Тақырып: Деректерді талдау және алдын ала өңдеу

Деректерді тазалау: бос ұяшықтармен және аномалиялармен жұмыс істеу

Деректерді тазалау — деректерді өңдеудің маңызды кезеңі болып табылады. Бұл кезеңде жиналған деректердегі кемшіліктер мен аномалиялар жойылып, оларды талдауға және модельдеуге дайындаймыз. Деректер жиынтығы көп жағдайда толық емес, қате немесе қайталанатын ақпараттармен болуы мүмкін. Мұндай мәселелер деректерді дұрыс талдай алмауға немесе дұрыс шешімдер қабылдауға кедергі болуы мүмкін.

Бос ұяшықтармен жұмыс істеу

Бос ұяшықтар деректер жиынында жиі кездеседі. Олар әртүрлі себептермен пайда болуы мүмкін: деректерді жинау кезінде ақпараттарды жіберіп алу, қателіктер, немесе басқа да себептер. Бос ұяшықтарды өңдеудің бірнеше тәсілі бар:

1. Бос ұяшықтарды жою:
   Бос ұяшықтарды жою дегеніміз — деректер жиынынан бос ұяшықтары бар жолдарды немесе бағандарды алып тастау. Бұл әдіс деректерді тазалауда жиі қолданылады, себебі бос ұяшықтар көптеген аналитикалық процестерде кедергі келтіруі мүмкін.Алайда, бұл тәсіл деректердің көп бөлігін жоғалтуға әкелуі мүмкін.
2. Бос ұяшықтарды толтыру:
   Бос ұяшықтарды толтыру үшін түрлі әдістер қолданылады:
   - Орташа мәнмен толтыру: Сандардың орташа мәнімен немесе медианасымен толтыру.
   - Алдыңғы немесе келесі мәнмен толтыру: Алдыңғы немесе келесі мәнмен толтыру (Forward/Backward Fill).
   - Модамен толтыру: Категориялық деректер үшін ең жиі кездесетін мәнмен толтыру.

Мысалы, егер мәліметтердің орташа мәнін есептесек:

$$\text{Mean}(V) = \frac{\sum_{i=1}^{n} V_i}{n}$$

мұндағы (V_i) — деректер жиынындағы жеке мәндер, (n) — деректердің саны.

Ал егер мәліметтердің орташа мәнін есептесек:

Егер n жұп болса (яғни, деректер саны тақ емес), онда медиана екі ортаңғы мәннің орташа арифметикалық мәні болады:

$$\text{Медиана} = \frac{v_{\left(\frac{n}{2}\right)} + v_{\left(\frac{n}{2}+1\right)}}{2}$$

Егер n тақ болса (яғни, деректер саны тақ болса), онда медиана тек ортаңғы мән болады:

$$\text{Медиана} = v_{\left(\frac{n+1}{2}\right)}$$

Аномалияларды өңдеу

Аномалиялар — деректер жиынында күтілмейтін немесе ерекше мәндер. Олар дұрыс емес өлшеулер, техникалық ақаулар немесе қателер болуы мүмкін. Аномалияларды өңдеу әдістері:

- Аномалияларды анықтау: Қате деректерді анықтау үшін әртүрлі статистикалық және визуализациялық әдістер қолданылады.
- Аномалияларды жою немесе түзету: Аномалияларды жою немесе оларды басқа мәндермен (мысалы, орташа мәндермен) ауыстыру.

.Деректерді қалыпқа келтіру және стандарттау

Деректерді қалыпқа келтіру мен стандарттау — деректердің ауқымын біркелкі ету және оларды модельдерге дайындау үшін қажетті кезеңдер. Бұл кезеңдер деректердің әртүрлі өлшемдерде және бірлікке ие болуына байланысты маңызды.

Қалыпқа келтіру (Min-Max Normalization)

Қалыпқа келтіру әдісі деректерді 0 мен 1 аралығында шкалаға келтіру үшін қолданылады. Бұл әдіс үшін келесі формула қолданылады:

$$V_{\text{normalized}} = \frac{V - V_{\text{min}}}{V_{\text{max}} - V_{\text{min}}}$$

мұндағы:

- (V) — өңделетін дерек,
- (V_min) — деректер жиынындағы ең кіші мән,
- (V_max) — деректер жиынындағы ең үлкен мән.

Стандарттау (Z-score Standardization)

Стандарттау әдісі деректерді нөлдік орташа мән мен бірлік стандартты ауытқуға келтіру үшін қолданылады. Бұл әдіс деректердің таралуын тұрақтандыруға көмектеседі. Стандарттау үшін келесі формула қолданылады:

$$Z = \frac{V - \mu}{\sigma}$$

мұндағы:

- (V) — деректің мәні,
- (mu) — деректер жиынындағы орташа мән,
- (sigma) — деректердің стандартты ауытқуы.

Бұл әдіс деректерді әртүрлі диапазондарда болған жағдайда да оларды салыстыруға мүмкіндік береді.

Деректерді оқу және бөлшектеу (Train/Test бөлу)

Деректерді оқу және бөлшектеу — модельдерді оқыту және бағалау үшін маңызды кезең. Бұл кезеңде деректер екі бөлікке бөлінеді: бір бөлігі модельді үйрету үшін (train), ал екіншісі модельдің тиімділігін тексеру үшін (test) қолданылады.

Деректерді оқу

Деректерді оқу — бұл деректерді дереккөздерден (мысалы, CSV файлдары, базалар) оқу. Бұл деректерді Python сияқты тілдерде өңдеу үшін Pandas, Numpy сияқты кітапханалар қолданылады.

Мысалы, Pandas арқылы деректерді оқу:

```
import pandas as pd

data = pd.read_csv('data.csv')
```

Деректерді бөлшектеу (Train/Test бөлу)

Деректер жиынын train және test жиынтықтарына бөлу маңызды. Әдетте, деректердің 70%-дан 80%-на дейінгі бөлігі үйретуге, ал қалғаны тексеруге бөлінеді.

Бұл үшін келесі әдістер қолданылады:

- Train/Test бөлу: Деректерді кездейсоқ бөлу арқылы үйрету және тестілеу жиынтықтарын алу.
- K-Fold кросс-валидация: Деректерді K бөлікке бөлу және әрбір бөлікке сәйкес келетін модельді бағалау.

Мысалы, Train/Test бөлу әдісі Python-да Scikit-learn кітапханасы арқылы жүзеге асырылады:

```
from sklearn.model_selection import train_test_split

X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)
```

Бұл жерде:

- X — деректердің ерекшеліктері (features),
- y — мақсатты айнымалы (target),
- test_size — тест жиынтығының мөлшері (мұнда 20%).

Қорытынды

Деректерді өңдеу және алдын ала өңдеу — бұл машиналық оқыту мен статистикалық талдаудағы өте маңызды кезеңдер. Деректердің сапасын қамтамасыз ету, оларды қалыпқа келтіру және оларды дұрыс форматта модельдерге беру — нәтижелі болашақ болжамдар үшін негіз болып табылады. Әрбір кезеңнің дұрыс жүзеге асырылуы талдаудың дәлдігіне және нәтиженің сапасына әсер етеді.

Қолданылған әдебиеттер

1. Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques* (3rd ed.). Elsevier.
2. Aggarwal, C. C. (2015). *Data Mining: The Textbook*. Springer.
3. Iglewicz, B., & Hoaglin, D. C. (1993). *How to detect and handle outliers*. Sage Publications.