14-ДӘРІС
ЭНТРОПИЯ ЖӘНЕ ТІЛДІҢ АРТЫҚТЫҒЫ. БІРЛІК АРАЛЫҚ

Дәрістің мақсаты: Ақпараттық энтропия ұғымын криптографиялық тұрғыдан түсіндіру, символ ықтималдығы мен ақпарат мөлшері арасындағы байланысты көрсету, тілдің артықтығын анықтау және бірлік аралықтың (unicity distance) мағынасын теориялық деңгейде түсіндіру.

Ақпарат мөлшері

Ақпарат теориясында оқиғаның ақпараттық мөлшері оның ықтималдығымен байланысты. Жиі болатын, алдын ала оңай болжанатын оқиға аз ақпарат береді, ал сирек кездесетін оқиға көбірек ақпарат береді. Шеннондық өлшем бойынша x оқиғасының ақпараттық мөлшері:

$$I(x) = -\log_2 p(x)$$

Мұндағы p(x) – оқиғаның ықтималдығы. Логарифм негізі 2 болғандықтан ақпарат мөлшері битпен өлшенеді. Егер оқиға міндетті түрде орындалса, p(x)=1 және I(x)=0 бит. Яғни толық белгілі оқиға жаңа ақпарат бермейді.

Мысалы, p(x)=1/2 болса:

$$I(x) = -\log_2(1/2) = 1 \text{ бит}$$

Ал p(x)=1/8 болса:

$$I(x) = -\log_2(1/8) = 3 \text{ бит}$$

Демек, ықтималдығы сегізден бір болатын оқиғаны бақылау ықтималдығы екіден бір болатын оқиғаға қарағанда көбірек ақпарат береді. Бұл қасиет ақпарат мөлшерінің оқиғаның сиректігін сандық түрде сипаттайтынын көрсетеді.

Энтропия ұғымы

Кездейсоқ X шамасының энтропиясы – оның барлық мүмкін мәндері бойынша орташа ақпарат мөлшері. Басқаша айтқанда, энтропия ақпарат көзінің белгісіздік деңгейін сипаттайды.

$$H(X) = -\Sigma\, p(x_i) \log_2 p(x_i)$$

Мұндағы $p(x_i)$ – X кездейсоқ шамасының x_i мәнін қабылдау ықтималдығы. Егер логарифм негізі 2 болса, энтропия битпен өлшенеді.

Энтропия жоғары болған сайын келесі символды немесе оқиғаны болжау қиынырақ. Энтропия төмен болған сайын кейбір мәндер басқаларға қарағанда жиі кездеседі және көздің мінез-құлқы болжамдырақ болады.

Энтропия жеке оқиғаның ақпарат мөлшері емес. I(x) нақты бір оқиға қанша ақпарат бергенін сипаттаса, H(X) – барлық мүмкін оқиғалар бойынша күтілетін орташа ақпарат мөлшері.

Тең ықтималды символдар

Егер алфавитте m символ болса және олардың барлығы бірдей ықтималдықпен пайда болса, әр символ үшін p=1/m болады. Бұл жағдайда белгісіздік ең жоғары және энтропия максимал мәнге жетеді:

$$H_max = \log_2 m$$

Мысалы, төрт символдан тұратын {A, B, C, D} алфавитіндегі барлық символдардың ықтималдығы 1/4 болсын. Онда:

$$H(X) = -4 \cdot (1/4) \cdot \log_2(1/4) = 2 \text{ бит/символ}$$

Бұл нәтиже $H_max=\log_2 4=2$ бит/символ формуласымен сәйкес келеді. Егер символдардың ықтималдықтары тең болмаса, нақты энтропия осы максимал мәннен төмен болады.

Энтропияны есептеу мысалы

X кездейсоқ шамасы екі мән қабылдасын: $p(0)=0,75$ және $p(1)=0,25$. Энтропия формуласына осы ықтималдықтарды қоямыз:

$$H(X) = -0,75\log_2 0,75 - 0,25\log_2 0,25$$

Жеке мүшелерді есептегенде:

$$-0,75\log_2 0,75 \approx 0,311$$
$$-0,25\log_2 0,25 = 0,500$$

Сонда:

$$H(X) \approx 0,311 + 0,500 = 0,811 \text{ бит}$$

Екі ғана мүмкін мән болғандықтан максимал энтропия 1 бит болар еді. Бірақ 0 символы 1 символына қарағанда үш есе жиі кездеседі. Сондықтан келесі символды болжау жеңілдейді және нақты энтропия 1 биттен төмен болады.

Тілдің артықтығы

Табиғи тілдегі символдар бірдей ықтимал емес. Кейбір әріптер жиі, кейбіреулері сирек кездеседі. Сонымен бірге әріптер мен сөздер бір-бірінен тәуелсіз пайда болмайды: олардың арасында грамматикалық, лексикалық және контекстік байланыстар бар.

