

№10 Дәріс. Сызықтық регрессия (Linear Regression)

Дәріс мақсаты:

Сызықтық регрессия модельдерінің теориялық негіздерін, математикалық аппаратын және практикалық қолдану әдістерін түсіндіру. Негізгі мақсаттар: деректер арасындағы тәуелділікті анықтау, регрессиялық модель құру, модель параметрлерін бағалау, нәтижені интерпретациялау және болжам жасау.

Кілттік сөздер:

Сызықтық регрессия, регрессия коэффициенті, тәуелді айнымалы, тәуелсіз айнымалы, корреляция, OLS, ең кіші квадраттар әдісі, болжам, қателік, детерминация коэффициенті, шығын функциясы, мультиколлинеарлық, overfitting, underfitting.

Дәріс жоспары:

1. Кіріспе: Регрессияның мәні мен мақсаты
2. Сызықтық регрессия түсінігі
3. Математикалық моделі
4. Ең кіші квадраттар әдісі (OLS)
5. Модель сапасын бағалау көрсеткіштері
6. Бір айнымалы және көп айнымалы регрессия
7. Қате талдауы және болжам
8. Регрессияның шектеулері және мәселелері
9. Практикалық қолдану мысалдары
10. Қорытынды

1. Кіріспе

Қазіргі таңда әлемде өндірілетін ақпарат көлемі қарқынды өсуде. Күн сайын компаниялар, ғылыми орталықтар, банктер, денсаулық сақтау ұйымдары, электрондық коммерция платформалары және әлеуметтік желілер үлкен көлемдегі деректерді жинақтап, сақтап және өңдеп жатыр. Бұл деректерден құнды ақпарат бөліп алу және бизнес, ғылым, техника салаларында дұрыс шешім қабылдау үшін заманауи аналитикалық әдістер қажет. Осындай әдістердің ең кең тарағандарының бірі — сызықтық регрессия (Linear Regression).

Сызықтық регрессия — статистика мен машиналық оқытудың іргелі әдісі, деректер арасындағы байланысты математикалық модель түрінде сипаттайды. Мысалы, пәтердің бағасын оның аумағына, орналасқан жеріне және қабат санына байланысты болжауға болады. Сол сияқты, адамның салмағын бойына қарай, өнім сатылымын жарнама бюджетіне қарай, немесе экономикалық көрсеткіштерді тарихи мәндерге қарай болжауға болады.

Бұл әдістің қарапайымдылығы оны көптеген ғылым салаларының негізгі құралына айналдырды: экономика, медицина, инженерия, маркетинг, экология және әлеуметтік ғылымдарда сызықтық тәуелділікті модельдеу

үшін қолданылады. Деректер ғылымында сызықтық регрессия көбіне алғашқы модель ретінде қолданылады, себебі оның нәтижесі түсінікті, параметрлері интерпретациялауға оңай, ал математикалық негіздері қарапайым.

Сонымен қатар, сызықтық регрессияны терең меңгеру — машиналық оқытудың күрделірек әдістерін үйренудің алғашқы маңызды қадамы. Ridge, Lasso, ElasticNet сияқты регрессияның кеңейтілген түрлері, сонымен қатар желілік модельдер, нейрондық желілер сияқты күрделі модельдер де регрессия концепцияларынан бастау алады.

Осы дәрісте біз сызықтық регрессияның теориялық негіздерін, математикалық моделін, есептеу тәсілдерін және практикалық қолдануын қарастырамыз. Сонымен қатар, модельдің шектеулерін түсініп, модель сапасын бағалау әдістерін оқимыз.

Мысалы:

- Тұрғын үй бағасын ауданына қарай болжау
- Студент бағасын оқу сағатына қарай болжау
- Бизнес табысын жарнама шығынына тәуелді анықтау
- Климат өзгерісін температура деректері арқылы зерттеу

Сызықтық модель қарапайым болғанымен, көптеген аналитикалық жүйелердің негізі.

2. Сызықтық регрессия түсінігі

Сызықтық регрессия — деректер арасындағы сызықтық тәуелділікті анықтап, ең жақсы түзу сызықты табуға бағытталған модель.

