Naive Bayes. Дискриминантты талдау. Логистикалық регрессия.

Жіктеу

Деректерді талдаушылар көбінесе автоматтандырылған шешімді қажет ететін мәселеге тап болады. Электрондық пошта фишинг әрекеті ме? Клиент басқа жеткізушіге барады ма? Интернет пайдаланушысы жарнаманы баса ма? Бұл сұрақтардың барлығы жіктеу мәселелеріне қатысты. Жіктеу болжаудың ең маңызды түрі болуы мүмкін: оның мақсаты - жазбаның нөл немесе бірлік екенін алдын-ала айту (фишинг / фишинг емес, басу/басу, өту/өту) немесе кейбір жағдайларда бірнеше санаттардың бірі (мысалы, кіріс хабарламаларыңызды сүзу). Gmail "негізгі", "әлеуметтік желілер", "промоакциялар" немесе "форумдар").

Көбінесе бізге қарапайым екілік классификациядан гөрі көп нәрсе қажет: Біз белгілі бір жағдайдың сыныпқа жататындығының болжамды ықтималдығын білгіміз келеді.

Жіктелетін жазбаның екілік санатын тағайындайтын модельге ие болудың орнына, көптеген Алгоритмдер мақсатты сыныпқа жататын ықтималдықтың (бейімділіктің) баллдық бағасын қайтара алады. Шын мәнінде, егер біз логистикалық регрессия туралы айтатын болсақ, R - де стандартты нәтижелер коэффициент логарифмінің шкаласында болады және оны бейімділікке айналдыру керек. Содан кейін бейімділіктің баллдық бағасын шешімге айналдыру үшін жылжымалы кесу шегін пайдалануға болады. Жалпы тәсіл келесідей болады:

1. Мақсатты сынып үшін шекті ықтималдылықты орнатыңыз, одан жоғары біз жазбаны осы сыныпқа тиесілі деп санаймыз.

2. Жазбаның мақсатты сыныпқа тиесілі болу ықтималдығын бағалау (кез-келген модель).

3. Егер бұл ықтималдық шекті ықтималдықтан жоғары болса, онда мақсатты сыныпқа жаңа жазба тағайындаңыз.

Кесу шегі неғұрлым жоғары болса, 1 деп болжанған жазбалар соғұрлым аз болады, яғни мақсатты сыныпқа жатады. Кесу шегі неғұрлым төмен болса, соғұрлым 1 деп жазылған жазбалар көп болады.

Бұл тарауда жіктеу мен бейімділікті бағалаудың бірнеше негізгі әдістері қарастырылған; класс - сификация үшін де, сандық болжау үшін де қолдануға болатын қосымша әдістер келесі тарауда сипатталған.

Екі санаттан артық?

Тапсырмалардың басым көпшілігі екілік жауаппен келеді. Алайда кейбір жіктеу міндеттері екіден көп нәтиженің жауабымен байланысты. Мысалы, клиенттің жазылу шартының мерзімі аяқталғаннан кейін үш нәтиже болуы мүмкін: клиент кетеді немесе "басқа жеткізушіге ауысады" (2 Y=), ай сайынғы шартқа ауыстырылады (1Y=) немесе қолтаңбалар- жаңа ұзақ мерзімді келісім-шарт жасайды (0 y =). Мақсат-болжау Yj = үшін 0,j= 1 немесе 2. Осы тараудағы жіктеу әдістерінің көпшілігін тікелей немесе екіден көп нәтиже беретін жауаптарға қарапайым бейімделумен қолдануға болады. Екіден көп нәтиже болған жағдайда да, тапсырма көбінесе шартты ықтималдықтардың көмегімен екілік тапсырмалар қатарына қайта өңделуі мүмкін. Мысалы,

Шарттың нәтижесін болжау үшін екілік болжаудың екі мәселесін шешуге болады: •

0 Y = немесе 0 Y екенін болжаңыз > ; •

егер 0 Y > берілсе, 1 екенін болжаңыз

Y= немесе 2 Y = .Екінші жағдайда, тапсырманы екі жағдайға бөлудің мағынасы бар: кли - лер басқа жеткізушіге ауыса ма, егер ол өтпесе, ол қандай келісімшартты таңдайды. Модельді сәйкестендіру тұрғысынан көп класты тапсырманы екілік тапсырмалар сериясына айналдыру жиі тиімді. Бұл, әсіресе, бір санат басқаларға қарағанда әлдеқайда кең таралған кезде тән.

