Пәні: Деректерді интеллектуалды талдау және визуализациялау

K-орташа алгоритмі (k-means). Визуализациялау тапсырмалары

Дәріс №10

Шаяхметова А.С., PhD, қауымдастырылған профессор
Жасанды интеллект және Big Data кафедрасы
2026-2027 оқу жылы

Мақсаты

k-орташа (k-means) кластерлеу алгоритмінің мәнін түсіндіру. Алгоритмнің жұмыс қадамдарын және аяқталу шартын көрсету. Алгоритмді екі өлшемді мысалда визуализациялау. Мәселелері мен кеңейтімдерін, қолдану аясын қарастыру.

Жоспары

1. k-орташа алгоритмі: мәні және тарихы
Кластерлеудің ең танымал әдісі, мақсат функциясы

2. Алгоритмнің жұмыс принципі
Негізгі идея, қадамдар, аяқталу шарты, параметрлер

3. Алгоритмді визуализациялау
Екі өлшемді жағдай, қадамдар тізбегі

4. Мәселелері, кеңейтімдері және қолданылуы
Жергілікті минимум, k-means++, терең оқыту

1. k-орташа алгоритмі: мәні және тарихы

Кластерлеудің ең танымал әдісі

Мақсат функциясы: квадраттық ауытқу

Тарихы: Штайнгаус, Ллойд, Маккуин

k-орташа әдісі (k-means)

Мәні
Кластерлеудің ең танымал әдісі

Бөлу
Көптеген векторларды k кластердің белгілі санына бөледі

Негізі
ЕМ алгоритмінің (гаусс қоспасын бөлу) нұсқасы

Орталықтар
Кластер орталықтары негізгі нүктелер деп те аталады

Әдістің өзі негізгі нүктелер әдісі деп те аталады

Алгоритмнің тарихы: схема

1950 жылдар

Гуго Штайнгаус, Стюарт Ллойд

→

Маккуиннің жұмысы

Әдіс ерекше танымал болды

→

Қазіргі уақыт

Кластерлеудің ең танымал әдісі

Әдісті 1950 жылдары математиктер ойлап тапқан, Маккуиннің жұмысынан кейін ерекше танымал болды

Мақсат функциясы

$$V = \Sigma \Sigma (x - \mu i)^2, \quad i = 1 \ldots k, \quad x \in Si$$

Белгі	Мағынасы
V	Квадраттық ауытқу, азайтылатын шама
Si	і-ші кластер
μi	Кластер орталығы (масса центрі)

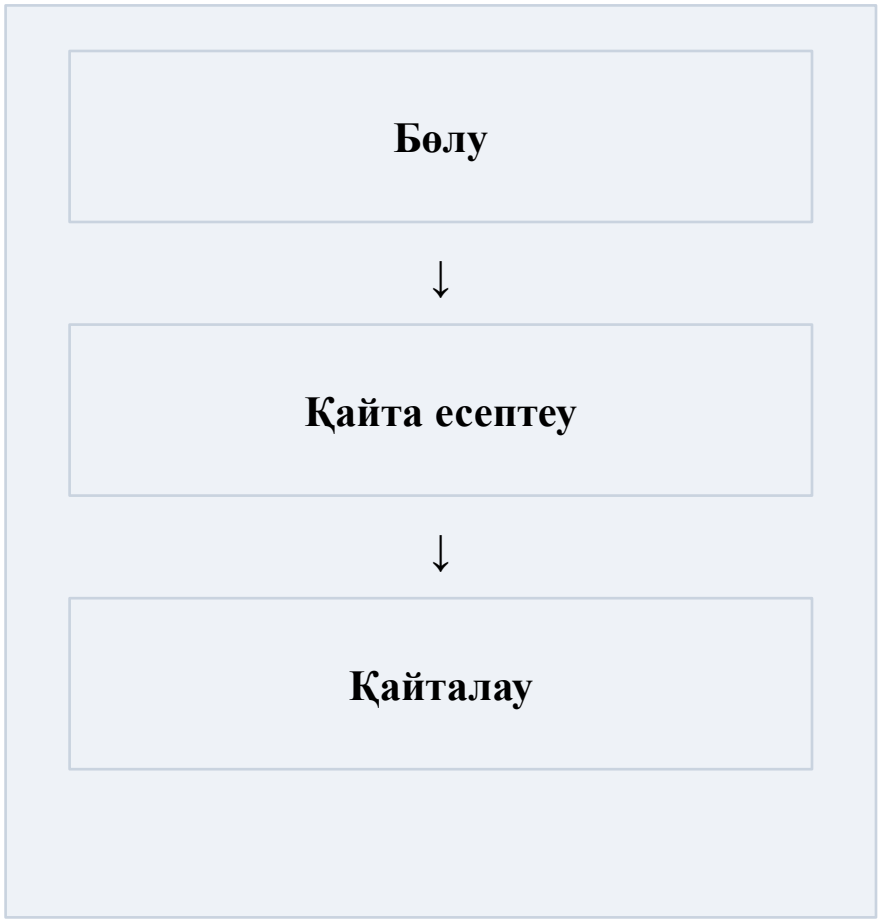
Алгоритм кластер нүктелерінің орталықтан квадраттық ауытқуын азайтады

