

Шешімдер Ағаштары және Ансамбльдік Әдістер

Машиналық оқытудың екі маңызды әдісі — Шешімдер ағаштары мен
Ансамбльдік әдістер туралы толық шолу.

Оқытушы: Сарсембаева Т.С.

Дәрістің Мақсаты мен Машиналық Оқытудағы Орны



Меңгеру мақсаты

Шешімдер ағашының құрылымы мен жұмыс қағидаларын, сондай-ақ Random Forest және Gradient Boosting сияқты ансамбльдік әдістердің мәнін, артықшылықтары мен айырмашылықтарын түсіну.



МО-дағы маңызы

Классификация және регрессия мәселелерін шешудегі осы әдістердің орны мен маңыздылығын анықтау.



Қысқаша шолу

Шешімдер ағашының қарапайымдылығы және ансамбльдік әдістердің дәлдік пен тұрақтылықты арттырудағы рөлі.

Шешімдер Ағашының Негіздері

Шешімдер ағашы — бақылау деректерін белгілерге қарай шарттармен бөлу арқылы шешім қабылдайтын модель.

1

Ұғымы

Деректер жинағын қарапайым шешім қабылдау ережелері арқылы бөлетін ағаш тәрізді құрылым. Әрбір түйін бір белгіге негізделген сұрақты, ал әрбір тармақ сұрақтың жауабын білдіреді.

2

Құрылымы

- **Түбір түйін (Root Node):** Ағаштың басталу нүктесі, барлық деректерді қамтиды.
- **Ішкі түйіндер (Internal Nodes) және тармақтар (Branches):** Деректерді әрі қарай бөлетін шарттарды білдіреді.
- **Жапырақ түйіндер (Leaf Nodes):** Соңғы шешімді немесе классификацияны көрсететін түйіндер.

3

Жұмыс қағидасы

Ағаш жоғарғы түйіннен бастап, деректерді әрбір ішкі түйіндегі шарттарға сәйкес бөліп отырады, соңында жапырақ түйіндеріндегі ақырғы сыныптарға немесе мәндерге жеткізеді.

Түйінді Бөлу Критерийлері: Тазалықты Арттыру

Шешімдер ағашын құрудағы басты мақсат — әрбір бөлуде деректердің "тазалығын" максималды түрде арттыру. Тазалық – бұл бір түйіндегі деректердің қаншалықты біртекті екенін көрсететін өлшем.

Джини индексі

Классификация үшін қолданылады. Топтағы кездейсоқ таңдалған екі элементтің әртүрлі сыныптарға жату ықтималдығын өлшейді. Индекс неғұрлым аз болса, тазалық соғұрлым жоғары болады.

$$Gini = 1 - \sum_{i=1}^C (p_i)^2$$

Мұнда p_i – i -ші сыныптың түйіндегі үлесі, C – сыныптар саны.

Энтропия және Ақпараттық Өсім

Энтропия – деректердің "тәртіпсіздік" немесе "белгісіздік" дәрежесін өлшейді. Жоғары энтропия дегеніміз – деректердің біртекті емес екендігі.

$$Entropy = - \sum p_i \log_2(p_i)$$

Ақпараттық өсім – бөлу нәтижесіндегі энтропияның төмендеу мөлшері. Бөлу арқылы ақпараттық өсім неғұрлым жоғары болса, сол бөлу соғұрлым тиімді болады.

Түйінді Бөлу Критерийлері: Ең жақсы бөлуді таңдау

Ағаштың әрбір түйінінде деректерді бөлу үшін ең жақсы белгі мен бөлу нүктесін таңдау қажет. Бұл таңдау тазалықты барынша арттыратын немесе энтропияны барынша төмендететін критерийге негізделеді.



Шешімдер Ағашының Артықшылықтары мен Кемшіліктері

Артықшылықтары

- **Интерпретациялау оңай:** Ағаш құрылымы визуалды түрде түсінікті және қарапайым. Адамдар үшін шешім қабылдау процесін бақылау оңай.
- **Алдын-ала өңдеу аз:** Деректерді нормалау немесе стандарттау қажет емес.
- **Әртүрлі деректермен жұмыс:** Сандық және категориялық деректермен жұмыс істей алады.
- **Масштабтауға төзімді:** Үлкен деректер жинақтарымен де тиімді жұмыс істейді.

Кемшіліктері

- **Артық оқыту (Overfitting):** Терең ағаштар оқу деректеріндегі шуды немесе ерекшеліктерді жаттап алып, жаңа деректерде нашар жұмыс істеуі мүмкін. Бұл модельдің жалпылау қабілетін төмендетеді.
- **Тұрақсыздық:** Оқыту деректеріндегі кішігірім өзгерістер ағаштың құрылымын түбегейлі өзгертіп жіберуі мүмкін. Бұл модельді сенімсіз етеді.
- **Оптималды ағаш табу:** Ең оптималды ағаш құрылымын табу NP-толық мәселе болып табылады.

Бұл кемшіліктерді жою үшін көбінесе **ансамбльдік әдістерге** жүгінеміз.

Ансамбльдік Әдістердің Идеясы

Ансамбльдік әдістер — бірнеше «әлсіз» модельдің нәтижелерін біріктіріп, одан да дәлірек әрі тұрақты болжамдар жасауға бағытталған қуатты тәсіл.