Naive Bayes (Аңғал Байес) алгоритмі

Аңғал Байес алгоритмі болжамды айнымалылардың мәндерін байқау ықтималдығын пайдаланады, егер нәтиже болса, болжамды айнымалылардың мәндерінің жиынтығы берілген жағдайда Y= і нәтиже беру ықтималдығын бағалау мақсаты.

Негізгі терминдер

Шартты ықтималдық (шартты бейімділік) басқа оқиға болған жағдайда қандай да бір оқиғаны (айталық, X =i) байқау ықтималдығы (айталық, Y=i); P (X i | Y i) ретінде жазылады .

Артқы ықтималдық (posterior probability) болжамды ақпарат ескерілгеннен кейін нәтиже ықтималдығы (оны есепке алмайтын нәтижелердің априорлық ықтималдығынан айырмашылығы)

Байес классификациясын түсіну үшін" аңғал емес " Байес классификациясын ұсынудан бастауға болады. Жіктелетін әрбір жазба үшін:

1. Бірдей протекторлық профилі бар барлық басқа жазбаларды табыңыз (яғни болжамды айнымалылардың мәндері бірдей).

2. Бұл жазбалардың қай сыныптарға жататынын және қай класс басым екенін анықтаңыз (яғни, мүмкін).

3. Бұл сыныпты жаңа жазбаға тағайындаңыз.

Жоғарыда келтірілген тәсіл болжамды айнымалылардың барлық мәндері бірдей деген мағынада жаңа жіктелетін жазба сияқты көрінетін үлгідегі барлық жазбаларды табуға дейін азаяды..

Неліктен нақты Байес классификациясы практикалық емес?

Болжалды айнымалылар саны айтарлықтай болған кезде, көптеген жіктелетін жазбалар дәл сәйкестіксіз болады. Мұны демографиялық айнымалылар негізінде дауыс беру нәтижелерін болжайтын модель контекстінде түсінуге болады. Тіпті әсерлі үлгіде соңғы сайлауда дауыс берген, алдыңғы сайлауда дауыс бермеген, үш қызы мен бір ұлы бар және ажырасқан АҚШ - тың орташа құлауынан жоғары табысы бар испандық ер адаммен жаңа жазба үшін бірде - бір сәйкестік болмайды. Бұл тек сегіз айнымалы — көптеген жіктеу тапсырмалары үшін өте аз Сан. Бес бірдей жиі санаты бар бір ғана жаңа айнымалыны қосу сәйкестік ықтималдығын 5 есе азайтады

Аңғал шешім

Аңғал Байес шешімінде біз ықтималдылықты есептеуді жіктелетін жазбамен сәйкес келетін жазбалармен шектемейміз. Оның орнына біз бүкіл деректер жиынтығын қолданамыз. Аңғал Байес модификациясы келесі түрге ие:

1. Екілік жауапқа қатысты Yi = (0 i = немесе 1) әрбір болжаушы үшін жеке шартты ықтималдықтарды бағалау (|)=J P XYi; бұл болжаушының мәні біз байқаған кезде жазбада болатындығы туралы Yi = . Бұл ықтималдық мәндердің үлесімен бағаланады Yi жазбаларының арасында j X = жаттығу жиынтығында.

2. Бұл ықтималдықтарды бір-бірімен, содан кейін Y =i - ге жататын жазбалардың үлесіне көбейтіңіз .

3. Барлық сыныптар үшін 1 және 2-қадамдарды қайталаңыз.

4. I сынып үшін 2-қадамда есептелген мәнді алып, оны барлық сыныптар үшін осындай мәндердің қосындысына бөлу арқылы і нәтижесінің ықтималдығын бағалаңыз.

5. Жазбаны осы болжаушы мәндер жиынтығы үшін ең жоғары ықтималдығы бар сыныпқа жатқызыңыз.