2. Алгоритмнің жұмыс принципі

Негізгі идея: орталықты қайта есептеу

Қадамдар мен бөлу

Аяқталу шарты



Негізгі идея: схема

Бастапқы орталықтар

k орталық таңдалады

→

Нүктелерді жақын орталыққа жатқызу

таңдалған метрика бойынша

→

Орталықты қайта есептеу

масса центрі есептеледі

→

Қайталау

бөліну тұрақтанғанша

Әр итерацияда масса центрі қайта есептеліп, векторлар таңдалған метрика бойынша жақын орталыққа бөлінеді

Алгоритм қадамдары

1. k бастапқы орталықты таңдау → **2. Нүктелерді жақын орталыққа бөлу** → **3. Орталықтарды қайта есептеу**

↓

Бөліну өзгерді ме?

Жоқ → Аяқталу

← Иә болса - 2-қадамға қайту

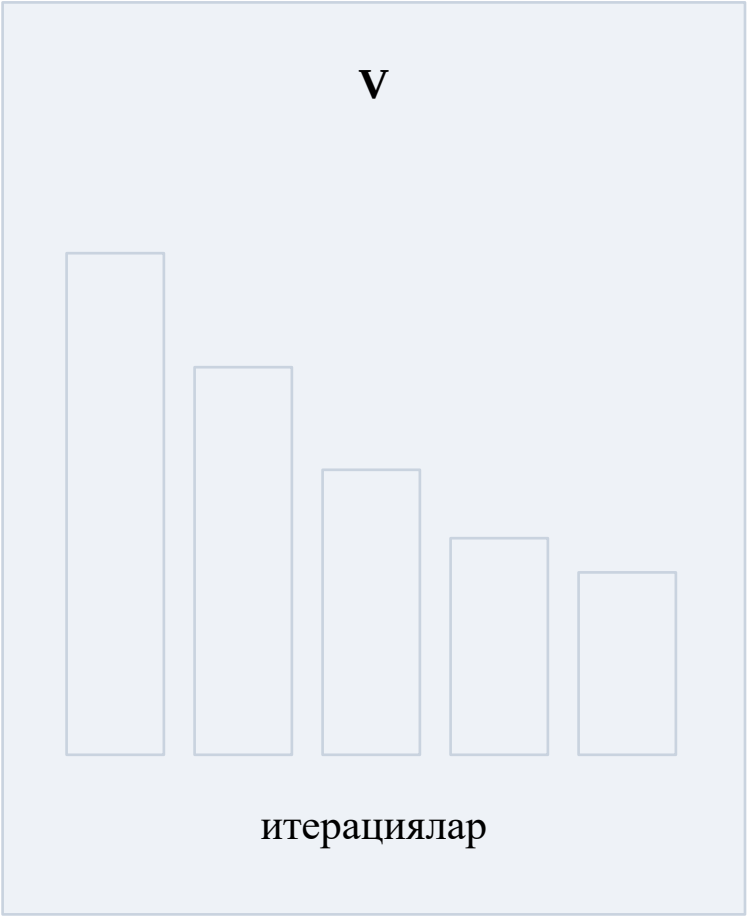
Аяқталу шарты

Алгоритм ішкі қашықтық өзгермегенде аяқталады

Әр қадамда V квадраттық ауытқу азаяды

Бөліністер саны ақырлы, сондықтан цикл мүмкін емес

Жинақталу уақыты жиынтық түріне байланысты



Алгоритм параметрлері

Параметр	Мағынасы
k	Кластерлер саны, алдын-ала беріледі
Метрика	Нүктелер арасындағы қашықтық өлшемі
Бастапқы орталықтар	Кездейсоқ таңдалады
V	Квадраттық ауытқу, азайтылатын шама
Векторлар таңдалған метрика бойынша жақын орталыққа бөлінеді	

