

Дәріс 3. Ақпарат көздерін талдау үшін ықтималдықты пайдалану. Байес формуласының ақпарат теориясындағы орны.

Дәріс жоспары.

1. Ақпарат көздерін ықтималдық тұрғысынан сипаттау
2. Ықтималдық ұғымы және оның ақпарат теориясындағы рөлі
3. Ақпарат мөлшерін анықтау және энтропия ұғымы
4. Байес формуласы және оның шығару тәсілі
5. Байес формуласын ақпарат көздерін талдауда қолдану

Ақпарат көздерін ықтималдық тұрғысынан сипаттау

Ақпарат көзі дегеніміз — белгілі бір ықтималдықпен символдар немесе хабарлар тізбегін шығаратын жүйе.

Мысалы:

Бір хабарлама x_i ықтималдығы $P(x_i)$ болсын.

Барлық ықтимал хабарламалар жиыны $X=\{x_1, x_2, \dots, x_n\}$

Сонда:

$$\sum_{i=1}^n P(x_i) = 1 \quad \sum_{i=1}^n P(x_i) = 1$$

Ақпарат теориясы ақпарат көзінің статистикалық сипаттамасын талдауда осы ықтималдық үлестірімін негізгі құрал ретінде қолданады.

Ықтималдық ұғымы және оның ақпарат теориясындағы рөлі

Ақпараттың мөлшері хабардың пайда болу ықтималдығына кері пропорционал: сирек хабар көбірек ақпарат береді. Шеннон енгізген ақпараттың бірлік өлшемі:

$$I(x_i) = -\log_2 P(x_i) \quad I(x_i) = -\log_2 P(x_i)$$

Мұнда:

$I(x_i)$ - хабардағы ақпарат мөлшері (битпен өлшенеді);

$P(x_i)$ - хабардың ықтималдығы.

Мысалы:

Егер хабардың ықтималдығы $P(x_i) = \frac{1}{2}$ болса, онда $I(x_i) = 1$ бит.

Ақпарат мөлшерін анықтау және энтропия ұғымы

Энтропия — ақпарат көзінің белгісіздік деңгейін сипаттайды.

Ақпарат теориясы — математиканың ақпаратты сақтау, түрлендіру және тасымалдау әдістерін зерттейтін бөлімі, кибернетиканың маңызды саласы. Ақпарат теориясының негізгі ұғымы ақпарат (хабар) болып есептеледі. Ақпарат теориясында берілетін не сақталатын хабарлар көзі бар деп ұйғарылады. Бұл хабарлардың шығуы мен қабылдануы кездейсоқ сипатта болғанымен, оның шығу ықтималдығы белгілі деп қарастырылады. Мұндай хабарлардың “анықталмау” дәрежесін сипаттау үшін Ақпарат теориясына энтропия деп аталатын ұғым ендіріледі.

Энтропия — қандай да бір тәжірибенің (сынақтың) әр түрлі нәтижемен аяқталатын анықталмағандық өлшемі. Ақпарат теориясы хабардың байланыс

арнасы арқылы бөгеуілдермен берілу тәсілдерін зерттейді. Әр түрлі ақпаратты жинау, оны түрлендіру, жеткізу және сақтау — кез келген басқару жүйесінің қызметінде кездесетін жағдай. Жүйенің әр түрлі бөліктерінің арасында ақпарат алмасу әрқашан болып тұрады. Кез келген ақпаратты бір жерден екінші жерге жеткізу үшін оның кодталынуы (таңбалануы), яғни арнаулы белгілерге (символдарға) және сигналдарға түрленуі керек. Сондықтан да Ақпарат теориясының негізгі мәселесі аз ғана белгілер арқылы көп мәнді ақпарат беруді тиімді түрде кодтау (таңбалау) тәсілін зерттеу болып табылады. Бұл — байланыс арнасында бөгеуіл болған не болмаған жағдайда шешілетін мәселе. Байланыс арнасы арқылы қабылдаушы буынға келіп жеткен таңбаланған ақпаратты алғашқы түріне келтіруді декодтау (таңбадан мағынаға көшіру) деп атайды. Ақпарат теориясындағы күрделі мәселенің бірі — ақпарат көзінде үздіксіз өндірілген ақпаратты байланыс арнасы арқылы басқа орынға дер кезінде және бөгеуілсіз жіберіп тұру үшін байланыс арнасының өткізгіштік сыйымдылығын анықтау. Бұл мәселені шешу ақпарат мөлшерін бағалай білуді қажет етеді. Ақпарат мөлшерін бағалау ықтималдық теориясының заңдарына негізделген. Ақпарат теориясының негізін 1948 — 1949 ж. американ ғалымы К.Э.Шеннон қалаған. Оның теориясының дамуына Ресей ғалымдары А.Н.Колмогоров (1903 — 1987), А.Я.Хинчин (1894 — 1959), ал іс жүзінде қолданылатын саласына В.А.Котельников т.б. үлкен үлес қосты.

