



KDD (Knowledge Discovery in Data) кезеңдері

Оқытушы: Сарсембаева Т.С.

Дәрістің мақсаты:

Студенттерге деректерден білімді ашу (KDD) процесінің кезеңдері мен олардың өзара байланысын түсіндіру; деректерді алдын ала өңдеу, үлгілеу және нәтижелерді бағалау тәсілдерінің мәнін ашу; талдаудың тиімділігін арттырудағы әр кезеңнің рөлін көрсету.

Негізгі сұрақтар:

1. KDD процесінің мәні және оның мақсаттары.
2. KDD кезеңдерінің жалпы құрылымы.
3. Деректерді алдын ала өңдеу: тазалау, біріктіру және трансформациялау.
4. Үлгілеу (Data Mining): классификация, регрессия, кластерлеу, ассоциация ережелері.
5. Бағалау және интерпретациялау кезеңдері.
6. KDD процесінің тиімділігін қамтамасыз ететін факторлар.

Қысқаша мазмұн (тезистер):

KDD (Knowledge Discovery in Data) – деректерден білім алу процесі, ол бірнеше өзара байланысты кезеңдерден тұрады: Деректерді таңдау (Selection) – талдауға қажетті деректерді анықтау және жинау. Деректерді алдын ала өңдеу (Pre-processing) – деректерді тазалау, толықтыру және шуларды жою. Деректерді трансформациялау (Transformation) – деректерді талдауға қолайлы форматқа келтіру (нормализация, агрегация, дискретизация). Деректерді өндіру (Data Mining/Modeling) – үлгілер мен заңдылықтарды табу (классификация, регрессия, кластерлеу, ассоциациялық талдау). Бағалау және интерпретациялау (Evaluation and Interpretation) – алынған нәтижелердің дұрыстығын бағалау және оларды білім ретінде ұсыну.

Деректерді алдын ала өңдеу кезеңі KDD процесінің ең жауапты бөлігі, себебі нәтижелердің сапасы көбіне осы кезеңге байланысты (80%-ға дейін әсер етеді).

Data Mining әдістері:

- Классификация — объектілерді белгілі топтарға жатқызу.
- Регрессия — үздіксіз шамаларды болжау.
- Кластерлеу — ұқсас элементтерді топтау.
- Ассоциация ережелері — айнымалылар арасындағы тәуелділіктерді анықтау.

Бағалау кезеңінде дәлдік (Accuracy), сенімділік (Confidence), қолдау (Support), ROC және AUC көрсеткіштері пайдаланылады.

Интерпретация және қолдану: алынған білім бизнес, медицина, қаржы, өндіріс және мемлекеттік басқару сияқты салаларда шешім қабылдауға негіз болады.

Кіріспе

KDD кезеңдеріне кіріспе

KDD (Knowledge Discovery in Data) – бұл деректер базасынан білімді ашудың кешенді процесі. KDD процесі әдетте бір-бірімен тығыз байланысты бірнеше кезеңдерден тұрады. Бұл үдеріс «лас» немесе шикі деректерді қабылдап, оларды талдауға дайындайды және соңында нақты шешім қабылдауға болатын құнды білімге айналдырады.

KDD процесінің негізгі қадамдары:

1. Деректерді таңдау (Selection): Талдауға қажетті деректерді анықтау және жинақтау.
2. Деректерді алдын ала өңдеу (Pre-processing): Деректерді тазалау және шулардан арылту.
3. Деректерді трансформациялау (Transformation): Деректерді талдауға ыңғайлы форматқа келтіру.
4. Деректерді өндіру (Data Mining/Үлгілеу): Арнайы алгоритмдер арқылы жасырын үлгілерді табу.
5. Бағалау және интерпретациялау (Evaluation and Interpretation): Табылған үлгілерді бағалау және білім ретінде ұсыну.

Бұл лекцияда KDD процесінің ең маңызды үш негізгі кезеңіне, яғни Деректерді алдын ала өңдеу, Үлгілеу (Data Mining) және Бағалауға тоқталамыз.