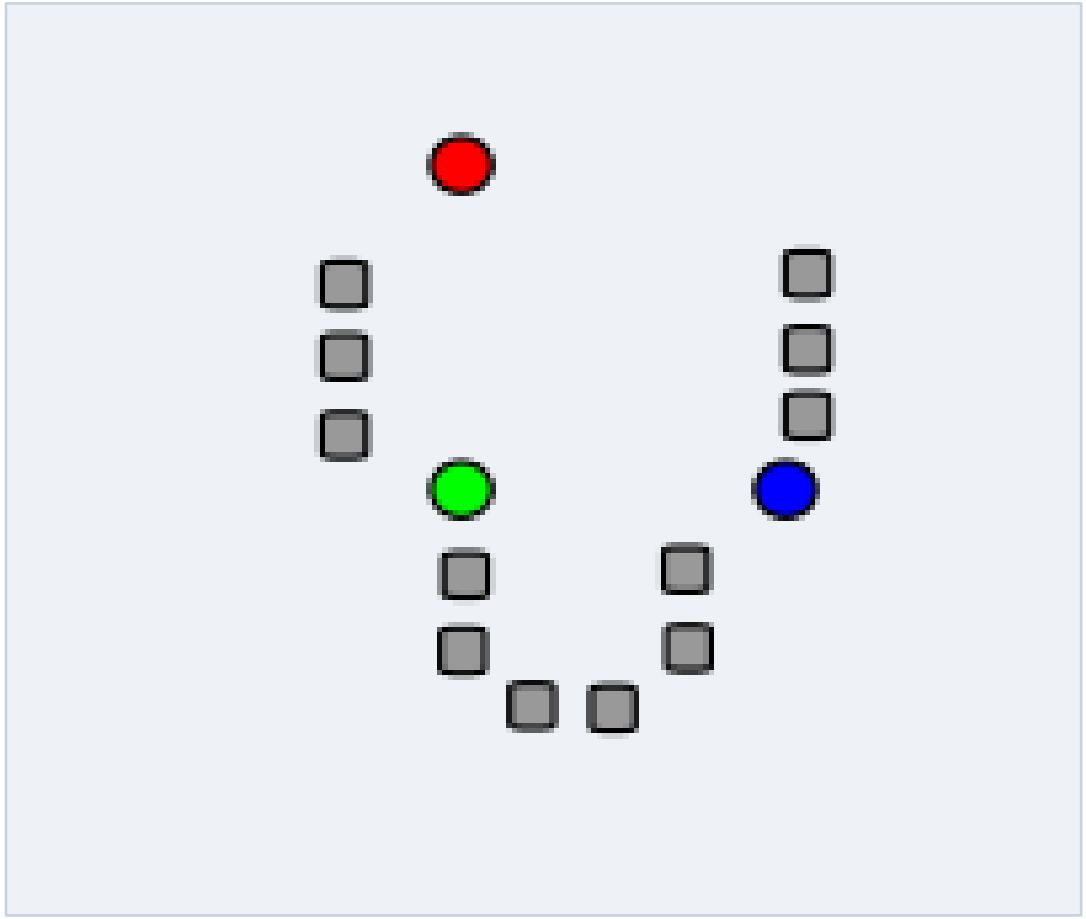
3. Алгоритмді визуализациялау

Екі өлшемді жағдай

Бастапқы нүктелер кездейсоқ таңдалады

Қадамдар тізбегі

1-қадам: бастапқы нүктелер

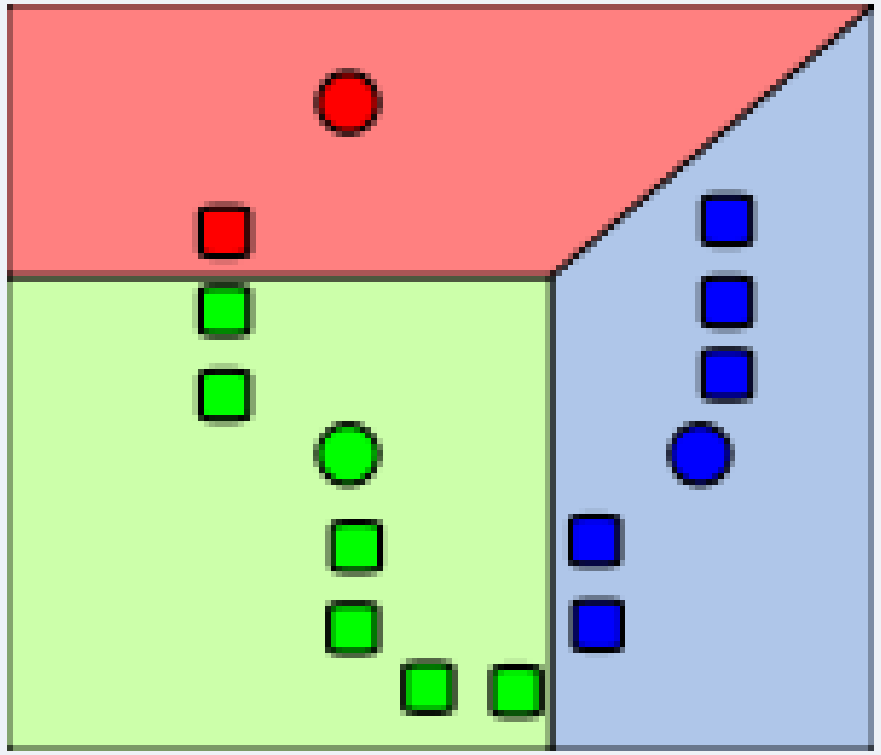


Екі өлшемді жағдай

Бастапқы орталықтар кездейсоқ таңдалады