Әлсіз Модельдер

Жеке-жеке алғанда болжамдық қабілеті шектеулі немесе дәл емес модельдер (мысалы, бір шешім ағашы).



Нәтижелерді біріктіру

Әрбір модельдің болжамын біріктіру арқылы (дауыс беру, орташаландыру) жалпы нәтижені күшейту.



Дәлірек және Тұрақты Нәтиже

Ансамбль жеке модельдерге қарағанда артық оқытуға төзімдірек және жалпылау қабілеті жоғары модель құрады.

Екі негізгі топ:

Бэггинг (Bagging)

Тәуелсіз модельдерді параллель оқыту және олардың нәтижелерін біріктіру. Әр модель деректер жиынтығының әртүрлі ішкі жиынтығында оқытылады (мысалы, Random Forest).

Бустинг (Boosting)

Модельдерді бірізді оқыту, мұнда әрбір келесі модель алдыңғы модельдің қателіктерін түзетуге бағытталған (мысалы, Gradient Boosting).

Random Forest (Кездейсоқ Орман) Алгоритмі

Random Forest — бэггинг стратегиясына негізделген ансамбльдік әдіс, ол көптеген шешімдер ағаштарынан тұрады.



Bootstrap Sampling (Бэггинг)

Бастапқы деректер жиынтығынан қайталап іріктеу арқылы бірнеше (N) ішкі жиынтықтар (bootstrap samples) құрылады. Әрбір ішкі жиынтық бастапқы деректер жиынтығымен бірдей мөлшерде болады, бірақ оның құрамы әртүрлі болуы мүмкін.



Ерекшеліктерді кездейсоқ таңдау

Әрбір шешім ағашын құру кезінде, әрбір түйінді бөлу үшін қолжетімді белгілердің кішігірім, кездейсоқ ішкі жиынтығы ғана қарастырылады. Бұл ағаштардың бір-бірінен тәуелсіз болуына көмектеседі.



Болжам жасау

Классификация мәселелері үшін барлық жеке ағаштардың болжамдары бойынша көпшілік дауыс қолданылады. Регрессия мәселелері үшін болжамдардың орташа мәні алынады.

Артықшылықтары:

Артық оқытуға төзімділік, жоғары тұрақтылық, өлшемділікке төзімділік (үлкен белгілер санымен жұмыс істей алады).

Feature Importance:

Random Forest моделі модель болжамына әрбір ерекшеліктің қосқан салымын анықтауға мүмкіндік береді, бұл модельді түсінуге көмектеседі.

Gradient Boosting (Градиентті Бустинг) Алгоритмі

Gradient Boosting — бустинг стратегиясына негізделген, модельдерді бірізді оқыту арқылы қателіктерді төмендетуге бағытталған тиімді алгоритм.

	Алғашқы модель Алғашқы әлсіз модельмен (көбінесе қарапайым шешім ағашымен) бастау.
	Қателіктерді есептеу Модельдің қателіктерін (қалдықтарын) есептеу, бұл дұрыс болжанбаған деректер.
	Қателерді болжау Келесі модельді алдыңғысының қателіктерін болжауға үйрету. Бұл қателіктерді азайтуға бағытталған.
	Қателіктерді төмендету Градиенттік түсу (Gradient Descent) принципі бойынша қателіктерді біртіндеп азайту. Жаңа болжам: $F_m(X) = F_{m-1}(X) + \eta \cdot h_m(X)$, мұнда $h_m(X)$ – жаңа модель, η – оқыту жылдамдығы.

Артықшылықтары:

Жоғары дәлдік, қателерді тиімді төмендету мүмкіндігі. Бәсекелестік машиналық оқыту турнирлерінде жиі жеңіске жетеді.

Кемшіліктері:

Артық оқытуға сезімталдығы жоғары (егер параметрлер дұрыс реттелмесе), параллельдендіру қиын (бірізділікке байланысты).

Мысал алгоритмдері:

XGBoost, LightGBM, CatBoost сияқты алгоритмдер Gradient Boosting-тің озық имплементациялары болып табылады.

Қолданылу Салалары және Қорытынды

Шешімдер ағаштары мен ансамбльдік әдістердің кең ауқымды қолданылуы олардың әмбебаптығы мен тиімділігін көрсетеді.

Қаржы Несие тәуекелін бағалау (Random Forest/Gradient Boosting).		Денсаулық сақтау Ауру диагностикасы (Шешімдер ағашы).
Маркетинг Клиенттердің кетуін болжау (Random Forest).		E-commerce Сатып алу ықтималдығын болжау (Gradient Boosting).

Қорытынды

- **Шешімдер ағашы:** Негізгі, интерпретацияланатын модель, қарапайымдылығымен және түсініктілігімен ерекшеленеді.
- **Random Forest:** Тұрақтылық пен артық оқытудан қорғау үшін тиімді, көптеген ағаштардың коллективтік даналығын пайдаланады.
- **Gradient Boosting:** Ең жоғары болжамдық дәлдік үшін қолданылады, бірақ параметрлерді мұқият реттеуді талап етеді.

Бақылау сұрақтары

1. Шешімдер ағашындағы "артық оқыту" мәселесін қалай түсіндіресіз?
2. Джини индексі мен ақпараттық өсімнің негізгі айырмашылығы неде?
3. Random Forest пен Gradient Boosting алгоритмдерінің негізгі айырмашылықтары мен артықшылықтарын атаңыз: