# K - Ең жақын көршілер. Ағаш үлгілері.

## Статистикалық Машиналық оқыту

Статистикадағы соңғы жетістіктер болжамды модельдеу — регрессия және жіктеу саласындағы неғұрлым қуатты автоматтандырылған әдістерді әзірлеуге арналды. Бұл әдістер статистикалық машиналық оқытудың неғұрлым жалпы әдіснамасының ажырамас бөлігі болып табылады және классикалық статистикалық әдістерден ерекшеленеді, өйткені олар деректермен басқарылады және деректерді сызықтық немесе басқа жалпы функциямен сипаттауға ұмтылмайды. Мысалы, ең жақын sose - dei әдісі өте қарапайым: ол жазбаны жазбалардың қаншалықты ұқсас екендігіне қарай жіктейді. Шешім ағаштарына қатысты ансамбльді оқытуға сүйенудің ең сәтті және кеңінен қолданылатын әдістері. Ансамбльдік оқытудың негізгі идеясы-болжамды қалыптастыру үшін бір модельге қарағанда көптеген модельдерді қолдану. Шешім ағаштары-болжамды айнымалылар мен нәтиже айнымалылары арасындағы байланыстар туралы ережелерді үйренуге арналған икемді және автоматты әдіс. Ансамбльдік оқытудың шешім ағаштарымен үйлесуі болжамды модельдеудің жоғары нәтижелі стандартты әдістеріне әкеледі екен.

## K- жақын көршілер

Жақын көршілердің K әдісі (KNN, ағылш. k-nearest neighbors) өте қарапайым идея жіктелетін немесе болжанатын әрбір жазба үшін:

1. K-ға ұқсас белгілері бар жазбаларды табу (яғни, дикторларға дейінгі ұқсас мәндер).

2. Жіктеу үшін: осы ұқсас жазбалардың арасынан мажоритарлық сыныпты анықтап, осы сыныпты жаңа жазбаға тағайындау.

3. Болжау үшін (KNN регрессиясы деп те аталады): осы ұқсас жазбалардың арасынан орташа мәнді тауып, жаңа жазбаның орташа мәнін болжау.

***Негізгі терминдер:***

- *Көрші (көрші) -болжалды мәндері басқа жазбаға ұқсас жазба.*
- *Қашықтықтың метрикалық көрсеткіштері (қашықтық метрикасы)- бір жазбаның екіншісінен қаншалықты алыс екенін бір санмен қорытындылайтын метрикалық көрсеткіштер.*
- *Стандарттау (стандарттау)- орташа мәнді алып тастап, стандартты ауытқуға бөлу.*

*Синоним: қалыпқа келтіру z-бағалау (z-score) стандарттаудан кейін алынған мән.*
*Синоним: стандартты бағалау.*

- *K - Жақын көршілердің алгоритмін есептеу кезінде ескерілетін көршілердің саны*

KNN-болжаудың / жіктеудің қарапайым әдістерінің бірі: сәйкес келетін модель жоқ. Бұл KNN пайдалану Автоматты процедура дегенді білдірмейді. Болжау нәтижелері белгілердің қалай талданғанына, ұқсастықтың қалай өлшенгеніне және K шамасы қандай болатынына байланысты. Сонымен қатар, барлық болжаушылар сандық түрде болуы керек. Біз осы әдістің жұмысын жіктеу мысалымен суреттейміз.

**Шағын мысал: несиені қайтармауды болжау.**

6.1-Кестеде. Lending Club инвестициялық несие компаниясында жеке несие деректерінің бірнеше жазбалары ұсынылған. Lending Club инвесторлар жеке тұлғаларға жеке несиелер беретін тең құқықты несиелеудің көшбасшысы болып табылады. Талдаудың мақсаты жаңа әлеуетті несиенің нәтижесін болжау болады: қайтару қарсы өтелген

6.1. Кесте. Lending Club инвестициялық несие компаниясының несие деректерінен бірнеше жазбалар мен бағандар

| Исход | Величина ссуды | Доход | Цель | Стаж работы | Домовладение | Штат |
|---|---|---|---|---|---|---|
| Погашено | 10 000 | 79 100 | Консолидация долга | 11 | ИПОТЕКА | NV |
| Погашено | 9600 | 48 000 | Переезд | 5 | ИПОТЕКА | TN |
| Погашено | 18 800 | 120 036 | Консолидация долга | 11 | ИПОТЕКА | MD |
| Невозврат | 15 250 | 232 000 | Малый бизнес | 9 | ИПОТЕКА | CA |
| Погашено | 17 050 | 35 000 | Консолидация долга | 4 | АРЕНДА | MD |
| Погашено | 5500 | 43 000 | Консолидация долга | 4 | АРЕНДА | KS |

Тек екі болжамды айнымалысы бар өте қарапайым модельді қарастырыңыз: dti, яғни қарыз бойынша төлемдердің кіріске қатынасы (ипотеканы қоспағанда) және payment_inc_ratio, яғни несие бойынша төлемдердің кіріске

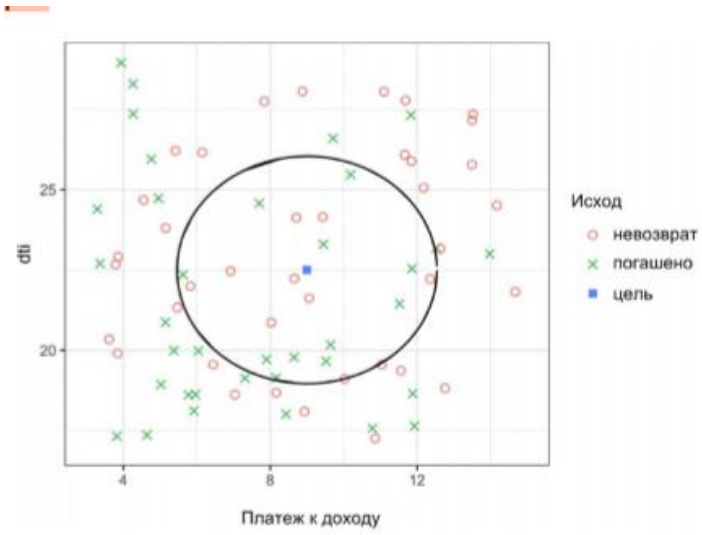
қатынасы. Екі қатынас 100-ге көбейтіледі. Белгілі екілік нәтижелері бар 200 loan 200 несиелерінің шағын жиынтығы негізінде (outcome200 болжауында берілген оның қарама - қарсылығын қайтармады) және 20-да орнатылған K, dti=22.5 және payment_inc_ratio=9 көмегімен болжанатын жаңа несиенің newloan ұпайы.

```
library(FNN) knn_pred <-
 knn(train=loan200, test=newloan, cl=outcome200, k=20)
 knn_pred == 'default'
[1] TRUE
```

KNN болжамы-несие қайтарылмайды.

R өзінің knn функциясына ие болғанымен, үшінші тараптың Fnn R бағдарламалық жасақтамасы (fast nearest neighbor-жылдам көрші) үлкен деректерге жақсырақ масштабтайды және көбірек икемділік береді.

6.2 -Суретте. бұл мысалдың визуалды көрінісі берілген. Болжауға болатын жаңа несие орталықтағы квадратпен ұсынылған. Шеңберлер (қайтарылмайтын) және крестер (сөндірілген) - бұл жаттығу деректері. Сопақ ең жақын 20 нүктенің шекарасын көрсетеді. Бұл жағдайда 14 қайтарылмайтын несие тек 6 өтелген несиемен салыстырғанда сопақ ішінде жатыр. Демек, несиенің болжамды нәтижесі қайтарылмайды.




6.2.-Сурет. Екі айнымалыны қолдана отырып, KNN несиені қайтармау алгоритмін болжау: қарыздың кіріске қатынасы және несие бойынша төлемдердің кіріске қатынасы

Қашықтықтың метрикалық көрсеткіштері

Ұқсастық (жақындық) қашықтықтың метрикалық көрсеткіші арқылы анықталады, яғни екі жазбаның қаншалықты алыс екенін өлшейтін функция (x1 , x2, ... x p) және (u1, u2 , ... up) бір-бірінен. Екі вектор арасындағы қашықтықтың ең танымал метрикалық көрсеткіші-евклидтік қашықтық. Екі вектор арасындағы евклидтік қашықтықты өлшеу үшін екіншісінен біреуін алып тастау керек, айырмашылықтарды квадраттап, оларды қорытындылап, квадрат түбірді алу керек

$$\sqrt{(x_1-u_1)^2+(x_2-u_2)^2+\ldots+(x_p-u_p)^2}.$$

Евклидтік қашықтық арнайы есептеу артықшылықтарын ұсынады. Бұл, әсіресе, үлкен деректер жиынтығы үшін өте маңызды, өйткені KNN k×n жұптық салыстыруды ұсынады, мұндағы n — жолдар саны.

Сандық мәліметтер үшін тағы бір қабылданған метрика-Манхэттен қашықтығы (немесе Минковский қашықтығы):

$$|x_1-u_1|+|x_2-u_2|+\ldots+|x_p-u_p|.$$

Евклидтік қашықтық екі нүкте арасындағы түзу сызыққа сәйкес келеді (олар айтқандай, "қарға қалай ұшады"). Манхэттен қашықтығы-бір уақытта бір бағытта қиылысатын екі нүкте арасындағы қашықтық (мысалы, тікбұрышты қалалық блоктар бойымен қозғалу). Осы себепті Манхэттен қашықтығы пайдалы жуықтау болып табылады, егер ұқсастық жолдағы нүктелік уақыт ретінде анықталса. Екі вектор арасындағы қашықтықты өлшеуде салыстырмалы түрде үлкен шкала бойынша өлшенетін айнымалылар (белгілер) бұл өлшемде басым болады. Мысалы, несие деректері үшін қашықтық тек ондаған немесе жүздеген мыңмен өлшенетін кіріс пен несие сомасының айнымалыларының функциясы болар еді. Салыстыру кезінде қатынастарға негізделген айнымалылар іс жүзінде жойылады. Бұл мәселе деректерді стандарттау арқылы шешіледі

**Бір белсенді күйі бар кодтаушы**

6.1 Кестедегі несиелер туралы мәліметтер бірнеше факторлық (жолдық) айнымалыларды қамтиды. Статистикалық және машиналық оқыту модельдерінің көпшілігі айнымалының бұл түрін кесте сияқты бірдей ақпаратты тасымалдайтын екілік жалған айнымалылар қатарына

түрлендіруді талап етеді. 6.2. Үй иесінің мәртебесін білдіретін бір айнымалының орнына: "ипотекамен иелік етеді", "ипотекасыз иелік етеді", "жалға алады" немесе "басқа", біз төрт екілік белбеуге келеміз. Біріншісі "ипотекамен — Y/N", екіншісі "ипотекасыз — Y/N" және т. б. болады. Бұл бір болжаушы, үй иесінің мәртебесі, осылайша статистикалық және машиналық оқыту алгоритмдерінде қолдануға болатын 1 және 0 векторын тудырады. "Бір белсенді күйді кодтау" (one hot encoding) тіркесі цифрлық интегралды Чип терминологиясынан шыққан, онда ол тек бір биттің оң (белсенді)болуына рұқсат етілген Чип конфигурациясын сипаттайды.