Тәуелді айнымалы: y

Тәуелсіз айнымалы: x

3. Математикалық моделі

Бір айнымалы регрессия:

мұнда $y = \beta_0 + \beta_1 x + \epsilon$

- β_0 — еркін мүше (intercept)
- β_1 — регрессия коэффициенті (сызықтың еңісі, slope)
- ϵ — қателік

Көп айнымалы регрессия:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \epsilon$$

4. Ең кіші квадраттар әдісі (OLS)

OLS тәсілін қолдану арқылы қателіктің квадраты минималданады.

Мақсат:

$$\min \sum (y_i - \hat{y}_i)^2$$

5. Модель сапасын бағалау көрсеткіштері

- MSE — орташа квадраттық қателік

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

RMSE

$$RMSE = \sqrt{MSE}$$

- **MAE** — орташа абсолюттік қателік
- R^2 — детерминация коэффициенті (0–1)

$$R^2 = 1 - \frac{SSR}{SST}$$

6. Бір айнымалы мен көп айнымалы

Түрі	Түсінігі
Simple Linear	$y = \beta_0 + \beta_1 x$
Multiple Linear	$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$

7. Қателер және шектеулер

Мәселелер:

- Мультиколлинеарлық
- Шектен тыс сәйкестендіру (overfitting)
- Түзу арқылы үлгіленбейтін байланыстар
- Шуды жоғары сезімталдығы

8. Практикалық қолдану салалары

Сала	Мысал
Банкинг	Несие төлем қабілетін болжау
Медицина	Диагноз болжамы
Экономика	Нарықтық баға болжау
Инженерия	Сенсорлық жүйе анализі
Маркетинг	Жарнама тиімділігін бағалау

Қорытынды

Сызықтық регрессия — статистикалық моделдеу мен машиналық оқытудың ең маңызды әдістерінің бірі. Оның қарапайымдылығы, интерпретациясының жеңілдігі және тиімділігі оны зерттеушілер мен инженерлер арасында кеңінен қолданылатын модельге айналдырды. Бұл модель тәуелді айнымалы мен тәуелсіз айнымалылар арасындағы сызықтық байланысты анықтап, нақты болжамдар жасауға мүмкіндік береді.

Дәріс барысында біз сызықтық регрессияның математикалық негіздерін, ең кіші квадраттар әдісі арқылы параметрлерді есептеу алгоритмін, модель сапасын бағалау көрсеткіштерін және практикалық қолдану мысалдарын қарастырдық. Сонымен қатар, регрессия моделінің шектеулері мен мүмкін болатын қателіктеріне тоқталдық, мысалы, мультиколлинеарлық, overfitting, сызықтық емес байланыстар және мәліметтердегі шудың әсері.

Сызықтық регрессия тек болжам жасау үшін ғана емес, деректердің құрылымын, тәуелділіктерді және әсер ету факторларын түсіну үшін де қолданылады. Мысалы, қандай факторлардың нәтижеге көбірек әсер ететінін, деректер арасында қандай байланыс бар екенін модель параметрлері арқылы интерпретациялау мүмкіндігі — осы әдістің артықшылығы.

Болашақта бұл білімді кеңейту үшін Ridge, Lasso регрессиясы, логистикалық регрессия, нейрондық желілер және жалпы машиналық оқыту әдістерін оқуға болады. Python тілінде Pandas, Scikit-learn кітапханаларымен тәжірибе жасап, шынайы деректермен жұмыс істеу дағдыларын дамыту ұсынылады.

Қорытындылай келе, сызықтық регрессия — қарапайым әрі қуатты құрал. Ол аналитикалық ойлау қабілетін дамытады және деректер ғылымы саласында сенімді бастапқы қадам болып табылады.

Бақылау сұрақтары

1. Сызықтық регрессия дегеніміз не?
2. Бір және көп айнымалы регрессияның айырмашылығы қандай?
3. OLS әдісінің мәні неде?
4. R^2 метрикасы нені көрсетеді?
5. Overfitting дегеніміз не?
6. Қандай салаларда сызықтық регрессия қолданылады?

Пайдаланылған әдебиеттер

1. James, G., Witten, D., Hastie, T., Tibshirani, R. *An Introduction to Statistical Learning*, Springer
2. Montgomery, D.C. *Introduction to Linear Regression Analysis*
3. Goodfellow I., Bengio Y., Courville A. *Deep Learning*
4. Bishop C. *Pattern Recognition and Machine Learning*
5. Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn & TensorFlow*