Түрлі-түсті нүктелер - орталықтар

2-қадам: нүктелерді бөлу

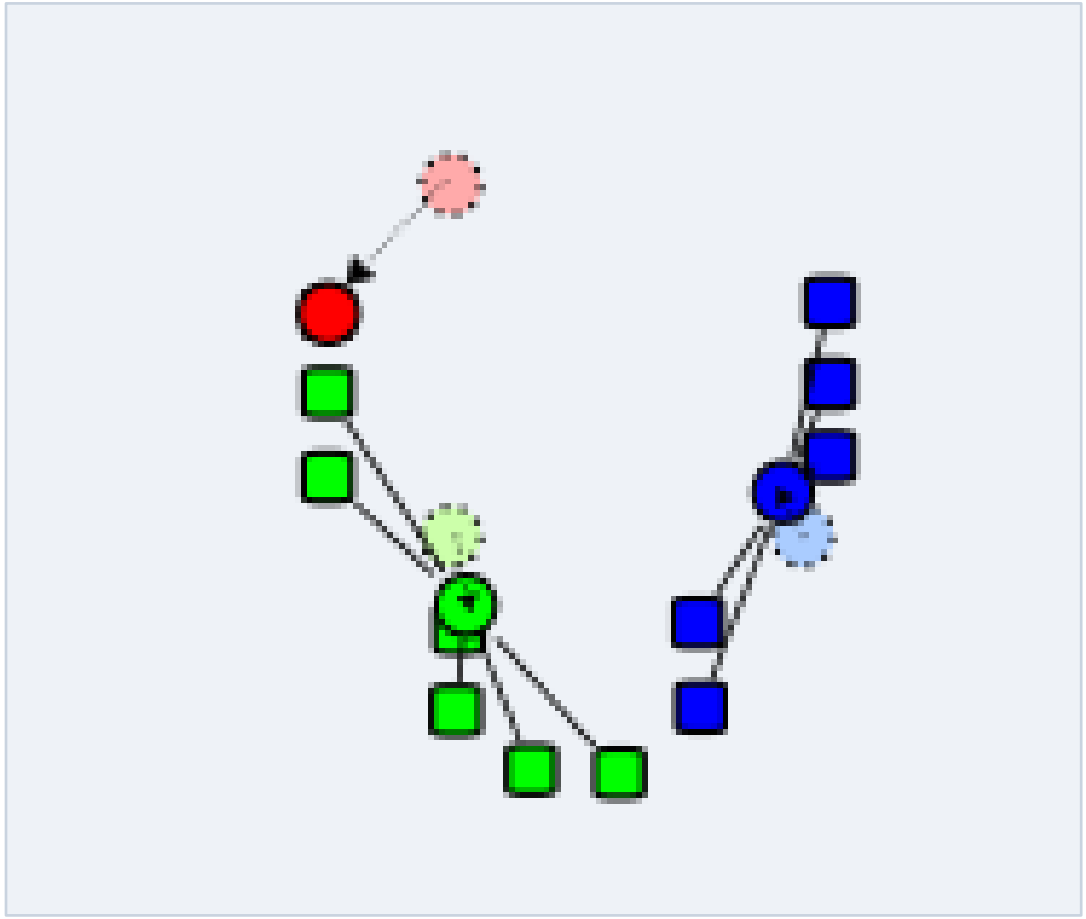


Нүктелер жақын орталыққа жатқызылады

Жазықтық бастапқы орталықтарға қатысты бөлінеді

Бөліну таңдалған метрика бойынша жүреді

3-қадам: жаңа орталықтар