6.2.-Кесте Сандық жалған айнымалы арқылы үй иесі туралы факторлық деректерді ұсыну

| ИПОТЕКА | ДРУГОЕ | ВЛАДЕЛЕЦ | АРЕНДА |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 1 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 |
| 0 | 0 | 0 | 1 |

**Стандарттау (қалыпқа келтіру, z-бағалау)**

Өлшеу нәтижесінде алынған мәліметтерде бізді көбінесе олардың мөлшері емес, олардың орташадан қаншалықты ерекшеленетіні қызықтырады. Стандарттау немесе қалыпқа келтіру процедурасы барлық айнымалыларды орташа мәнді алып тастау және стандартты ауытқуға бөлу арқылы ұқсас шкалаларға орналастырады. Осылайша, біз айнымалының бастапқы өлшеу шкаласына байланысты модельге шамадан тыс әсер етпейтініне кепілдік береміз.

$$z = \frac{x - \bar{x}}{s}.$$

Стандарттау нәтижесінде алынған шамалар әдетте стандартты бағалау немесе z-бағалау деп аталады. Өлшеу деректері одан әрі"орташадан стандартты ауытқуларда" қолданылады. Осылайша, айнымалының модельге әсері оның бастапқы өлшеу шкаласына әсер етпейді.

KNN және басқа да бірнеше процедуралар үшін (мысалы, негізгі компоненттерді талдау және кластерлеу) процедураны қолданар алдында

деректерді стандарттауды ескеру өте маңызды. Бұл идеяны көрсету үшін KN date және payment_inc_ratio көмегімен кемелер туралы мәліметтерге қолданылады және басқа екі өзгермелі: revol_bal — өтініш берушіге доллармен қол жетімді жалпы жаңартылатын несие және revol_util — пайдаланылған несиенің пайызы. Жаңа болжамды жазба төменде көрсетілген:

*new loan*
*payment_inc_ratio dti revol_bal revol_util*
*1 2.3932 1 1687 9.4*

Доллармен есептелетін rival_ball шамасы басқа айнымалыдан әлдеқайда үлкен. Ln функциясы nn атрибуты сияқты ең жақын көршілердің индексін қайтарады.Индекс және оны loan_df деректер кадрындағы ең жақын бес жолды көрсету үшін пайдалануға болады:

*loan_df <- model.matrix(~ -1 + payment_inc_ratio + dti + revol_bal + revol_util, data=loan_data)*
*knn_pred <- knn(train=loan_df, test=newloan, cl=outcome, k=5)*
*loan_df[attr(knn_pred,"nn.index"),]*
*payment_inc_ratio dti revol_bal revol_util*
*36054 2.22024 0.79 1687 8.4*
*33233 5.97874 1.03 1692 6.2*
*28989 5.65339 5.40 1694 7.0*
*29572 5.00128 1.84 1695 5.1*
*20962 9.42600 7.14 1683 8.6*

Бұл көршілердегі rival_ball мәні оның жаңа жазбадағы мәніне өте жақын, бірақ басқа болжамды айнымалылар шашыраңқы және көршілерді анықтауда маңызды рөл атқармайды.
Мұны әр айнымалы үшін z бағасын есептейтін scale R-функциясын қолдана отырып, стандартталған деректерге қолданылатын KNN-мен салыстырайық:
*loan_std <- scale(loan_df) knn_pred <-*
*knn(train=loan_std, test=newloan_std, cl=outcome, k=5)*
*loan_df[attr(knn_pred,"nn.index"),]*
*payment_inc_ratio dti revol_bal revol_util*
*2081 2.61091 1.03 1218 9.7*
*36054 2.22024 0.79 1687 8.4*
*23655 2.34286 1.12 523 10.7*

*41327 2.15987 0.69 2115 8.1*
*39555 2.76891 0.75 2129 9.5*

Ең жақын бес көрші барлық айнымалыларда әлдеқайда ұқсас, бұл ақылға қонымды нәтиже береді. Нәтижелер бастапқы мектепте көрсетілгенін ескеріңіз, бірақ KNN талданған мәліметтерге және болжанған жаңа несиеге қолданылды.