Мысалы, сөздің алғашқы бірнеше әрпі белгілі болған кезде келесі әріпті немесе сөзді белгілі дәрежеде болжауға болады. Осындай болжанатын құрылым тілдің артықтығын қалыптастырады.

Артықтықтың пайдалы жағы да бар: хабарламадағы кейбір символдар жоғалса немесе қате берілсе, адам контекст арқылы бастапқы мағынаны кейде қалпына келтіре алады. Бірақ криптография тұрғысынан бұл болжамдылық классикалық шифрларды талдауда статистикалық заңдылықтардың көзі болуы мүмкін.

Сондықтан классикалық криптоталдауда символдар жиілігі, қайталанатын тіркестер және тіл құрылымы маңызды болған. Қазіргі криптографиялық алгоритмдер ашық мәтіннің осындай статистикалық қасиеттерін шифрмәтінде байқатпауға бағытталады.

Абсолюттік және салыстырмалы артықтық

Алфавиттің барлық символдары тең ықтималды болған жағдайда H_max максимал энтропия алынады. Нақты ақпарат көзінің энтропиясы H болса, олардың айырмасы бір символға келетін абсолюттік артықтықты сипаттай алады:

$$D = H_max - H$$

Мұндағы D – абсолюттік артықтық. Ол бит/символ бірлігімен берілуі мүмкін.

Салыстырмалы артықтық максимал энтропияның қандай бөлігі нақты ақпаратпен пайдаланылмайтынын көрсетеді:

$$R = 1 - H / H_max$$

Мысалы, $H_max=4$ бит/символ, ал нақты энтропия $H=3$ бит/символ болсын. Онда:

$$D = 4 - 3 = 1 \text{ бит/символ}$$
$$R = 1 - 3/4 = 0,25$$

Демек, осы модель бойынша салыстырмалы артықтық 25%-ға тең. R мәні 0-ге жақын болса, көз максимал энтропияға жақын және символдарды болжау қиынырақ. R өскен сайын құрылымдық болжамдылық артады.

Кілт энтропиясы

Криптографияда энтропия ұғымы кілттің белгісіздігін бағалау үшін де қолданылады. Егер кілт K мүмкін мәннің ішінен бірдей ықтималдықпен таңдалса, оның энтропиясы:

$$H(K) = \log_2 K$$

Егер кілт n биттен тұрса және 2^n мүмкін кілттің барлығы бірдей ықтимал болса:

$$H(K) = \log_2(2^n) = n \text{ бит}$$

Мысалы, барлық 256 мүмкіндік бірдей ықтимал болатын 8 биттік кілттің теориялық энтропиясы 8 битке тең.

Бірақ кілт ұзындығы мен нақты кілт энтропиясы әрқашан бірдей емес. Егер пайдаланушы кілтті кездейсоқ емес, қайталанатын немесе оңай болжанатын үлгі бойынша таңдаса, мүмкін мәндердің нақты таралуы біркелкі болмайды. Мұндай жағдайда кілттің тиімді белгісіздігі оның формалды ұзындығынан төмен болуы мүмкін.

Сондықтан криптографиялық кілттерді сенімді кездейсоқ генерациялау кілт кеңістігінің мүмкіндігін толық пайдалануға көмектеседі.

Бірлік аралық ұғымы

Бірлік аралық (unicity distance) – классикалық криптография мен ақпарат теориясындағы теориялық ұғым. Ол шифрмәтін ұзындығы артқан сайын тілдің артықтығы мен кілт кеңістігі туралы ақпараттың бірегей дұрыс кілтті басқа ықтимал кілттерден ажыратуға жеткілікті бола бастайтын шамамен шекарасын сипаттайды.

Бұл ұғымды нақты шабуыл белгілі бір ұзындықтан кейін міндетті түрде кілтті табады деп түсінуге болмайды. Бірлік аралық – нақты есептеу уақытының немесе нақты криптоталдау алгоритмінің көрсеткіші емес. Ол ақпараттық-теориялық модель шеңберіндегі бағалау.

Шифрмәтін өте қысқа болғанда бірнеше әртүрлі кілт мағыналы болып көрінетін мәтіндер беруі мүмкін. Шифрмәтін ұзарған сайын тілдің статистикалық шектеулері көбейеді және жалған кілттерге сәйкес келетін мағыналы нұсқалардың саны азая түседі. Бірлік аралық осы процесті теориялық тұрғыдан сипаттайды.

Жуық байланыс

Классикалық модельдерде бірлік аралық кілт энтропиясының тілдің бір символға келетін артықтығына қатынасымен жуық байланыстырылуы мүмкін:

$$N_0 \approx H(K) / D$$

Мұндағы N_0 – бірлік аралықтың жуық мәні, H(K) – кілт энтропиясы, D – бір символға келетін тіл артықтығы.