KDD кезеңдерін түсіндіретін болсақ:

Деректерді алдын ала өңдеу – бұл KDD процесінің ең көп уақытты қажет ететін және маңызды кезеңі, себебі шикі деректер әдетте толық емес, қателі, шулы немесе сәйкес келмейтін форматта болады. Талдаудың сапасы 80% дейін осы кезеңге байланысты.

1. Деректерді тазалау (Data Cleaning)

Деректерді тазалаудың мақсаты – деректердегі қателіктерді, қайшылықтарды және толық еместіктерді жою.

Жетіспейтін мәндерді өңдеу:

- Жою: Жетіспейтін мәндері бар қатарларды (жазбаларды) толығымен алып тастау. Бұл деректердің көп бөлігін жоғалтуы мүмкін.
- Толтыру: Жетіспейтін мәндерді қолмен, орташа мәнмен, медианамен, модамен немесе ең ықтимал мәндерді болжайтын арнайы модельдер (мысалы, регрессия немесе машиналық оқыту) арқылы толтыру.

Шуды жою (Noise Removal):

Шу – бұл өлшеу немесе жинау кезінде туындаған кездейсоқ қателер немесе дисперсиялар.

- Тегістеу: Деректерді тегістеу (binning) әдісі, регрессиялық модельдер немесе кластерлеу арқылы шуды азайту.

Қайшылықтарды өңдеуде әртүрлі дереккөздерде бір объектіге қатысты әртүрлі деректер болса, оларды сәйкестендіру және түзеуді айтады.

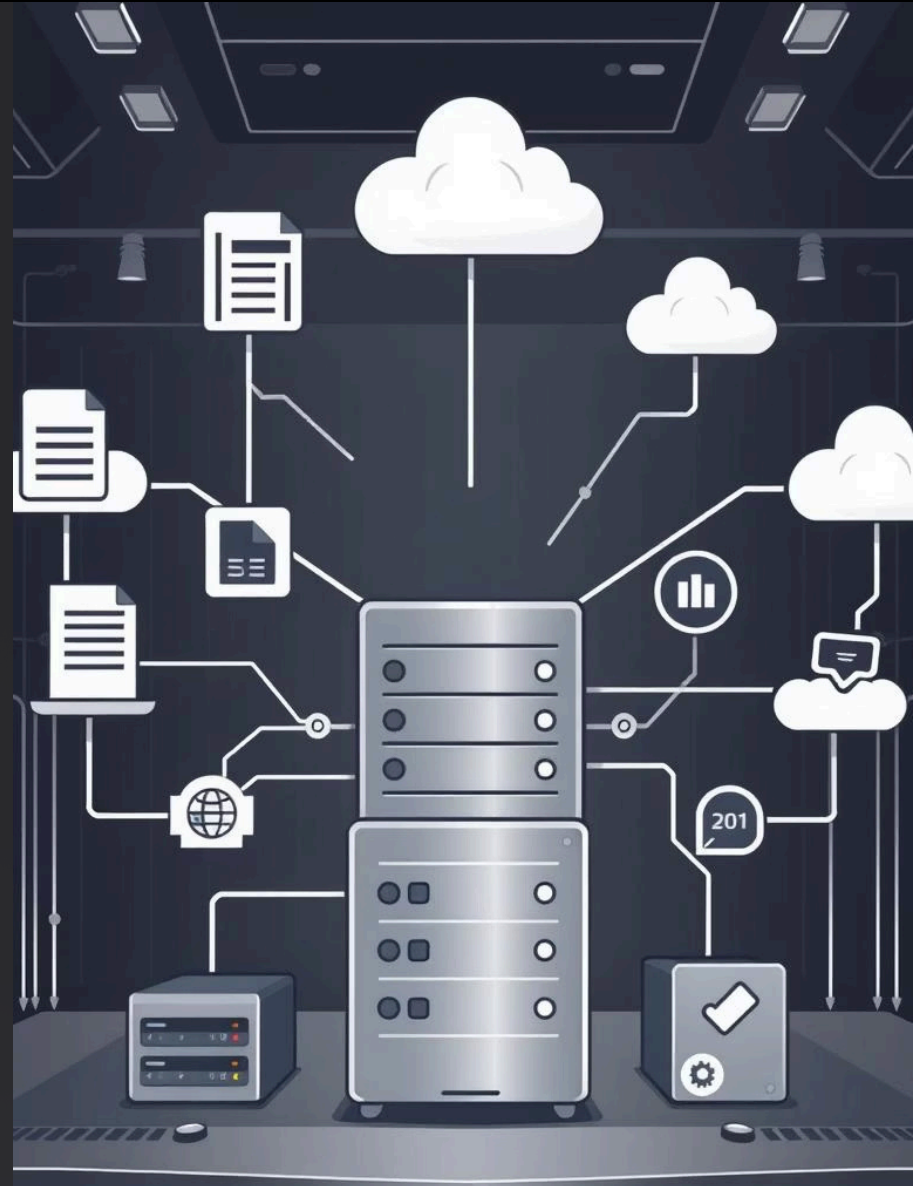
2.1 сурет Деректерді тазалаудың әдістері