Егер барлық хабарлар тең ықтимал болса, энтропия ең жоғары мәнге ие болады.

Егер хабардың бірі 100% ықтималдықпен пайда болса, энтропия нөлге тең.

Байес формуласы және оның шығару тәсілі

Байес формуласы шартты ықтималдық арасындағы байланысты сипаттайды:

$$P(A|B)=P(B|A) \cdot P(A)P(B)P(A|B)$$

Мұндағы:

$P(A|B)P(A|B)P(A|B)$ - оқиға В орындалғанда -ның ықтималдығы (апостериорлық ықтималдық);

$P(B|A)P(B|A)P(B|A)$ - оқиға А болғанда В-ның ықтималдығы (априорлық шартты ықтималдық);

$P(A)P(A)P(A)$ - А-ның априорлық ықтималдығы;

$P(B)P(B)P(B)$ - барлық мүмкін жағдайлар үшін ВВВ-ның ықтималдығы.

Бұл формула жаңа ақпарат пайда болғанда ықтималдықты жаңарту үшін қолданылады.

Байес формуласын ақпарат көздерін талдауда қолдану

Ақпарат көздерінен алынған мәліметтер көбінесе сенімсіз немесе бұрмаланған болады.

Байес тәсілі жаңа ақпаратты пайдалана отырып, бастапқы (априорлық) білімімізді нақтылауға мүмкіндік береді.

Қолдану мысалдары:

Шуды бар байланыс арнасында дұрыс хабарды қалпына келтіру;
Ақпараттық сүзгілер мен классификаторларда (мысалы, "спам/спам емес" хаттарды анықтау);

Сенімділік пен ықтималдық модельдері бар криптожүйелерде шешім қабылдау.

Информацияның өлшем бірлігі

Аз информациясы бар хабарды көп информациясы бар айы- ратын критерийлерді қалай анықтауға болады?

Екі хабарды салыстырайық: A – күн шығады және B – ертең күн тұтылады. Бірінші (A) оқиғаның шындығында ешкімнің кү- мәні жоқ, ал екінші хабар (B) қызығушылық тудырады (баспа- сөзде кең қарастырады, белгілі ғалымдардан сұхбат алынады).

Осыдан мынадай нәтиже шығаруға болады: аз априорлы (тәжі- рибе алдында) ықтималдылығы бар оқиғалар көп информация әкеледі. Яғни информация I және оқиғаның ықтималдылығы P мынадай қатынаспен байланысқан:

Белгісіз f функцияның түрін анықтау үшін жиынның өлшемі ретінде алынған информация мөлшердің аддитивтік қасиеті бар деп есептейміз, яғни екі когеренті 1 және 2 көзден түскен жалпы информация бөлек үлестің қосындысына тең болуы керек $I = I_1 + I_2$. Осы шарт орындалу үшін f функция келесі қасиетке ие болуы керек:

мұндағы P_1 және P_2 – екі тәуелсіз оқиғаның орындалу ықтимал- дылықтары. Осындай текті қасиеті бар жалғыз математикалық функция – ол логарифмдік функция, сондықтан

Біздің жағдайымыз үшін A оқиғасымен салыстырғанда B оқиғасында жатқан информация мөлшері $I \square B/A \square$, (1.2)-ге негіз- деліп ол келесі түрде жазылады:

мұндағы $P(A)$, $P(A, B)$ – сәйкесінше A оқиғаның және B мен A оқиғалардың бірге орындалу ықтималдылығы. $B = A$ оқиғаның пайда болуын A оқиға пайда болды деп мағыналасуға болады.

$I \square A/ A \square \square I \square A \square$ саны A оқиғасында жатқан $I \square A \square$ информацияның мөлшерін анықтайды:

(1.3) формуладағы логарифмнің негізіне (r) байланысты информация мөлшері *бит* (логарифмнің негізі $r = 2$), *дит* (негізі $r = 10$), *нат*-пен (Эйлер саны негізі болатын натурал логарифм жағдайында $r = e$) өлшенеді.

Көбінесе ықтималдылығы бірдей екі оқиғаның біреуі орын- далғаны туралы мәліметтегі мөлшерді информация мөлшерінің бірлігі ретінде алады. Мұндай информация мөлшерінің бірлігі *бит* немесе *екілік бірлігі* деп аталады. Мысалы, симметриялы тиынды лақтырған нәтижесінде алынған информация мынаған тең:

Бит – ол өте аз бірлік, сондықтан көбінесе одан 8 есе көп шама қолданады: *байт* (byte). Байтты үлкен Б әріппен белгілей- ді. Басқа стандартты бірліктеріндегідей бит пен байт бірліктері- нің туындылары бар: кило (К), мега (М), гига (Г), тера (Т), пета (П) және т.б. Бірақ бит пен байттар үшін олар 10- ның дәрежесін емес, 2-нің дәрежесін білдіреді:

кило – $2^{10} = 1024 \square 10^3$,

мега – $2^{20} \square 10^6$, гига – $2^{30} \square 10^9$, тера – $2^{40} \square 10^{12}$, пета – $2^{50} \square 10^{15}$.