## K Таңдау

K таңдау KNN өнімділігі үшін өте маңызды. Ең оңай таңдау K =1 орнату , бұл 1-ші жақын көршінің фикаторына сәйкес келеді. Болжам интуитивті: ол жаңа болжамды жазбаға ұқсас жазбаның жаттығу жиынтығында болу негіздері болып табылады. Негіз ретінде қабылдау K =1 сирек ең жақсы таңдау болып табылады; сіз әрқашан K >1 жақын көршілерді пайдалану арқылы жоғары өнімділікке ие боласыз. Жалпы айтқанда, егер k мәні тым төмен болса, онда біз артық жарысты тудыруы мүмкін: модельге деректердегі шуды қосу арқылы. Жоғары k мәндері жаттығу деректерінде қайта жарамдылық қаупін азайтатын тегістеуді қамтамасыз етеді. Екінші жағынан, егер K тым жоғары болса, онда біз деректердің шамадан тыс тегістелуіне әкеліп соқтырамыз және KNN - дің деректердегі жергілікті құрылымды түсіру мүмкіндігін жіберіп аламыз - оның басты артықшылықтарының бірі. Қайта сәйкестендіру мен өте тегістеу арасындағы жақсы тепе - теңдікті сақтайтын k мәні әдетте дәлдік метрикалық көрсеткіштерімен және атап айтқанда, кешіктірілген деректермен немесе кросс - валидациямен бақылау үлгісіне негізделген дәлдікпен анықталады. K-ны жақсы білуге қатысты жалпы ереже жоқ-бәрі негізінен деректердің табиғатына байланысты. Шуы аз жоғары құрылымдалған деректер үшін кіші k мәндері жақсы жұмыс істейді. Сигналдарды өңдеу аймағынан терминді ала отырып, деректердің бұл түрі кейде жоғары сигнал/кедергі қатынасы (SNR, signal-to-noise ratio) деп аталады. Жоғары SNR деректерінің мысалдары әдетте қолжазба мен сөйлеуді тануға арналған деректер болып табылады. Несие деректері сияқты құрылымы төмен (SNR деректері төмен) шулы деректер үшін үлкенірек k мәндері орынды болады.әдетте, k мәндері 1-ден 20-ға дейінгі диапазонға түседі. Дауыс беру кезінде дауыстардың теңдігін болдырмау үшін тақ сан сирек таңдалмайды.

## KNN әдісі атрибут құрастырушы ретінде

KNN әдісі өзінің қарапайымдылығы мен интуитивті табиғатына байланысты танымал болды. Тиімділік тұрғысынан KNN әдетте неғұрлым жетілдірілген жіктеу әдістерімен салыстырғанда конкурентке қабілетті емес. Практикалық жағдайларда модельдерді сәйкестендіру кезінде KNN басқа жіктеу әдістерімен көп сатылы процесте "жергілікті білімді" қосу үшін қолданыла алады. 1. KNN деректерде орындалады және әрбір жазба үшін класс - сификация нәтижесі (немесе сыныптың квази ықтималдығы) қалыптасады. 2. Бұл нәтиже жазбаға жаңа белгі ретінде қосылады, содан кейін деректерде тағы бір жіктеу әдісі орындалады. Бастапқы болжау ре-белбеулер осылайша екі рет қолданылады. Алдымен бұл процесс мультиколлинеарлыққа байланысты проблеманы тудыратынына күмәндануға болады, өйткені кейбір болжаушылар оны екі рет қолданады (бөлімді қараңыз. "Мультиколлинеарлық" 4 тарау). Бұл проблема емес, өйткені екінші кезең моделіне енгізілген ақпарат өте Жергілікті, тек бірнеше көрші жазбалардан алынған және сондықтан артық ақпарат емес, Қосымша ақпарат болып табылады.

Мысалы, Кинг округінің тұрғын үй қорының деректерін қарастырыңыз. Үйді сату бағасын белгілеу кезінде жылжымайтын мүлікті сату агенті бағаны жақында сатылған ұқсас үйлерге негіздейді, олар"сату - ана-журналдар" деп аталады. Негізінде, жылжымайтын мүлік агенттері KNN-дің қолмен нұсқасын орындайды: ұқсас үйлердің сату бағасына қарап, олар Boo - det үйі не үшін сатылғанын бағалай алады. Біз KNN - ді соңғы сатылымдарға өзгерту арқылы жылжымайтын мүлік саудасының маманына еліктейтін статистикалық модель үшін жаңа белгі жасай аламыз. Болжалды мән сату бағасы болып табылады және қолданыстағы болжамды айнымалыларға орналасу орны, жалпы шаршы метр, құрылым түрі, жер көлемі және жатын бөлмелері мен жуынатын бөлмелердің саны кіруі мүмкін. Біз KNN арқылы қосатын жаңа болжау айнымалысы (белгі) - әрбір жазба үшін KNN болжаушысы (жылжымайтын мүлік агенттеріндегі аналогтық сатылымдарға ұқсас). Болжалды мән сандық болғандықтан, жақын көршілердің орташа K (KNN регрессиясы деп аталады) мажоритарлық мемлекеттік лосингтің орнына қолданылады. Сол сияқты, несие деректері үшін біз несие беру процесінің әртүрлі жақтарын білдіретін белгілерді жасай аламыз. Мысалы, келесі код үзіндісі қарыз алушының несиелік қабілетін білдіретін белгіні жасайды:

*borrow_df <- model.matrix(~ -1 + dti + revol_bal + revol_util + open_acc + delinq_2yrs_zero + pub_rec_zero, data=loan_data)*

```
borrow_knn <- knn(borrow_df, test=borrow_df, cl=loan_data[, 'outcome'], prob=TRUE, k=10)
prob <- attr(borrow_knn, "prob")
 borrow_feature <- ifelse(borrow_knn=='default', prob, 1-prob)
summary(borrow_feature)
   Min. 1st Qu.  Median    Mean 3rd Qu.   Max.
0.0000   0.4000  0.5000  0.5012  0.6000 1.0000
```

нәтиже-бұл жағдайдың сенімділігін болжайтын белгі, қарыз алушы несие тарихына сүйене отырып, несиені қайтармайды.

***К көршілеріне арналған негізгі идеялар:***

- *Жақын көршілердің k әдісі (KNN) жазбаны ұқсас жазбаларға жататын сыныпқа жатқызу арқылы жіктейді.*
- *Ұқсастық (қашықтық) евклидтік қашықтықпен немесе басқа ұқсас метрикалық көрсеткіштермен анықталады.*
- *Жазбаны салыстыратын жақын көршілердің Саны, к, алгоритмнің әртүрлі К мәндерін қолдана отырып, жаттығу деректерінде қаншалықты жақсы нәтиже көрсететіндігімен анықталады.*
- *Әдетте, болжамды айнымалылар стандартталған, нәтижесінде үлкен масштабты айнымалылар метрикалық қашықтық көрсеткішінен басым болмайды.*
- *Болжалды модельдеуде KN бірінші кезеңде жиі қолданылады және болжамды мән деректерге екінші (KNN емес) кезеңде модельдеудің болжаушысы ретінде қосылады.*

## Ағаш модельдері

Ағаш модельдері, жіктеу және регрессия ағаштары деп аталады. Шешім ағаштары, немесе жай ағаштар-бұл тиімді және танымал жіктеу әдісі, бастапқыда 1984 жылы Лео Брейман жасаған ағаш модельдері және олардың күшті ағындары кездейсоқ ормандар және бустинг регрессия үшін де, жіктеу үшін де деректер ғылымында ең көп қолданылатын және қуатты болжамды модельдеу құралдарының негізін құрайды.

***Негізгі терминдер:***

** Рекурсивті сегменттеу (recursive partitioning) әрбір қорытынды кіші бөлімде барынша біртекті нәтижелерді жасау мақсатында деректерді бөлімдер мен кіші бөлімдерге бірнеше рет бөлу.*
** Бөлу нүктесіндегі мән (split value) жазбаларды осы болжаушы кішірек және бөлу нүктесіндегі мәннен үлкен жерлерге бөлетін болжаушы мәні.*
** Түйін (түйін) шешім ағашында немесе тиісті тармақталу ережелерінің жиынтығында түйін графикалық немесе ереже түрінде бөлу нүктесіндегі мәнді көрсету болып табылады.*
** Жапырақ (жапырақ) "егер-онда" форматындағы ережелер жиынтығының соңы немесе ағаштың бұтақтары, яғни жапыраққа әкелетін ережелер ағаштың кез келген жазбасы үшін жіктеу ережелерінің бірін қамтамасыз етеді.*
** Жоғалту (жоғалту) бөлу процесінде белгілі бір кезеңдегі қате жіктеу нәтижелерінің Саны; шығындар неғұрлым көп болса, гетерогенділік соғұрлым көп болады.*
** Гетерогенділік (impurity) деректер бөліміндегі сыныптардың араласу дәрежесі (аралас неғұрлым көп болса, гетерогенділік соғұрлым көп болады).*
*Синонимдер: гетерогенділік, арамдық.*
*Антонимдер: біртектілік, тазалық, біртектілік.*
** Кесу (pruning) қайта орнатуды азайту мақсатында толық өсірілген ағаштың бұтақтарын трансляциялық кесу процесі.*

**Ағаш Моделі**-бұл түсіну және жүзеге асыру үшін "егер-онда - басқаша" түрінің импликация ережелерінің жиынтығы. Регрессия мен логистикалық регрессиядан айырмашылығы, ағаштар деректердегі күрделі өзара әрекеттесулерге сәйкес келетін жасырын үлгілерді (суреттер, үлгілер) анықтау қабілетіне ие. Сонымен қатар, CNN немесе аңғал Байес классификаторынан айырмашылығы, қарапайым ағаш үлгілерін оңай түсіндіруге болатын болжаушылар арасындағы байланыстар тұрғысынан көрсетуге болады.

***Қарапайым мысал***
R - де ағаш тәрізді сәндеуге арналған екі негізгі бағдарламалық пакет бар — lei-rpart және tree. Rpart пакетінің көмегімен модель payment_inc_ratio және borrower_score айнымалыларын қолдана отырып, кемелер туралы 3000 жазбаның үлгісіне бейімделеді

*library(rpart)*

*loan_tree <- rpart(outcome ~ borrower_score + payment_inc_ratio, data=loan_data, control = rpart.control(cp=.005))*
*plot(loan_tree, uniform=TRUE, margin=.05)*
*text(loan_tree)*

Алынған ағаш 6.3-суретте көрсетілген. Бұл жіктеу ережелері иерархиялық ағашты айналып өтіп, тамырдан бастап, жапырақ тигенше орнатылады.


6.3.-сурет. Несие деректеріне сәйкес келетін қарапайым ағаш үлгісінің ережелері

Әдетте, ағаш төңкеріліп көрсетіледі, осылайша тамыр жоғарғы жағында, ал жапырақтары төменгі жағында болады. Мысалы, егер біз қарыз алушының borrower_score ұпайы 0,6-ға тең және төлемдер мен payment inc_ratio кірісінің коэффициенті 8,0-ге тең несие алсақ, онда біз сол жақтағы параққа келеміз және несие өтеледі деп болжаймыз.
Ағаштың құрылымдық басып шығарылған нұсқасын жасау да қиын емес:
*loan_tree*
*n= 3000*
*node), split, n, loss, yval, (yprob)*
** denotes terminal node*
*1) root 3000 1467 paid off (0.5110000 0.4890000)*
*2) borrower_score>=0.525 1283 474 paid off (0.6305534 0.3694466)*
*4) payment_inc_ratio< 8.772305 845 249 paid off (0.7053254 0.2946746) **
*5) payment_inc_ratio>=8.772305 438 213 default (0.4863014 0.5136986)*
*10) borrower_score>=0.625 149 60 paid off (0.5973154 0.4026846) **
*11) borrower_score< 0.625 289 124 default (0.4290657 0.5709343) **
*3) borrower_score< 0.525 1717 724 default (0.4216657 0.5783343)*