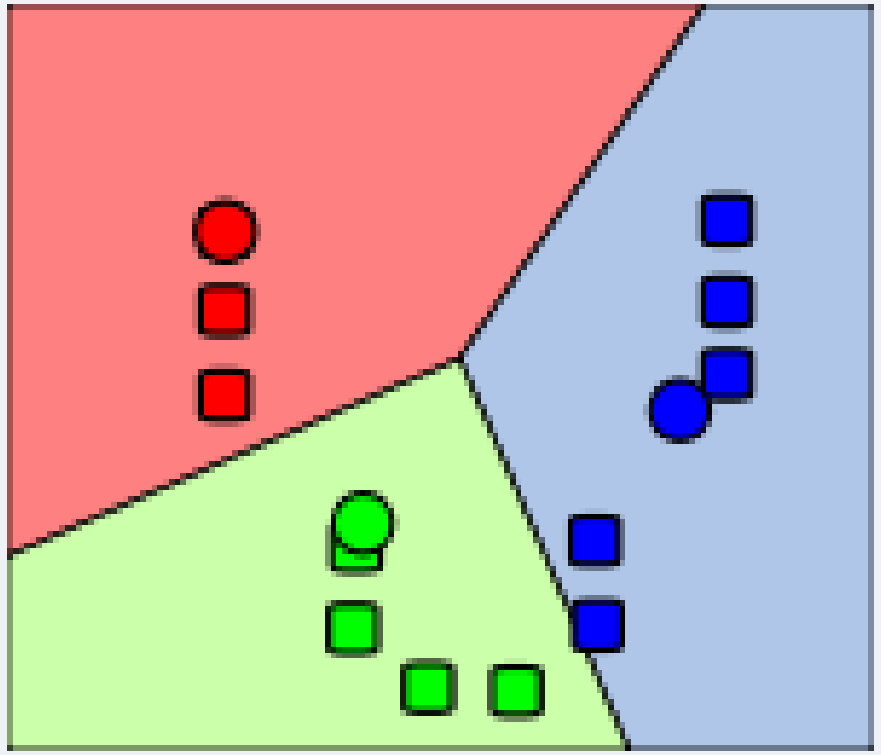


Жаңа кластер орталықтары есептеледі

Масса центрі ізделеді

Орталықтар жаңа орынға жылжиды

4-қадам: қайталау

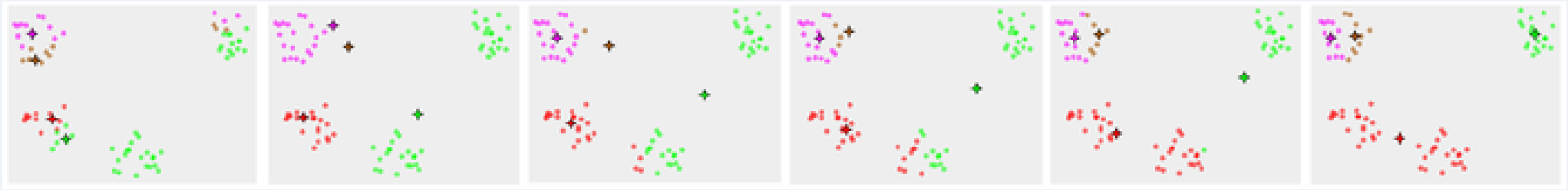


Қадамдар қайталанады

Алгоритм жинақталғанша жалғасады

Бірінші қадамнан басқалары қайталанады

Қадамдар тізбегі



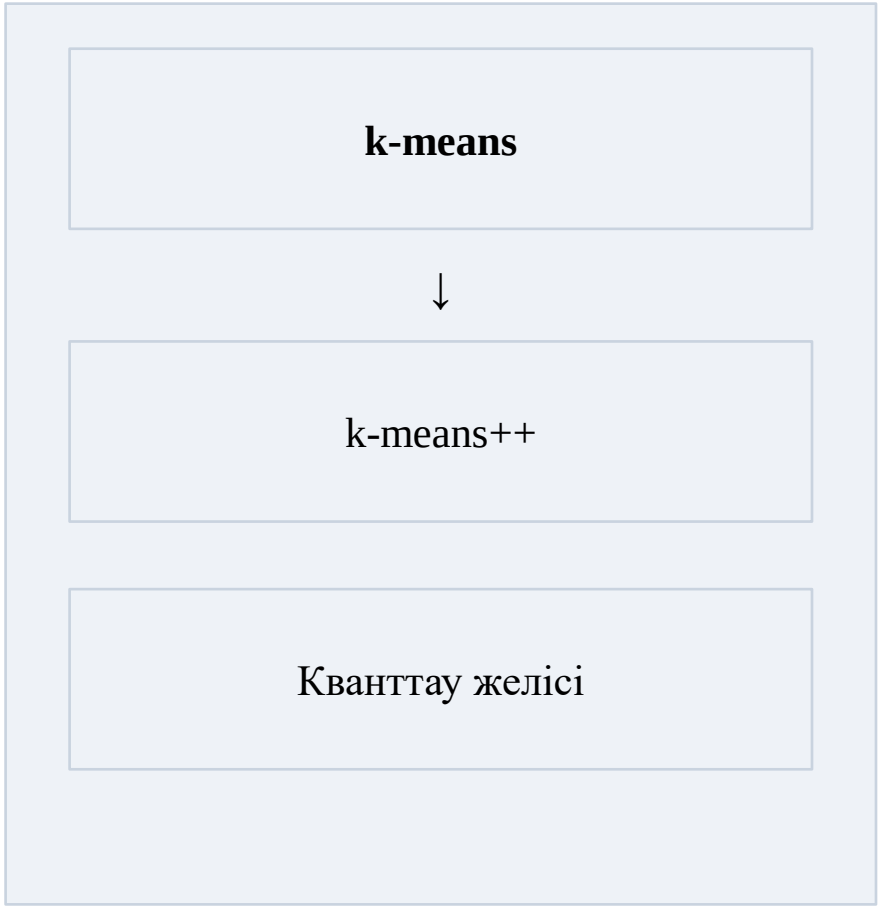
Қадам	Не болады
1	Бастапқы орталықтар кездейсоқ таңдалады
2	Нүктелер жақын орталыққа жатқызылады
3	Масса центрі бойынша жаңа орталықтар есептеледі
4	Әр итерацияда орталықтар жылжиды, соңғы суретте бөліну тұрақтанады