Мысалы, 1КБ = 8 Кбит = 1024 Б = 8192 бит.

Белгі немесе сигналдардың реалды әлемнің объектілерімен бір мағыналық байланыстығына негізделген информация *семантикалық* немесе *мағыналық* деп аталады. Мәліметтегі белгілер (сигналдар) реті және олардың байланысында жататын информация *синтактикалық* деп аталады. Семантикалық информацияны алу тәжірибелік маңызы жоғары екендігі айқын. Бірақ се-

мантикалық информацияның эффективті өлшемі жоқ. Ал синтактикалық информацияны өлшеуге болады, оның тәжірибелік құндылығы семантикалық информация (алушыға керек) берілген сигналдар тізбегінде жатқанында. Белгілі уақыт аралығында көп белгі берілсе, онда орташа (барлық мәліметтің ансамблі бойынша) көп мағыналық информация беріледі. Белгілер жалпы ғылымында (*семиотикада*) информацияның семантикалық және синтактикалық аспектілерінен басқа синтактикалық және прагматикалық аспектілері зерттеледі. Бірінші жағдайда объектілерді белгілеуге арналған белгіні таңдау (беру) сұрақтары, екіншіде информацияның тәжірибе жүзінде қолдануы, қойылған мақсатқа жетуге оның құндылығы қарастырылады. Қазіргі уақытта прагматикалық информацияның өлшем теориясы инженерлік қолдану деңгейіне дейін жеткізілмеген.

Информацияның жалпы келесі түрлерін айтуға болады (толық келесі тарауда жазылған): *дискретті (сандық)* және *үздіксіз (аналогты)*. Дискретті информация қандай да бір шаманың тізбектік дәл мәнімен, ал үздіксіз информация үздіксіз үдерісімен сипатталады. Үздіксіз информацияны, мысалы, атмосфералық қысымның көрсеткіші немесе автомобильдің жылдамдық көрсеткіші көрсетеді. Дискретті информацияны кез келген сандық көрсеткіштен алуға болады: электронды сағат, магнитофонның санағышы және т.б.

Дискретті информацияны өңдеу өте ыңғайлы, бірақ үздіксіз информация тәжірибеде жиі кездеседі, сондықтан үздіксіз информацияны дискреттіге (*дискреттеу үдерісі*) және керісінше айналдыра білу қажет. Дискреттеу үдерісі толық екінші тарауда қарастырылған. Модем (бұл сөз модуляция және демодуляция сөздерінен тұрады) осындай аударуға арналған құрал: ол компьютерден алынған сандық берілгендерді дыбысқа немесе электромагнитті тербелістерге – дыбыстың көшірмесіне аударады және керісінше.

Үздіксіз информацияны дискретті түріне аударғанда үздіксіз шаманың мәнін анықтау периодын ($T = 1/\square$) анықтайтын

$\square$ *дискреттеу жиілігі* маңызды (1.1-сурет).

Дискреттеу жиілігі жоғары болған сайын үздіксіз информацияны дискреттіге аудару дәлірегі артады. Бірақ бұл жиілік өскен сайын осындай аудару кезінде алынған дискретті берілгендердің де өлшемі және де оларды өңдеу, жіберу, сақтау қиыншылығы өседі.





1.1-сурет

Дискреттеу дәлдігін арттыру үшін оның жиілігін шексіз өсіру міндетті емес. Бұл жиілікті Найквист (Nyquist) теоремасымен анықталатын шекке дейін өсіру дұрыс.

Кез келген үздіксіз шама бірнеше, бір-біріне беттестірілген, гармониктер деп аталатын және келесі түрдегі функциялармен анықталатын $A \sin t$, мұндағы A – амплитуда, ω – жиілік, t – уақыт, ϕ – фаза, толқындық үдерістермен сипатталады.

Найквист теоремасы. Дәл дискреттеу үшін сигналдың жиілігі дискреттелетін шамаға кіретін гармониктің ең жоғары жиілігінің екі есесінен көп болуы керек.