*6) payment_inc_ratio< 9.73236 1082 517 default (0.4778189 0.5221811)*
*12) borrower_score>=0.375 784 384 paid off (0.5102041 0.4897959) **
*13) borrower_score< 0.375 298 117 default (0.3926174 0.6073826) **
*7) payment_inc_ratio>=9.73236 635 207 default (0.3259843 0.6740157) **

Ағаштың тереңдігі шегініспен көрсетілген. Әрбір түйін берілген сегменттегі басым нәтижемен анықталған алдын ала жіктеуге сәйкес келеді. "Теру" - бұл сегменттегі алдын-ала жіктеу нәтижесінде пайда болатын қате жіктеу нәтижелерінің саны. Мысалы, 2-түйінде 1467 жазбаның жалпы санынан 474 қате жіктеу нәтижелері болды. Жақшадағы мәндер сәйкесінше өтелген және қайтарылмайтын кемелер туралы жазбалардың үлесін білдіреді. Мысалы, қайтарылмауды болжайтын 13-түйінде жазбалардың 60% — дан астамы қайтарылмайтын несиелер болып табылады.

## Рекурсивті сегменттеу алгоритмі

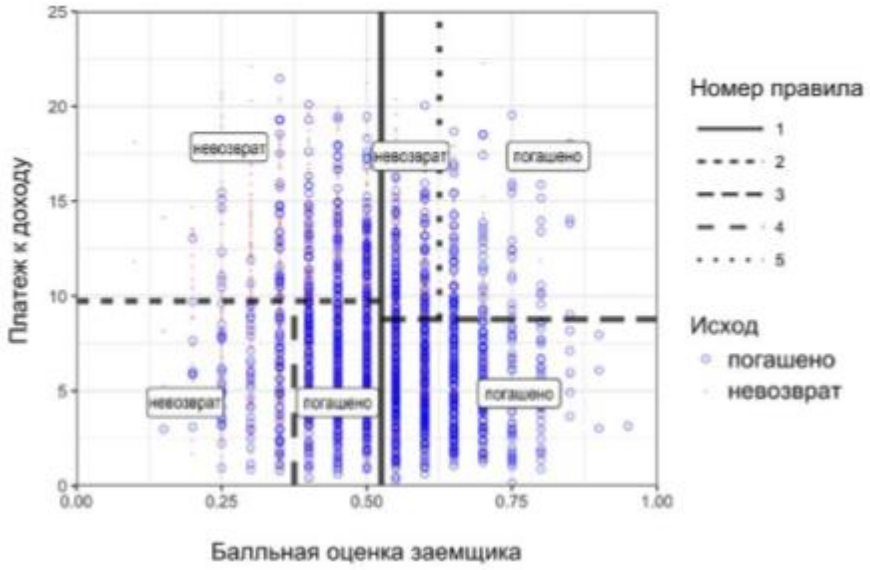
Шешім ағашын құру үшін рекурсивті сегменттеу алгоритмі өте қарапайым және интуитивті. Деректер деректерді салыстырмалы түрде біртекті сегменттерге бөлу үшін қолдан келгеннің бәрін жасайтын болжаушылардың мәндері арқылы бірнеше рет бөлінеді. 6.4 -Суретте. суреттегі ағаш үшін жасалған сегменттердің 6.3-суреті ұсынылған.. Бірінші ереже
borrower_score >= 0.525 графикте 1 нөмірімен көрсетілген. Екінші ереже payment_inc_ratio < 9.732 оң жақ аймақты екіге бөледі.
Бізде Y жауап айнымалысы және 1, үшін j x болжамды айнымалыларының P жиынтығы бар делік ...,jP = . Жазбалары бар A сегменті үшін рекурсивті сегменттеу алгоритмі A ны екі ішкі сегментке бөлудің жақсы әдісін табады:

1. Әрбір болжамды айнымалы үшін X j : •
X j ішінен әрбір S j мәні үшін :

жазбаларды Xj < Sj мәндерімен бір сегментке және қалған жазбаларға, мұндағы Xj≥S j басқа сегментке жатқызыңыз;

әрбір A ішкі сегментіндегі сыныптардың біртектілігін өлшеңіз • * сыныптың максималды сегментішілік біртектілігін тудыратын js мәнін таңдаңыз.
2. J x айнымалысын және сыныптың максималды сегментішілік біртектілігін тудыратын j s бөлу мәнін таңдаңыз. Енді рекурсивті бөлікке кезек келеді: 1. A-ны барлық деректер жиынтығымен инициализациялаңыз. 2. A-ны екі ішкі сегментке, 1 A және 2 A-ға бөлу үшін сегменттеу алгоритмін қолданыңыз .

3. 1 А және 2 а ішкі сегменттерінде 2-қадамды қайталаңыз .
4. Алгоритм сегменттердің біртектілігін жеткілікті түрде жақсартатын кез-келген қосымша сегмент құру мүмкін болмаған кезде аяқталады.




6.4.-Сурет. Несие деректеріне сәйкес келетін қарапайым ағаш үлгісінің ережелері

Соңғы нәтиже 6.4-суреттегідей деректерді сегменттеу., Р-өлшемдерін қоспағанда, әр сегмент осы сегменттегі жауаптың көпшілік дауыс беруіне байланысты 0 немесе 1 нәтижесін болжайды.