Бұл аңғал Байес алгоритмін ықтималдық теңдеуі ретінде де жазуға болады Yi = нәтижесін бақылаңыз алдын-ала дикторлардың мәндерінің жиынтығы болған жағдайда

$X_1, ..., X_p$:

$$P(X_1, X_2, ..., X_p).$$

Мағынасы

$P(X_1, X_2, ..., X_p)$. бұл ықтималдық 0 мен 1 арасында болатындығына және Y - ге тәуелді болмайтындығына кепілдік беретін түзету коэффициенті:

$$P(X_1, X_2, ..., X_p) =$$
$$= P(Y=0)\left(P(X_1|Y=0)P(X_2|Y=0)\cdots P(X_p|Y=0)\right) +$$
$$+P(Y=1)\left(P(X_1|Y=1)P(X_2|Y=1)\cdots P(X_p|Y=1)\right).$$

Неліктен бұл формула "аңғалдық" деп аталады? Шындығында, біз бақыланатын нәтиже жағдайында болжаушылардың векторының нақты шартты ықтималдығы жеке шартты ықтималдықтардың өнімімен жақсы бағаланады деген қарапайым болжамды қабылдадық $P(X_j|Y=i)$. Басқаша айтқанда, бағалауда $P(X_j|Y=i)$ орнына $P(X_1, X_2, ..., X_p|Y=i)$ біз X_k для $k \neq j$. үшін барлық басқа болжамды k x айнымалыларынан тәуелсіз екенін қабылдаймыз .

R-де аңғал Байес моделін бағалау үшін бірнеше бағдарламалық жасақтама пакеттері қолданылуы мүмкін. Бұдан әрі ұсынылған код үзіндісі klaR пакетін пайдаланып модельге сәйкес келеді

```
library(klaR)
naive_model <- NaiveBayes(outcome ~ purpose_ + home_ + emp_len_,
                          data = na.omit(loan_data))
naive_model$table
$purpose_
          var
grouping   credit_card debt_consolidation home_improvement major_purchase
  paid off   0.1857711          0.5523427       0.07153354     0.05541148
  default    0.1517548          0.5777144       0.05956086     0.03708506
          var
grouping      medical      other small_business
  paid off 0.01236169 0.09958506     0.02299447
  default  0.01434993 0.11415111     0.04538382

$home_
          var
grouping    MORTGAGE        OWN       RENT
  paid off 0.4966286 0.08043741 0.4229340
  default  0.4327455 0.08363589 0.4836186

$emp_len_
          var
grouping    > 1 Year   < 1 Year
  paid off 0.9690526 0.03094744
  default  0.9523686 0.04763140
```

На выходе из модели будут условные вероятности $P(X_j | Y = i)$. Модель может использоваться для предсказания исхода новой ссуды:

```
new_loan
        purpose_    home_  emp_len_
1 small_business MORTGAGE  > 1 Year
```

В данном случае модель предсказывает невозврат ссуды:

```
predict(naive_model, new_loan)
$class
[1] default
Levels: paid off default

$posterior
      paid off   default
[1,] 0.3717206 0.6282794
```

Болжам сонымен қатар posterior - дан кейінгі несиені қайтару ықтималдығын қайтарады. Аңғал Байес классификаторы офсеттік бағалауды шығарумен

танымал. Алайда, мақсат жазбаларды сәйкестендіру болып табылатын жерде-бірақ ықтималдығы

1 Y = , ықтималдықтың біржақты бағалауы қажет емес және аңғал Байес классификаторы жақсы нәтижелерге әкеледі

Сандық болжау айнымалылары

Анықтамадан біз Байес классификаторы тек категориялық болжаушылармен жұмыс істейтінін көреміз (мысалы, сөздердің, сөз тіркестерінің, таңбалардың және басқа талданатын параметрлердің болуы немесе болмауы болжау мәселесінің негізінде жатқан спам классификациясы). Аңғал Байес классификаторын қолдану үшін екі тәсілдің бірі сандық болжаушыларға қабылдануы керек:

аралық топтарға бөлу және сандық болжаушыларды категориялық болжаушыларға түрлендіру және одан әрі алдыңғы бөлімнен алгоритмді қолдану;

ықтималдық моделін қолданыңыз-мысалы, қалыпты үлестіру (бөлімді қараңыз. 2 — тараудың" қалыпты таралуы") - шартты ықтималдылықты бағалау үшін $P(X_j \mid Y = i)$.

Аңғал Байес алгоритмінің негізгі идеялары

- *Аңғал Байес алгоритмі категориялық (факторлық) болжаушылармен және нәтижелермен жұмыс істейді.*
- *Ол сұраққа жауап береді: нәтижелердің әр санатында қандай болжау категориялары болуы мүмкін?*
- *Бұл ақпарат болжаушылардың мәндерін ескере отырып, нәтижелер санаттарының ықтималдығын бағалау үшін одан әрі аударылады.*

Дискриминантты талдау

Дискриминантты талдау-бұл ең алғашқы статистикалық классификатор; оны р. а. Фишер 1936 жылы "евгеника шежіресі" журналында жарияланған мақалада ұсынған (евгеника жылнамалары)

Негізгі терминдер

Коварианс (covariance)

Айнымалының басқасымен бірге өзгеру дәрежесін көрсететін метрикалық көрсеткіш (яғни оның шамасы мен бағыты ұқсас).