4. Мәселелері, кеңейтімдері және қолданылуы

Жергілікті минимум

k-means++ кеңейтімі

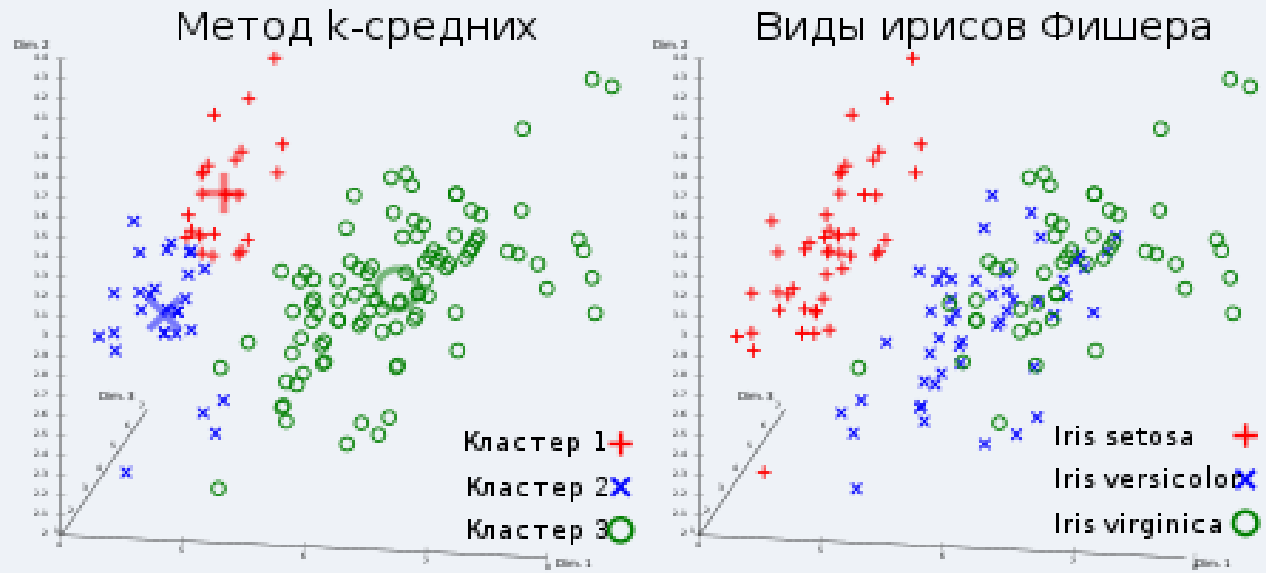
Терең оқытуда қолдану



k-means мәселелері

Мәселе	Сипаттамасы
Жергілікті минимум	Жаһандық минимумға кепілдік жоқ
Бастапқы орталықтар	Нәтиже таңдауға байланысты, оңтайлы таңдау белгісіз
Кластер саны	k алдын-ала білінуі керек
Жергілікті минимумда нәтиже деректердің айқын кластерлік құрылымына қайшы келуі мүмкін	

Фишер иристері: кластерлеу нәтижесі



Фишер иристері үшін k-орташа нәтижесі

Үлкен маркерлер – кластер орталықтары

Нәтиже иристердің нақты түрлерімен салыстырылады

Кеңейту және вариация: схема

k-means

↓ ↓

k-means++: бастапқы орталықтарды оңтайлы таңдау

Векторлық кванттау желісі (Кохонен желісінің нұсқасы)