## Біртектілікті немесе гетерогенділікті өлшеу

Ағаш тәрізді модельдер Y = 0 немесе Y =1 нәтижесін болжайтын А сегменттерін (жазбалар жиынтығын) рекурсивті түрде жасайды . Алдыңғы алгоритмнен сегменттегі сыныптың тазалығы деп аталатын біркелкілікті өлшеу әдісі қажет екенін көруге болады. Немесе сол сияқты, біз сегменттің гетерогенділігін өлшеуіміз керек. Аңыздардың дәлдігі-бұл сегменттің ішіндегі қате жіктелген жазбалардың р үлесі, ол 0-ден (мінсіз) 0,5-ке дейін (таза кездейсоқ болжам).
Дәлдік гетерогенділіктің жақсы өлшемі емес екен. Оның орнына гетерогенділіктің тағы екі шарасы қабылданды-Гетерогенділік коэффициенті Джини және энтропия, немесе ақпарат. Бұл (және басқа) шаралар гетерогенді жаңалықтар екіден көп сыныптары бар жіктеу тапсырмаларына қолданылады, біз екілік жағдайға назар аударамыз. А жазбалар жиынтығы үшін Джинидің гетерогенділік коэффициенті келесідей:

$$I(A) = p(1-p).$$

Энтропиялық Өлшем келесі формуламен берілген:

$$I(A) = -p\log_2(p) - (1-p)\log_2(1-p).$$

6.5 -суретте. Джиннидің гетерогенділік өлшемі (қайта масштабталған) және энтропия өлшемі ұқсас, ал энтропия орташа және жоғары дәлдік деңгейлері үшін гетерогенділіктің жоғары бағаларын береді.

Гетерогенділіктің метрикалық көрсеткіші бұрын сипатталған сегменттеу алгоритмінде қолданылады. Әрбір ұсынылған деректерді бөлу үшін гетерогенділік бөлу нәтижесінде алынған әрбір сегмент үшін есептеледі. Содан кейін өлшенген орташа мән есептеледі және (әр қадамда) ең төменгі өлшенген орташа мәнді беретін кез келген сегмент таңдалады.

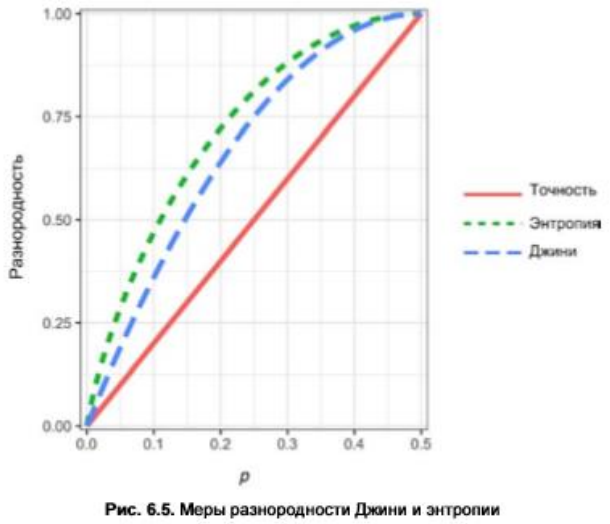


Рис. 6.5. Меры разнородности Джини и энтропии

6.5. -сурет. Джини мен энтропияның гетерогенділік шаралары

**Ағаштың өсуін тоқтату**

Ағаш өсіп келе жатқанда, бөлу ережелері егжей - тегжейлі болады және ағаш бірте - бірте деректердегі нақты және сенімді байланыстарды анықтайтын "үлкен" ережелерді танудан тек шуды көрсететін "кішкентай" ережелерге ауысады. Толық өсірілген ағаш мүлдем таза парақтарға әкеледі, сондықтан ол үйретілген деректерді жіктеуде 100% дәлдікке әкеледі. Бұл дәлдік, әрине, иллюзиялық — біз тым жақын сәйкестікті орындадық (бөлімдегі "мещысу

мен дисперсия Арасындағыисаға келу "жазбасын қараңыз. Осы тараудың басында "k таңдау") жаңа деректерде анықтағымыз келетін сигналға емес, жаттығу деректеріндегі шуға бейімделген деректерге.

- Бізге жаңа деректерге қорытындыларды қорытындылайтын кезеңде ағаш өсіруді қашан тоқтату керектігін анықтаудың қандай да бір әдісі қажет. Деректерді бөлуді тоқтатудың екі жалпы қабылданған әдісі бар.

- Егер алынған ішкі сегмент немесе терминал парағы тым кішкентай болса, сегменттің бөлінуіне жол бермеңіз. Rpart-та бұл шектеулер тиісінше 20 және 7 әдепкі мәндері бар minsplit және minbucket параметрлерімен басқарылады.