Дискриминантты функция (discriminant function)

Болжалды айнымалыларға қолданылған кезде сыныптардың бөлінуін барынша арттыратын Функция.

Дискриминантты салмақтар (discriminant weights)

Белгілі бір сыныпқа жату ықтималдығын есептеу үшін қолданылатын кемсітушілік функцияны қолдану нәтижесінде алынған бағалар.

Дискриминантты талдау бірнеше мамандандырылған әдістерді қамтығанымен, сызықтық дискриминантты талдау (linear discriminant analysis, LDA) кең таралған. Фишер ұсынған бастапқы әдіс LDA-дан сәл өзгеше болды, бірақ механизм негізінен өзгеріссіз қалды. Бүгінгі күні Lda ағаш модельдері және логистикалық регрессия сияқты күрделі мамандандырылған әдістердің пайда болуымен аз қолданылады. Алайда, LDA-ны кейбір қосымшаларда әлі де кездестіруге болады және басқа кеңінен қолданылатын әдістермен байланысы бар (мысалы, негізгі компоненттерді талдау; бөлімді қараңыз. "7-тараудың" негізгі компоненттерін талдау"). Сонымен қатар, дискриминантты талдау болжаушының маңыздылығын қамтамасыз ете алады және белгілерді таңдаудың есептеу тиімді әдісі ретінде қолданылады

Ковариациялық матрица

Дискриминантты талдауды түсіну үшін алдымен екі немесе бірнеше айнымалылар арасындағы ковариация ұғымын енгізу қажет. Ковариация x және z Екі айнымалысы арасындағы байланысты өлшейді. x және z (бөлімді қараңыз. 1-тараудың" орташа"). Ковариация ., X және z арасындағы xss келесідей орнатылады формула

$$s_{x,z} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(z_i - \bar{z})}{n-1},$$

мұндағы n - жазбалардың саны (n орнына 1 n − ге бөлетінімізді ескеріңіз: "еркіндік дәрежесі және n немесе n - 1" бөлімін қараңыз?"1 тараулар).

Корреляция коэффициенті сияқты (бөлімді қараңыз. 1-тараудың" корреляциясы"), оң мәндер оң байланыс туралы, ал теріс мәндер кері байланыс туралы айтады. Алайда Корреляция 1 − ден 1 - ге дейінгі мәндермен шектеледі, ал ковариация x және Z айнымалыларымен бірдей өлшеу шкаласында болады. X және z үшін Σ Ковариация матрицасы диагональдағы 2 X s және 2 z s жеке айнымалыларының дисперсияларынан тұрады (мұнда жол мен баған — бұл бірдей айнымалы) және диагональдардан тыс тұрған айнымалы жұптар арасындағы ковариа - ций.

$$\hat{\Sigma} = \begin{bmatrix} s_x^2 & s_{x,z} \\ s_{x,z} & s_z^2 \end{bmatrix}.$$

Фишердің сызықтық дискриминанты

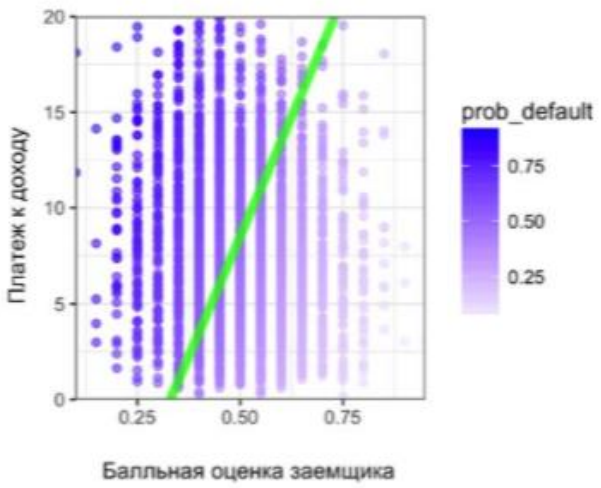
Қарапайымдылық үшін біз алдын - ала айтқымыз келетін жіктеу мәселесіне назар аударамыз екілік нәтиже y тек екі үздіксіз сандық белдік (x, z) көмегімен. Техникалық тұрғыдан алғанда, дискриминантты талдау болжамды айнымалылар қалыпты үлестірілген емес шамалар болып табылады деп болжайды, бірақ іс жүзінде бұл әдіс қалыпты жағдайдан, сондай-ақ екілік болжаушылардан шексіз ауытқуларда да жақсы жұмыс істейді. Фишердің сызықтық дискриминанты бір жағынан топтар арасындағы вариацияны, екінші жағынан топтар ішіндегі вариацияны ажыратады. Атап айтқанда, жазбаларды екі топқа бөлуге тырысып, LDA SS (топ ішіндегі вариацияны өлшейтін) ішіндегі квадраттардың "топ ішіндегі" қосындысына қатысты SS (осы екі топ арасындағы вариацияны өлшеу) арасындағы квадраттардың қосындысын "арасында" максимизациялауға бағытталған. Бұл жағдайда бұл екі топ y = 0болатын жазбаларға (x0 , z0) және y =1 болатын жазбаларға (11, xz) сәйкес келеді . Бұл әдіс сызықтық комбинацияны табуға мүмкіндік береді