Мысалы, оқу есебі ретінде H(K)=40 бит, ал D=2 бит/символ болсын. Онда:

$$N_0 \approx 40 / 2 = 20 \text{ символ}$$

Бұл нәтиже қабылданған теориялық модельде шамамен 20 символдан кейін статистикалық ақпарат бірегей кілтті ажыратуға жеткілікті бола бастау мүмкіндігін сипаттайды. Бірақ ол 20 символ болғанда кілт міндетті түрде практикалық түрде табылады дегенді білдірмейді.

Формула тіл моделіне, кілттердің таралуына, шифр түріне және басқа болжамдарға тәуелді. Сондықтан оны әмбебап нақты шекара ретінде емес, ақпараттық-теориялық түсіндіру құралы ретінде қолдану керек.

Криптоталдаумен байланысы

Классикалық шифрларда тілдің статистикалық артықтығы криптоталдаушыға қосымша ақпарат бере алады. Мысалы, моноалфавиттік ауыстыру шифрында символдардың орындары өзгермегендіктен, жиі кездесетін ашық мәтін символдарына сәйкес шифрмәтін символдары да жиі кездесуі мүмкін. Бұл жиілік талдауына негіз болады.

Егер шифр ашық мәтіндегі статистикалық құрылымды жеткілікті дәрежеде жасырмаса, ұзын шифрмәтін талдау үшін көбірек материал береді. Осы тұрғыдан тілдің артықтығы классикалық криптоталдауда маңызды рөл атқарады.

Қазіргі шифрлардың мақсаты – шифрмәтінде ашық мәтіннің пайдалы статистикалық іздерін қалдырмау және шифрмәтінді сыртқы бақылаушы үшін кездейсоқ деректерге ұқсас ету. Бұған көпраундты түрлендіру, шашырату, араластыру және кілтке тәуелді күрделі операциялар арқылы қол жеткізіледі.

Негізгі шамаларды салыстыру

Шама	Формула	Мағынасы
Ақпарат мөлшері	$I(x)=-\log_2 p(x)$	Жеке оқиғаның ақпараттылығы
Энтропия	$H(X)=-\Sigma p(x_i)\log_2 p(x_i)$	Орташа белгісіздік
Максимал энтропия	$H_max=\log_2 m$	Тең ықтималды m символ үшін максимум
Абсолюттік артықтық	$D=H_max-H$	Бір символға келетін артықтық
Салыстырмалы артықтық	$R=1-H/H_max$	Артықтықтың салыстырмалы үлесі
Кілт энтропиясы	$H(K)=\log_2 K$	Кілттің теориялық белгісіздігі
Бірлік аралық	$N_0 \approx H(K)/D$	Бірегей кілтті ажыратуға қатысты теориялық бағалау

Қорытынды

Ақпарат мөлшері мен энтропия ақпарат көзіндегі белгісіздікті сандық бағалауға мүмкіндік береді. Тең ықтималды символдарда энтропия максимал болады, ал ықтималдықтар біркелкі болмаған сайын энтропия төмендейді. Табиғи тілдің статистикалық және құрылымдық болжамдылығы артықтықты қалыптастырады.

Криптографияда бұл ұғымдар кілт белгісіздігін, тіл статистикасының криптоталдауға әсерін және бірлік аралық сияқты теориялық шамаларды түсіндіруге көмектеседі. Бірлік аралық нақты шабуыл нәтижесін емес, кілт энтропиясы мен тіл артықтығы арасындағы ақпараттық-теориялық байланысты сипаттайды.

Бақылау сұрақтары

1. Ақпарат мөлшері қалай анықталады?
2. Шеннон энтропиясының формуласы қандай?
3. Энтропия қай жағдайда максимал болады?
4. Тең ықтималды m символ үшін H_max неге тең?
5. Тілдің артықтығы дегеніміз не?
6. Абсолюттік артықтық қалай анықталады?
7. Салыстырмалы артықтық формуласы қандай?
8. Кілт энтропиясы нені сипаттайды?
9. Бірлік аралық дегеніміз не?
10. Бірлік аралықты неге нақты шабуыл кепілдігі ретінде қарастыруға болмайды?

Әдебиеттер тізімі:

1. Ахметов Б.С., Корченко А.Г., Сиденко В.П., Дрейс Ю.А., Алимсеитова Ж.К. Қолданбалы криптология: шифрлау әдістері. Оқулық. Алматы: ҚазҰТЗУ, 2016. – 500 б.

2. Ахметов Б.С., Кузнецов А.А., Краснобаев В.А., Алимсеитова Ж.К., Кузнецова Т.Ю. Криптография негіздері: сандар теориясының, топтардың, өрістердің, сақиналардың элементтері. Алматы: АЭжБУ, 2020.
3. Васильева И.Н. Криптографические методы защиты информации. Москва: Юрайт, 2024.
4. Лось А.Б., Нестеренко А.Ю., Рожков М.И. Криптографические методы защиты информации для изучающих компьютерную безопасность. Москва: Юрайт, 2024.
5. Омассон Ж.-Ф. О криптографии всерьез. Практическое введение в современное шифрование. Москва: ДМК Пресс, 2021.