k-means++ кластер орталықтарының бастапқы мәндерін оңтайлы таңдауға бағытталған

Терең оқытуда қолдану: схема

Суреттің 16×16 бөліктері → k=256 кластерге бөлу → Центроидтер

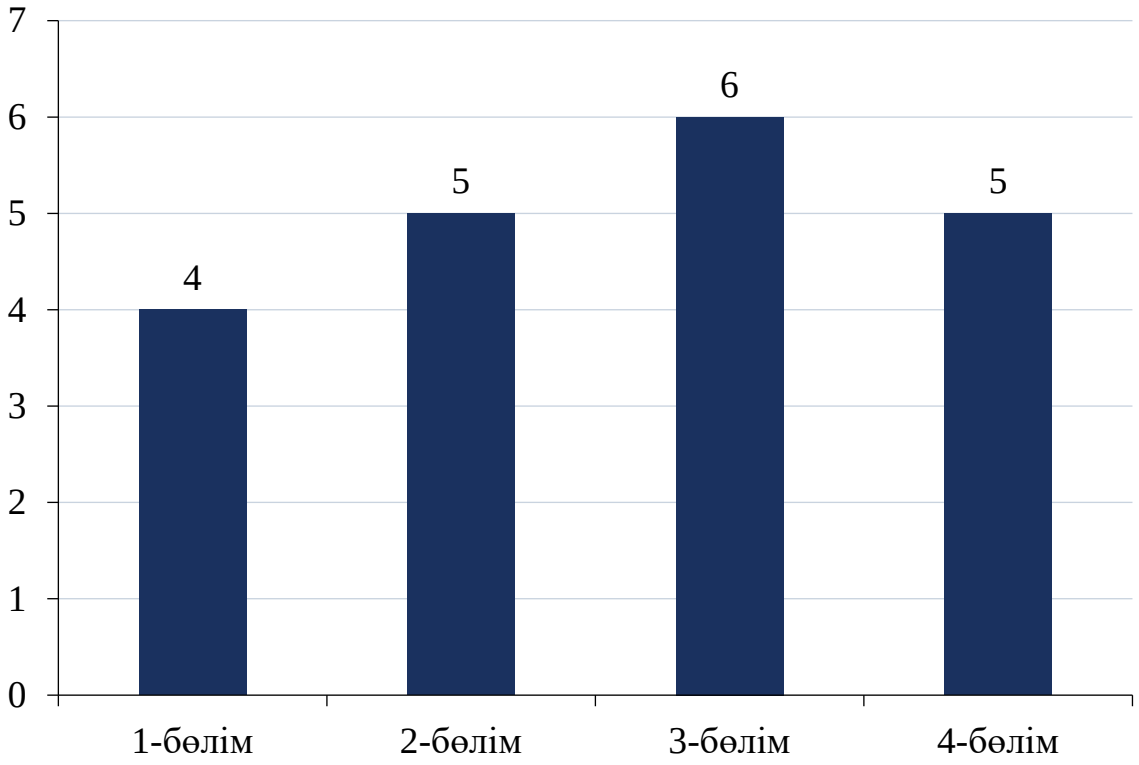
↓

Сүзгілер (жинақтау ядролары) ← Конволюциялық нейрондық желі

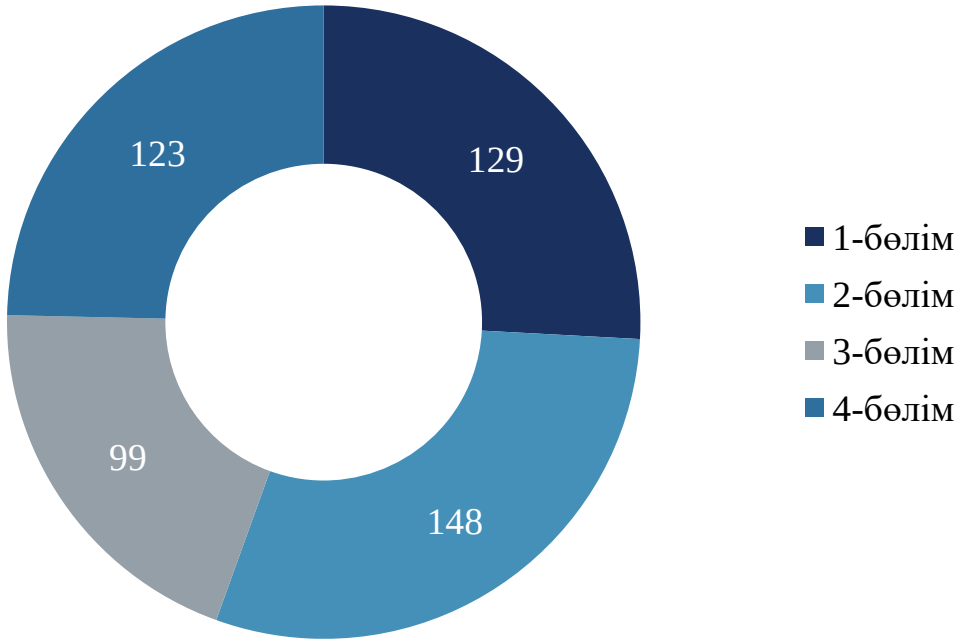
Оқыту k-орта әдісімен мұғалімсіз жүзеге асады

Диаграммалар

Бөлімдер бойынша слайд саны



Бөлімдер бойынша мәтін көлемі (сөз саны)



Қорытынды

1. k-means - нүктелерді k кластерге бөлетін ең танымал әдіс

2. Орталықтарды қайта есептеу мен қайта бөлу бөліну тұрақтанғанша қайталанады

3. Әдіс бастапқы орталықтарға және k таңдауына тәуелді, жергілікті минимумға тоқтауы мүмкін

4. k-means++ және терең оқытудағы қолданыс әдістің мүмкіндігін кеңейтеді

K-орташа алгоритмі (k-means). Визуализациялау тапсырмалары

Дәріс №10