- квадрат $w_x x + w_z z$, бұл квадраттар қосындысының қатынасын барынша арттырады:

$$\frac{SS_{\text{между}}}{SS_{\text{внутри}}}.$$

Квадраттардың топ аралық қосындысы-екі топтық орташа мәннің квадраттық арақашықтығы, ал квадраттардың топ ішіндегі қосындысы — ковариациялық

матрицаға өлшенген әр топтың ішіндегі орташа мәннің таралуы. Интуитивті түрде квадраттардың топаралық қосындысын көбейту және квадраттардың топ ішіндегі қосындысын азайту арқылы бұл әдіс осы екі топ арасындағы ең үлкен бөлуді береді деп болжауға болады



5.1. -сурет. Екі айнымалыны қолдана отырып, несиені қайтармау туралы LDA-ға негізделген болжам-қарыз алушының несиелік ұпайы және төлемдер мен кірістердің арақатынасы

Дискриминантты функцияның салмағын қолдана отырып, LDA болжау кеңістігін қалың сызықпен көрсетілгендей екі аймаққа бөледі. Сызықтан әлдеқайда ұзағырақ орналасқан болжамдардың сенімділік деңгейі жоғары (яғни, ықтималдығы 0,5-тен әлдеқайда жоғары).

Дискриминантты талдаудың негізгі идеялары

- *Дискриминантты талдау үздіксіз немесе категориялық болжаушылармен, сондай-ақ категориялық нәтижелермен жұмыс істейді.*
- *Ковариациялық матрицаны қолданған кезде ол әр түрлі кластарға сәйкес келетін жазбаларды ажырату үшін қолданылатын сызықтық дискриминантты функцияны есептейді.*
- *Бұл функция Раяның бағалау сыныбын анықтайтын әрбір жазба үшін (әрбір мүмкін сынып үшін бір салмақ) таразыны немесе баллдық ұпайларды алу үшін жазбаларға қолданылады.*

Логистикалық регрессия

Логистикалық регрессия бір ерекшелікті қоспағанда, бірнеше сызықтық регрессияға ұқсас — нәтиже екілік. Бұл есепті сызықтық модельді орнатуға болатын түрге келтіру үшін әртүрлі түрлендірулерді қолданады. Дискриминантты талдау сияқты және жақын көршілердің K әдісінен және аңғал Байес алгоритмінен айырмашылығы, логистикалық регрессия - бұл ақпараттық-центрлік тәсілге қарағанда құрылымдық модельді тәсіл. Бұл әдіс өзінің жоғары есептеу жылдамдығының және жаңа деректерді жедел бағалауға мүмкіндік беретін модельдің шығуының арқасында танымал болды.

Негізгі терминдер

Логит-түрлендіру (logit) сынып қа жату ықтималдығын көрсететін Функция (0-1 диапазонында) ± ∞ диапазонына . Синонимдер: коэффициенттердің логарифмі (төменде қараңыз), логит.

Коэффициенттер (odds) "табысқа" (1) "сәтсіздікке" (0) қатынасы.

Коэффициент логарифмі (log odds) түрлендірілген модельдегі жауап (қазір сызықтық), ол артқа қарай көрсетіледі ықтималдық.

Екілік нәтиже айнымалысынан сызықтық модельдеуге болатын, содан кейін екілік нәтижеге қайтарылатын нәтиже айнымалысына қалай ауысуға болады?

