

№12 Дәріс. Машиналық оқытудағы шешім ағаштары

Дәріскер: Даркенбаев Д.К. PhD, қауымдастырылған
профессор



Дәрістің мақсаты

Студенттерге шешім ағаштары алгоритмдерінің құрылымы мен қағидаларын, классификация және регрессия міндеттеріндегі қолданылуын, модельді құру принциптері мен артықшылықтары мен кемшіліктерін түсіндіру.

Кілттік сөздер

Шешім ағашы, Decision Tree, Classification Tree, Regression Tree, Gini индексі, Энтропия, Information Gain, Overfitting, Pruning, CART, ID3, C4.5, Random Forest, Feature importance.

Дәріс жоспары

1. Шешім ағашына жалпы түсінік
2. Шешім ағаштарының түрлері
 - o Классификация ағаштары
 - o Регрессия ағаштары
3. Ақпараттық критерийлер (бөліну метрикалары)
 - o Энтропия
 - o Information Gain
 - o Gini индексі
4. Ағаштың құрылу принципі
5. Артықшылықтары мен кемшіліктері
6. Араласып кету (Overfitting) мәселесі
 - o Pruning (кесу әдістері)
7. Шешім ағаштарын жақсарту әдістері
8. Python (sklearn) мысалы
9. Қорытынды
10. Бақылау сұрақтары

Кіріспе

Қазіргі заманда ақпарат көлемінің үздіксіз артуы, цифрландыру процестерінің жеделдеуі, ақылды жүйелердің дамуы — деректерге негізделген шешім қабылдаудың маңызын айрықша арттырды. Data Science және Machine Learning салаларында дамыған модельдер мен алгоритмдер күнделікті өмірімізге еніп, бизнес, медицина, қаржы, өндіріс, білім беру және мемлекеттік басқару салаларында кең қолданыс табуда.

Машиналық оқыту әдістері ішінде **шешім ағаштары (Decision Trees)** ерекше орын алады. Себебі олар ғылыми тұрғыдан да, практикалық қолданыста да қарапайым, түсінікті және қолдануға жеңіл. Бұл модель адам логикасына ұқсас түрде «сұрақ–жауап» принципімен жұмыс істейді, сондықтан ол интуитивті түрде түсінікті әрі нәтижелерді түсіндіру оңай.

Шешім ағаштары классификация мен регрессия міндеттерін шешуде тиімді, олар деректердің түріне қатты тәуелді емес, сандық та, категориялық та ақпаратпен жұмыс істей алады. Сонымен қатар Feature scaling (нормализация, стандарттау) қажет емес, бұл модельді қолдануды одан әрі жеңілдетеді.

Бүгінде шешім ағаштары тек өз бетінше ғана емес, сондай-ақ Random Forest, Gradient Boosting, XGBoost, CatBoost сияқты қуатты ансамбльдік алгоритмдердің негізі болып табылады. Бұл әдістер әлемдегі ең тиімді машиналық оқыту модельдері қатарына кіреді.

Осы дәрісте шешім ағаштарының негізгі принциптері, ақпараттық өлшемдер (энтропия, Gini индексі), модельді кесу (pruning), артықшылықтары мен кемшіліктері, және нақты өмірдегі қолданылу мысалдары қарастырылады. Бұл тақырыпты игеру студентке болашақ күрделі алгоритмдерді оңай меңгеруге негіз болады.

Теориялық бөлім



1. Шешім ағашына түсінік

Шешім ағашы — деректерді бөлу арқылы шешім қабылдайтын құрылым.



Түбір түйін

root node



Ішкі түйіндер

decision nodes



Жапырақтар

leaf nodes — нәтиже

Әрбір бөлігі:

- Түбір түйін (root node)
- Ішкі түйіндер (decision nodes)
- Жапырақтар (leaf nodes) — нәтиже



1. Шешім ағаштарының түрлері

Түрі	Міндеті	Шығыс мәні
Classification Tree	Классификация	Сынып (0/1/...)
Regression Tree	Регрессия	Нақты сан

1. Ақпараттық критерийлер



Энтропия

Деректердің тәртіпсіздігін өлшейді. Неғұрлым біртекті болса, энтропия төмен.



Information Gain

Бөлінуден кейін энтропияның қанша азайғанын көрсетеді. Жоғары мән — жақсы бөліну.



Gini индексі

Кездейсоқ таңдалған элементтің қате классификациялану ықтималдығын өлшейді.