2. Деректерді біріктіру (Data Integration)

Бірнеше дереккөздерден (мысалы, бірнеше дерекқордан, файлдардан) алынған деректерді біртұтас, біркелкі деректер қоймасына (Data Warehouse) немесе жиынтығына біріктіру.

- Мәселелер: Бірдей объектілердің әртүрлі аталуы (схемалық интеграция мәселесі) немесе өлшем бірліктерінің сәйкессіздігі. Бұл мәселелерді шешу үшін сәйкестендіру және біріктіру ережелері қолданылады.

2.2 сурет Деректерді біріктіру



3. Деректерді трансформациялау (Data Transformation)

Деректерді өндіру алгоритмдерінің тиімді жұмыс істеуі үшін оларды талдауға қолайлы форматқа түрлендіру.

1. Нормализациялау (Normalization): Деректердің мәндерін белгілі бір диапозонда сақтау. Бұл машиналық оқытудың көптеген алгоритмдері (мысалы, нейрондық желілер, K-NN) үшін маңызды.

Min-Max нормализациясы:

1. Агрегация (Aggregation): Деректерді жинақтау, мысалы, транзакциялық деректерді күнделіктіден айлық жиынтыққа ауыстыру.
2. Дискретизациялау (Discretization): Үздіксіз сандық атрибуттарды шектеулі санды дискретті интервалдарға немесе категорияларға айналдыру.