Бұл теореманы қолданудың мысалы ретінде лазерлі компакт-дискілерді қарастыруға болады, оларда дыбыс информациясы сандық түрде сақталады. Дискреттеу жиілігі артқан сайын дыбыс дәлірек болады және де бір дискіге мұндай информация аз жазылады. Бірақ жай адамның құлағы жиілігі 20 КГц-ке дейін дыбысты айырады, сондықтан одан көп жиілікті дыбысты жазу мәнсіз. Найквист теоремасы бойынша дискреттеу жиілікті 40 КГц-тен аз алмау керек. Дискретті информацияны үздіксіз

түрге аударғанда анықтаушы ретінде осы аударудың жылдамдығы болады: жылдамдық артқан сайын үздіксіз шама жоғары жиілікті болады. Бірақ осы шамада жоғары жиіліктер көп болған сайын онымен жұмыс істеу қиындайды.

Үздіксіз информацияны дискретті түрге аударуға арналған құралдардың жалпы қысқаша аталуы – АСТ (аналогты-сандық түрлендіргіш, АЦП – аналого-цифровой преобразователь, ADC – Analog to Digital Converter, A/D), ал дискретті сигналды үздіксізге аударушы құрал – САТ (сандық-аналогты түрлендіргіш, ЦАП – цифро-аналоговый преобразователь, DAC – Analog to Digital Converter, D/A).

Информацияны беру радиотехникалық жүйесі (ИБ РТЖ) информацияны ғана жіберуге арналған, объект, энергия немесе сигналды емес. Осындай жүйелерде электромагнитті толқындар (сигналдар) көзден тұтынушыға *информацияның тасымалдау-шысы* ғана ретінде қолданылады. Бұл – радиосигналдарды жіберу және қабылдауға арналған басқа жүйелерден басты айыр-машылығы.

Информацияны жіберу радиотехникалық жүйелерге немесе басқаша айтылатын *байланысты типті жүйелерге* радиобайланыс жүйелер (соның ішінде ұялы және радиотелефондық байланыс), радиохабар мен теледидар жүйелер (соның ішінде жерсеріктік байланыс), радиотелеметриялық жүйелер (радиоканалмен өлшеулердің нәтижесін жіберуге арналған) және т.б. жүйелер жатады. *Өлшеуші типті жүйелер* – радиолокациялық және радионавигациялық жүйелер, траекториялық өлшеу жүйелер, қоршаған орта параметрлерін өлшеуге арналған жүйелер және т.б. ИЖ РТЖ-нен айырмашылығы: қоршаған орта және объектімен әсерлесу кезінде пайдалы информация сигналға беттестіріледі (немесе сигналда пайда болады) және осы объектілер мен сигналдың таралу ортасының параметрлері мен қасиеттерін бейнелейді. Қабылдау орнында пайдалы информация сигналдан электромагнитті өрістің сәйкесінше параметрлерін өлшеу арқылы алынады.

6. Мысал

Бір дерек көзі екі хабар жібереді:

x_1 : «0» – ықтималдығы 0.7

x_2 x_2 : «1» – ықтималдығы 0.3

Байланыс арнасында қателік ықтималдығы $P(\text{қате})=0.1$ $P(\text{қате}) = 0.1P(\text{қате})=0.1$.

Хабар ретінде «1» қабылданса, оның шын мәнінде «1» болған ықтималдығы:

$$P(x_2|1 \text{ қабылданды}) = \frac{P(1 \text{ қабылданды}|x_2)P(x_2)}{P(1 \text{ қабылданды}|x_1)P(x_1) + P(1 \text{ қабылданды}|x_2)P(x_2)}$$

Мұнда:

$$P(1 \text{ қабылданды}) = P(1 \text{ қабылданды}|x_1)P(x_1) + P(1 \text{ қабылданды}|x_2)P(x_2)$$

Есептеу нәтижесінде Байес тәсілі қабылданған сигналдың шын мәнін ықтималдық тұрғысынан бағалауға мүмкіндік береді.

Қорытынды

Ықтималдық теориясы ақпарат көздерін талдау мен сипаттауда негізгі рөл атқарады.

Энтропия — ақпарат көзі шығаратын хабарлардың белгісіздігін өлшейтін негізгі көрсеткіш.

Байес формуласы жаңа ақпарат түскен сайын ықтималдықты жаңарту мен шешім қабылдауда кең қолданылады.

Ақпарат теориясы мен ықтималдық әдістері байланыс жүйелерін, криптожүйелерді және деректерді талдау әдістерін дамытуға негіз болады.

Пайдаланылған әдебиеттер тізімі

1. Асанов С. *Физикалық энциклопедия*. – Алматы: Қазақ энциклопедиясы, 2003.
2. Шеннон К. *Математическая теория связи*. – М.: Наука, 1963.
3. Клаузиус Р. *О движущей силе тепла и законах, пригодных для этого движения*. – Берлин, 1865.
4. Больцман Л. *Избранные труды по статистической физике*. – М.: Наука, 1984.