Логистикалық жауап функциясы және логиттік түрлендіру

Негізгі компоненттер - бұл логистикалық жауап функциясы және логиттік түрлендіру, бұл ықтималдылықты (көрсетілген) көрсетуге мүмкіндік береді шка - ле 0-1) сызықтық модельдеуге сәйкес келетін кеңейтілген шкалаға. Бірінші қадам - нәтиже айнымалысын екілік белгі ретінде емес, белгісі "1" болатын p ықтималдығы ретінде көрсету. Біз болжамды айнымалылардың сызықтық функциясы ретінде p модельдеуді қалауымыз мүмкін:

$$p = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q.$$

Алайда, бұл модельге сәйкес келу p 0 мен 1 арасында болатынына кепілдік бермейді, бұл ықтималдық болуы керек. Оның орнына біз болжаушыларға логистикалық жауап немесе кері логит түрлендіру функциясын қолдану арқылы p модельдейміз:

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)}}.$$

Бұл түрлендіру p 0 мен 1 арасында қалуын қамтамасыз етеді. Бөлгіштен экспоненциалды өрнек алу үшін сенімнің орнына біз коэффициенттерді немесе артықшылықтарды қарастырамыз. Барлық ойыншыларға таныс коэффициенттер-бұл "табыстың" (1) "сәтсіздікке" (0) қатынасы. Ықтималдықтар тұрғысынан коэффициенттер-бұл оқиғаның орын алмау ықтималдығына бөлінген оқиғаның ықтималдығы. Мысалы, егер сіз атқа ие болу ықтималдығы 0,5 болса, онда "жеңе алмау" ықтималдығы 1 0,5 0,5 − = болады, ал коэффициенттер 1,0 болады.

$$\text{Шансы } (Y = 1) = \frac{p}{1 - p}.$$

Мүмкіндіктердің ықтималдығын кері функция арқылы алуға болады

$$p = \frac{\text{Шансы}}{1 + \text{Шансы}}.$$

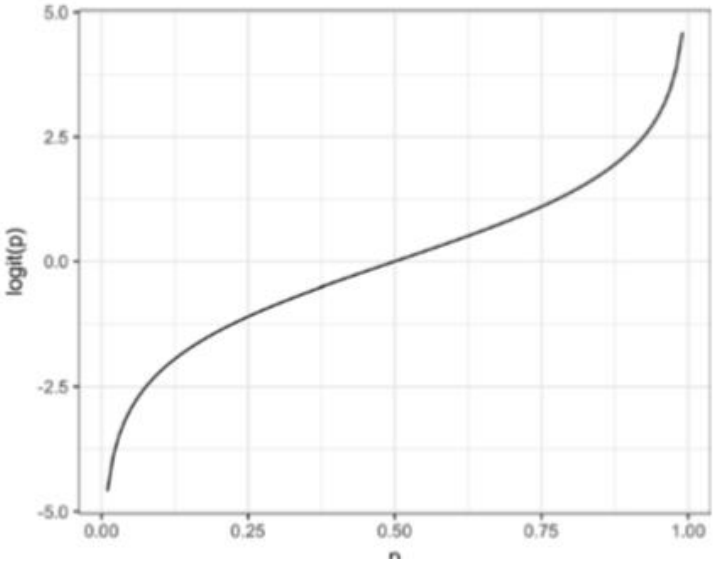
Біз бұл формуланы бұрын көрсетілген логистикалық жауап функциясымен біріктіреміз және аламыз:

$$\text{Шансы } (Y = 1) = e^{\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q}.$$

Соңында, тең белгінің оң және сол жағында орналасқан өрнектердің логарифмін алып, болжаушылардың сызықтық функциясын қамтитын өрнекті аламыз:

$$\log\left(\text{Шансы } (Y = 1)\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q.$$

Коэффициент логарифмінің функциясы, логит функциясы деп аталады, p ықтималдығын (0, 1) интервалдан кез - келген мәнге (,−∞+∞) көрсетеді, сурет. 5.2. Түрлендіру циклі аяқталды; біз ықтималдылықты болжау үшін сызықтық модельді қолдандық, оны өз кезегінде кесу ережесін қолдану арқылы сынып белгісіне көрсетуге болады — кесу мүйізінен үлкен ықтималдығы бар кез келген жазба 1 ретінде жіктеледі.



Сурет. 5.2. Сызықтық модельге сәйкес келетін мектепке ықтималдықты көрсететін Функция (логит)