастығын максималды ету және әртүрлі кластерлер арасындағы айырмашылықты арттыру.

Мысалдар:

- Клиенттерді сатып алу әдетіне қарай сегментациялау;
- ДНҚ деректерін топтау;
- Құжаттарды мазмұнына қарай топқа бөлу;
- Аномалияны анықтау.

3. K-means алгоритмі

Негізгі идеясы:

Алдын ала берілген K санына сәйкес деректерді топтарға бөлу және әр топтың ортасын (centroid) анықтау.

Алгоритм қадамдары

1. K кластер санын таңдау
2. Бастапқы центр нүктелерін кездейсоқ таңдау
3. Әр объектіні ең жақын центрге қосу
4. Әр кластер үшін жаңа центрді есептеу
5. Центрлер өзгермейінше 3–4 қадамдарды қайталау

Екі нүкте арасындағы қашықтық (Евклид):

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

4. Қашықтық өлшемдері

- Евклид қашықтығы (ең жиі)
- Манхэттен қашықтығы
- Косинустық қашықтық

5. K мәнін таңдау (Elbow Method)

K мәні дұрыс таңдалмаса:

- Топтар тым көп $\rightarrow$ модель күрделенеді
- Топтар тым аз $\rightarrow$ ақпарат жоғалады

Elbow (шынтақ) әдісі:

Әртүрлі K мәндер үшін шығын функциясын есептеу және графиктегі иілу нүктесін таңдау.

6. Артықшылықтары мен кемшіліктері

Артықшылықтары:

- Қарапайым және түсінікті
- Үлкен деректерде тез жұмыс істейді

- Практикада кең қолданылатын әдіс
- Кемшіліктері:*
- К мәнін алдын ала беру керек
 - Шу (noise) мен аномалияларға сезімтал
 - Тек дөңгелек кластерлер үшін тиімді

7. Қолдану салалары

Сала	Қолданылуы
Маркетинг	Клиент сегментациясы
Медицина	Пациент топтарын анықтау
Экология	Климаты ұқсас аймақтарды бөлу
Cybersecurity	Аномалияларды табу
NLP	Тақырыптық модельдер

8. Қорытынды

Кластерлеу — деректер арасындағы табиғи құрылымды табуға мүмкіндік беретін қуатты құрал. Оның ішінде **К-means алгоритмі** — қарапайымдылығы мен жылдамдығы арқасында ең кең таралған кластерлеу әдісі.

К-means көптеген қолданбалы міндеттерде өзін тиімді көрсетеді, мысалы клиенттерді сегменттеу, мінез-құлық модельдерін талдау, аномалияларды анықтау, медициналық диагноздарды топтау және басқа да көптеген салаларда.

Сонымен қатар, К-means-тің шектеулері де бар: ол кластер санын алдын ала таңдауды талап етеді және деректер таралуының сфералық болуын қалайды. Сондықтан зерттеуші әрқашан дұрыс параметр таңдау, нормализациялау және нәтижелерді талдау қажет.

Жалпы алғанда, К-means — машиналық оқытудың негізгі құралдарының бірі және оны меңгеру аналитикалық ойлау мен деректерді тиімді талдауға үлкен мүмкіндік береді.

9. Бақылау сұрақтары

1. Кластерлеу дегеніміз не?
2. К-means алгоритмінің негізгі идеясы қандай?
3. Алгоритмнің негізгі қадамдарын атаңыз.
4. К мәні қалай таңдалады?
5. К-means артықшылықтары мен кемшіліктері қандай?
6. Евклид қашықтығы қандай формуламен есептеледі?
7. Қандай салаларда қолданылады?

10. Пайдаланылған әдебиеттер

1. James G., Witten D., Hastie T., Tibshirani R. *An Introduction to Statistical Learning*.
2. Bishop C. *Pattern Recognition and Machine Learning*.

3. Han J., Kamber M. *Data Mining: Concepts and Techniques*.
4. Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*.
5. Andrew Ng — Machine Learning Course Notes (Stanford).
6. Python Scikit-Learn Documentation.