Егер жаңа сегмент гетерогенділікті "айтарлықтай" төмендетпесе, сегментті бұзбаңыз. Rpart-та бұл CP күрделілік параметрімен басқарылады, яғни. ағаштың қаншалықты күрделі екендігі-неғұрлым күрделі болса, соғұрлым ср білімі артады. Іс жүзінде ср ағаштың өсуін шектеу үшін қолданылады, бұл ағаштың қосымша күрделілігіне (қосымша бөлімдеріне) айыппұл салу арқылы. Бірінші әдіс ерікті ережелерді қамтиды және барлау кезеңінде жұмыс істеу үшін пайдалы болуы мүмкін, бірақ біз оңтайлы мәндерді оңай анықтай алмаймыз (яғни. жаңа деректермен болжамды дәлдікті барынша арттыратын мәндер). CP күрделілік параметрінің көмегімен біз ағаштың қандай өлшемі жаңа деректермен жақсы нәтиже беретінін бағалай аламыз. Егер CP күрделілік параметрі тым аз болып шықса, онда ағаш сигналға емес, шуылға бейімделіп, деректерге қайта өңделеді. Екінші жағынан, егер ср тым үлкен болса, онда ағаш тым кішкентай болып шығады және болжау күші аз болады. Rpart-та әдепкі мән 0,01 құрайды, дегенмен үлкен деректер жиынтығы жағдайында сіз оны тым үлкен деп санайсыз. Алдыңғы мысалда ср 0,005-ке орнатылды, өйткені әдепкі мән бір реттік ағашқа әкелді. Барлау талдауында бірнеше мәндерді сынау жеткілікті. Оңтайлы CP параметрінің анықтамасы орын ауыстыру мен дисперсия арасындағыроманың мысалы болып табылады (бөлімдегі "орын ауыстыру мен дис - Персия Арасындағыромаға келу" жазбасын қараңыз. Осы тараудың басында "k таңдау"). Сіз үшін ең көп қабылданған әдіс - CP параметрінің сәйкес мәнінің шамамен бағасын есептеу кросс-тексеру арқылы жүзеге асырылады (бөлімді қараңыз. 4-тараудың" кросс-тексеру"):
1. Деректерді оқу және тексеру (жалған мәліметтермен бақылау үлгісі) жиынтықтарына бөліңіз.
2. Ағашты жаттығу деректерімен өсіріңіз.

3. Оны дәйекті түрде кесіңіз, әр қадамда cp жазыңыз (жаттығу деректерін пайдалану).
4. Тексеру деректеріндегі ең аз қатеге (жоғалтуға) сәйкес келетін cp белгілеңіз.
5. Деректерді жаттығу және тексеру жиынтықтарына қайта бөлу және ағаш өсіру, бұтақтарды кесу және жазу процесін қайталау
6. Бұл процесті қайта-қайта орындаңыз және әр ағаш үшін ең аз қатені көрсететін cp параметрлерін орташалаңыз.
7. Бастапқы деректерге немесе болашақ деректерге оралып, алынған оңтайлы CP параметріне тоқталып, ағашты өсіріңіз.
Rpart-та cptable аргументін cp мәндерінің кестесін және олармен байланысты кросс-валидация қатесін (R-дегі xerror) құру мақсатында пайдалануға болады, одан кросс - версияның ең төменгі қателігі бар CP мәнін анықтауға болады.

***Үздіксіз шаманы болжау***

Ағашқа негізделген үздіксіз шаманы болжау (яғни регрессия) бірдей логика мен процедураны ұстанады, тек әр Ішкі сегменттегі орташа квадраттық ауытқулармен (квадраттық қателіктермен) өлшенетін гетерогенділік және болжамды өнімділік орташа квадраттық қатенің квадрат түбірімен бағаланады (RMSE)

## Ағаштар қалай қолданылады

Ұйымдардағы MO - deley әзірлеушілерінің ең үлкен кедергілерінің бірі-олар қолданатын әдістерге жататын "қара жәшік" құбылысы, бұл ұйымның басқа элементтерінің қарсылығына негіз береді. Осыған байланысты ағаш үлгісінің екі тартымды аспектісі бар.

Ағаш модельдері қандай айнымалылардың маңызды екендігі және олардың бір - бірімен байланысы туралы түсінік алу үшін деректерді тексерудің көрнекі құралын ұсынады. Ағаштар дикторға дейінгі айнымалылар арасында сызықтық емес байланыстарды түсіре алады.

Ағаш модельдері деректерді терең талдау жобасын жүзеге асыру немесе "сату" үшін қарапайым адамдарға тиімді түрде берілуі мүмкін ережелер жиынтығын ұсынады. Алайда, болжамға келетін болсақ, бірнеше ағаштардың нәтижелерін пайдалану бір ағашты пайдаланудан гөрі тиімдірек болады. Атап айтқанда, кездейсоқ орман мен бустер ағаштарының алгоритмдері

әрқашан дерлік болжамды дәлдік пен тиімділіктен асып түседі (бөлімді қараңыз. "Бэггинг және кездейсоқ орман" және "Бустинг" осы тарауда), бірақ жоғарыда аталған жалғыз ағаш үшін мыжылған артықшылықтар жоғалады.

***Ағаш модельдеріне арналған негізгі идеялар***:

- *Шешім ағаштары жіктеу немесе болжау ережелерінің жиынтығын тудырады.*
- *Ережелер деректердің дәйекті сегменттерге бөлінуіне сәйкес келеді.*
- *Әрбір сегмент немесе бөлу дикторға дейінгі айнымалының белгілі бір мәнімен байланысты және деректерді осы предикатордың мәні бөлу нүктесіндегі мәннен жоғары немесе төмен болатын жазбаларға бөледі.*
- *Әр кезеңде ағаш алгоритмі әр Ішкі сегменттегі нәтиженің гетерогенділігін азайтатын бөлу нүктесін таңдайды.*
- *Әрі қарай бөлу мүмкін болмаған кезде, ағаш толығымен өсірілген болып саналады және әрбір терминал түйінінде немесе жапырақта бір класс бар; осы Ережелер (бөлу) жолымен жүретін жаңа жағдайларға осы класс тағайындалады.*
- *Толық өсірілген ағаш деректерге қайта сәйкес келеді және шудың орнына сигнал алатындай етіп кесілуі керек.*
- *Кездейсоқ ормандар мен бұталы ағаштар сияқты бірнеше ағаш алгоритмдері жақсы болжамды өнімділік береді, бірақ жалғыз ағаштардың ережелерге негізделген коммуникативтік қабілетін жоғалтады*