2.3 сурет Деректерді трансформациялау

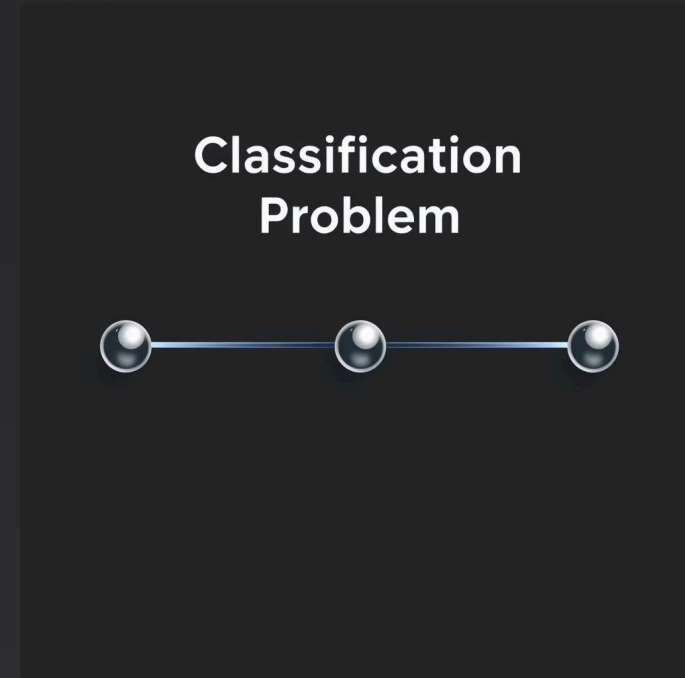
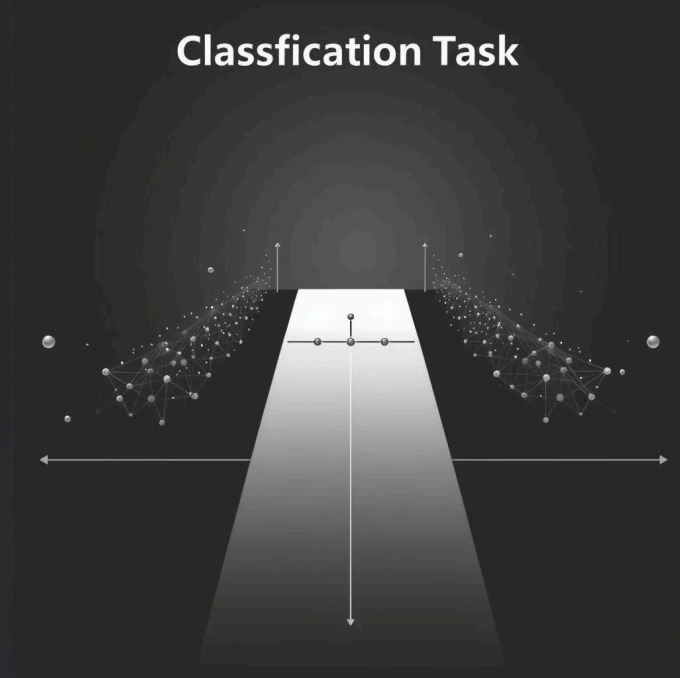
2.4 сурет Деректерді алдын ала өңдеу кезеңдерінің схемасы

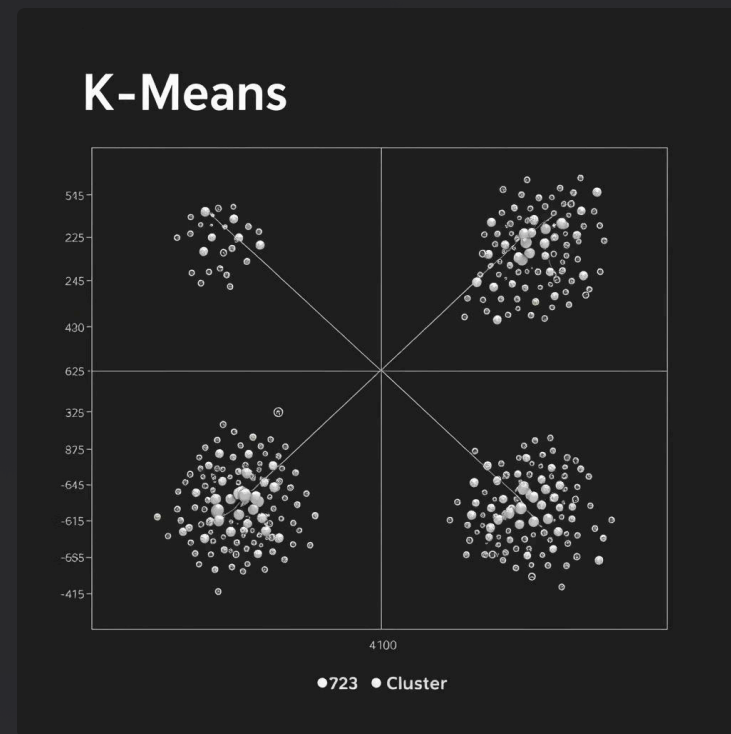
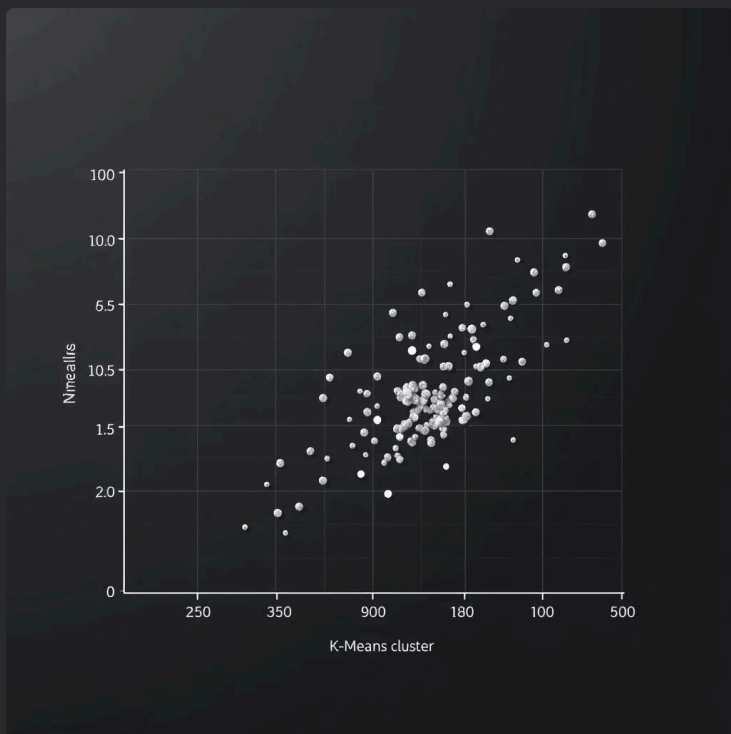
Деректерді өндіру немесе Үлгілеу (Data Mining/Modeling).