3. Барлық жапырақтарға дейін қайталау

Әрбір бөлік үшін процесті рекурсивті түрде қайталау





1. Артықшылықтары мен кемшіліктері

Артықшылық

- Түсінікті, интерпретациясы жеңіл
- Категориялық, сандық деректермен жұмыс істейді
- Feature scaling қажет емес
- Жылдам

Кемшілік

- Overfitting
- Шуды сезімтал қабылдайды
- Кішкене өзгеріс → үлкен өзгеріс
- Ағаш терең болса күрделенеді

OVERFEX MODEL



TRAINING

TEST DATA

SIMPLE
MODEL

TEEN DATA

1. Overfitting — басты проблема

Шешімі: Pruning (кесу)



Pre-pruning

depth, leaves, min samples



Post-pruning

cost complexity pruning



Pre-pruning: Ағаш өсу кезінде шектеулер қою (тереңдік, жапырақтар саны, минималды үлгілер)

Post-pruning: Толық ағаш құрғаннан кейін артық тармақтарды кесу



1. Ensemble әдістері

Random Forest

Көптеген шешім ағаштарын біріктіріп, дауыс беру арқылы нәтиже шығарады

Gradient Boosting

Қателерді түзету арқылы кезекпен ағаштар құрады

XGBoost, LightGBM, CatBoost

Оптимизацияланған, жылдам және қуатты boosting алгоритмдері

Бұл әдістер шешім ағаштарының күшін арттырып, overfitting мәселесін азайтады және жоғары дәлдікке қол жеткізеді.

Python (Scikit-learn) мысалы

```
from sklearn.tree import DecisionTreeClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score

# Деректерді бөлу
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)

# Модельді құру
model = DecisionTreeClassifier(max_depth=5, min_samples_split=10)
model.fit(X_train, y_train)

# Болжам жасау
y_pred = model.predict(X_test)

# Дәлдікті бағалау
accuracy = accuracy_score(y_test, y_pred)
print(f"Accuracy: {accuracy}")
```

❏ **max_depth:** Ағаштың максималды тереңдігі

min_samples_split: Бөлу үшін қажетті минималды үлгілер саны



Қорытынды

Шешім ағаштары — машиналық оқытудағы ең маңызды және кеңінен қолданылатын әдістердің бірі. Олар қарапайымдылығымен, нәтижелерді түсіндіру мүмкіндігімен және әртүрлі типтегі мәліметтермен тиімді жұмыс істеуімен ерекшеленеді. Қарапайым логикалық құрылым арқылы күрделі процестердің шешімдерін табуға мүмкіндік береді.

Сонымен қатар шешім ағаштары — қазіргі заманғы көптеген күшті ансамбльдік әдістердің негізі. Random Forest және Gradient Boosting алгоритмдерінің жоғары нәтижелі болуы — шешім ағаштарының тиімділігі мен әмбебаптығын дәлелдейді.

Шешім ағаштарының басты проблемасы — **overfitting**, яғни модельдің тым күрделеніп, үйрену деректеріне шамадан тыс бейімделуі. Дегенмен, pruning әдістері, параметрлерді дұрыс баптау және ансамбльдік қосымшалар арқылы бұл мәселе тиімді шешіледі.

Қорыта айтқанда, шешім ағаштары:

- Интерпретациясы анық,
- Есептеу тұрғысынан тиімді,
- Әртүрлі салаларда қолданылатын,
- Базалық, әрі іргелі ML алгоритмі болып саналады.

Осы модельді жақсы меңгеру — экономикалық талдау, қаржылық модельдеу, биомедицина, өндірістік автоматтандыру, қауіпсіздік жүйелері және интеллектуалды басқару жүйелері сияқты салаларда нақты тапсырмаларды шешуге мүмкіндік береді.

Шешім ағаштарын толық түсіну — болашақта күрделі машиналық оқыту және жасанды интеллект алгоритмдерін меңгерудің сенімді қадамы.

Бақылау сұрақтары

01

Шешім ағашы деген не?

02

Классификация және регрессия ағаштарының айырмашылығы қандай?

03

Gini индексінің формуласы қандай?

04

Энтропия қай кезде қолданылады?

05

Overfitting деген не?

06

Pruning әдісі не үшін қолданылады?

07

Decision Tree қандай жағдайда тиімді емес болуы мүмкін?

Пайдаланылған әдебиеттер

1. Hastie, Tibshirani, Friedman — The Elements of Statistical Learning
2. James et al. — An Introduction to Statistical Learning
3. Aurélien Géron — Hands-On Machine Learning with Scikit-Learn & TensorFlow
4. Scikit-Learn Documentation — <https://scikit-learn.org>
5. Tom Mitchell — Machine Learning, McGraw-Hill

❏ **Ескерту:** Бұл әдебиеттер машиналық оқыту саласындағы классикалық және қазіргі заманғы еңбектер болып табылады. Олар шешім ағаштары мен басқа ML алгоритмдерін тереңірек зерттеуге мүмкіндік береді.