Бұл KDD процесінің өзегі. Деректерді өндіру (Data Mining) – алдын ала өңделген деректерден қызықты, пайдалы және бұрын белгісіз үлгілерді шығарып алу үшін интеллектуалды әдістерді қолдану. Бұл кезеңде математикалық және машиналық оқыту алгоритмдері қолданылады.

1. Негізгі Талдау Міндеттері:

- Классификация (Classification): Деректерді берілген үлгілерге сүйене отырып, алдын ала анықталған топтардың біріне жатқызу. Мысалы, электрондық поштаны «Спам» немесе «Спам емес» деп жіктеу. (Қолданылатын әдістер: Шешім ағаштары, Наивті Байес, SVM).
- Регрессия (Regression): Үздіксіз сандық нәтижелерді болжау. Мысалы, үйдің бағасын немесе температураны болжау. (Қолданылатын әдістер: Сызықтық регрессия, Нейрондық желілер).
- Кластерлеу (Clustering): Ұқсас деректер элементтерін топтастыру. Топтар (кластерлер) алдын ала анықталмаған. Мысалы, клиенттерді мінез-құлқына қарай сегменттеу. (Қолданылатын әдістер: K-Means, DBSCAN).
- Қауымдастық ережелерін табу (Association Rule Mining): Деректердегі айнымалылар арасындағы өзара байланыстарды немесе үлгілерді табу. Мысалы, «Нан сатып алған клиенттердің көпшілігі сүт те сатып алады». (Қолданылатын әдістер: Apriori алгоритмі).





2. Үлгілерді Таңдау (Model Selection):

Талдау мақсатына сәйкес ең қолайлы алгоритм (үлгі) таңдалады. Мысалы, болжам жасау үшін регрессия, ал сегменттеу үшін кластерлеу таңдалады.

3. Үлгіні Оқыту (Model Training):

Таңдалған алгоритм (үлгі) деректер жиынтығының бір бөлігінде (оқыту жиынтығы) оқытылады. Бұл кезеңде модель деректердегі жасырын заңдылықтарды «үйренеді».

Бағалау және Интерпретациялау (Evaluation and Interpretation)

Деректерді өндіру кезеңінде табылған үлгілерді немесе құрылған модельдерді бағалау және олардың шынайы жағдайдағы пайдалылығын анықтау.

1. Модельді Бағалау (Model Evaluation):

1. Дәлдік (Accuracy): Модельдің болжамдарының дұрыс пайыздық көрсеткіші.
2. Сенімділік (Confidence) және Қолдау (Support): Қауымдастық ережелері үшін қолданылады. Қолдау – ережені қолдайтын транзакциялардың жалпы пайыздық үлесі, ал Сенімділік – ереженің қаншалықты дұрыс орындалатыны.
3. Қателік Матрицасы (Confusion Matrix): Классификациялық модельдердің жұмысын бағалау үшін қолданылатын кесте (дұрыс оң, жалған оң, жалған теріс, дұрыс теріс нәтижелерді көрсетеді).
4. ROC қисығы және AUC мәні: Классификациялық модельдердің әртүрлі табалдырықтардағы жұмысын көрсететін графикалық құрал.

[Қателік Матрицасының түсініктемесі]

Дұрыс Теріс (True Negative, TN): 61 (Қатерлі деп дұрыс болжанған)

Жалған Оң (False Positive, FP): 2 (Қатерсіз деп қате болжанған)

Жалған Теріс (False Negative, FN): 2 (Қатерлі деп қате болжанған)

Дұрыс Оң (True Positive, TP): 106 (Қатерсіз деп дұрыс болжанған)

2. Білімді Интерпретациялау:

Бағалау кезеңінен өткен, сенімді үлгілер «білім» ретінде қабылданады. Бұл білімнің маңыздылығын және оны іскери мақсаттарға қолдану мүмкіндігін түсіндіру. Мысалы: Кластерлеу нәтижесінде «жаңа, жас және техникалық құралдарға бейім» клиенттер тобы анықталды. Бұл білімді маркетинг бөлімі жаңа өнімдерді ұсыну үшін пайдаланады.

3. Білімді Қолдану (Deployment):

Табылған білімді іскери шешім қабылдау процестеріне немесе ақпараттық жүйелерге енгізу. Бұл KDD процесінің соңғы, бірақ іс жүзіндегі ең маңызды кезеңі.

2.10 сурет KDD процесінің негізгі циклдік схемасы

Қосымша Ақпарат: Деректерді Жинақтау (Data Gathering/Selection) кезеңінің егжей-тегжейі

Деректерді жинақтау кезеңі – KDD процесінің бастамасы. Бұл кезеңде талдауға арналған мақсатты деректер жиынтығы (target dataset) қалыптастырылады.

Негізгі қадамдары:

1. Бизнес мақсатты анықтау: Талдау арқылы қандай мәселені шешу керек екенін немесе қандай сұраққа жауап іздеу керектігін нақтылау. (Мысалы: «Клиенттердің неліктен компаниядан кететінін (churn) анықтау»).
2. Дереккөздерді анықтау: Ішкі (CRM, ERP жүйелері, транзакциялық дерекқорлар) және сыртқы (әлеуметтік желілер, ашық деректер, нарықтық зерттеулер) дереккөздерді анықтау.
3. Деректерді таңдау: Талдау мақсатына сәйкес келетін, өзекті жазбалар мен атрибуттарды іріктеп алу. Бұл кезеңде үлкен деректер базасынан кішігірім, бірақ нақты мәселеге қатысты жиынтық қалыптасады.

Нәтижесі: Талдаудың келесі кезеңдеріне дайын, мақсатты деректер жиынтығы.

Қорытынды

KDD (Knowledge Discovery in Data) процесі, оның ішінде деректерді алдын ала өңдеу, үлгілеу (Data Mining) және бағалау кезеңдері, бүгінгі таңдағы ақпараттық технологиялар мен бизнес-аналитиканың негізгі тірегі болып табылады. Бұл жүйелі қадамдар шикі деректердің құнды, әрі іс жүзінде қолдануға болатын білімге айналуын қамтамасыз етеді. KDD процесінің тиімділігі тек алгоритмдердің күрделілігіне ғана емес, ең алдымен, деректерді алдын ала сапалы өңдеуге байланысты.

Бақылау сұрақтары:

1. KDD процесінің негізгі кезеңдерін атаңыз.
2. Деректерді алдын ала өңдеу не үшін қажет және оның негізгі әдістері қандай?
3. Нормализация мен дискретизацияның айырмашылығы неде?
4. Data Mining кезеңінде қолданылатын негізгі әдістерді атаңыз.
5. Модельдің сапасын бағалау қандай көрсеткіштермен өлшенеді?
6. KDD нәтижелерін қолдану салаларына мысал келтіріңіз.