giáo trình lý thuyết thông tin

Biên tập bởi:
duongvanhieu

giáo trình lý thuyết thông tin

Biên tập bởi:

duongvanhieu

Các tác giả:

phantantai

lequyetthang

duongvanhieu

Phiên bản trực tuyến:
http://voer.edu.vn/c/f89a66f8

MỤC LỤC

giới thiệu tổng quan giáo trình lý thuyết thông tin

GIỚI THIỆU TỔNG QUAN

GIÁO TRÌNH LÝ THUYẾT THÔNG TIN

MỤC ĐÍCH

Giáo trình này sẽ cung cấp cho người đọc những khối kiến thức cơ bản của lý thuyết thông tin như: Độ do lượng tin (Measure of Information), Sinh mã tách được (Decypherable Coding), Kênh truyền tin rời rạc không nhớ (Discrete Memoryless Channel) và Sửa lỗi trên kênh truyền (Error Correcting Codings).

Liên quan đến Độ đo lượng tin, giáo trình sẽ trình bày các khái niệm cơ bản về thông tin, entropy, một số công thức, tính chất, các định lý quan trọng của entropy và cách tính lượng tin.

Về Sinh mã tách được, giáo trình sẽ giới thiệu đến người học các vấn đề về yêu cầu của bài toán sinh mã, giải mã duy nhất, cũng như mã tức thời và giải thuật kiểm tra mã tách được. Các định lý quan trọng được đề cập trong nội dung này là: Định lý Kraft (1949), Định lý Shannon (1948) và Định lý sinh mã Huffman.

Về kênh truyền tin rời rạc không nhớ, giáo trình sẽ giới thiệu mô hình kênh truyền theo 2 khía cạnh vật lý và toán học. Các khái niệm về dung lượng kênh truyền, phân lớp kênh truyền, định lý về dung lượng kênh truyền, cũng như các khái niệm trong kỹ thuật truyền tin và phương pháp xây dựng lược đồ giải mã tối ưu cũng được trình bày trong môn học này.

Vấn đề Sửa lỗi (hay **xử lý mã sai) trên kênh truyền** là một vấn đề rất quan trọng và được quan tâm nhiều trong môn học này. Các nội dung được giới thiệu đến các bạn sẽ là Nguyên lý Khoảng cách Hamming, các định lý về Cận Hamming, phương pháp kiểm tra chẵn lẻ, các lược đồ sửa lỗi, Bảng mã Hamming và Bảng mã xoay vòng.

Hơn nữa, hầu hết các vấn đề nêu trên đều được đưa vào nội dung giảng dạy ở các bậc Đại học của một số ngành trong đó có ngành Công nghệ thông tin. Do đó, để có một tài liệu phục vụ công tác giảng dạy của giáo viên cũng như việc học tập và nghiên cứu của sinh viên, chúng tôi mạnh dạn biên soạn giáo trình này nhằm giúp cho sinh viên có một tài liệu tự học và nghiên cứu một cách hiệu quả.

yêu cầu

YÊU CẦU

Sau khi học xong môn này, sinh viên phải có được những khả năng sau:

Hiểu các khái niệm về về thông tin, Entropy, Entropy của một phân phối, Entropy của nhiều phân phối, Entropy có điều kiện, Độ đo lượng tin. Vận dụng giải quyết các bài toán về xác định lượng tin.

Biết khái niệm về mã tách được, mã không tách được, bảng mã tối ưu. Hiểu Định lý Kraft (1949), Định lý Shannon (1948), Định lý sinh mã Huffman và phương pháp sinh mã Huffman. Vận dụng để sinh bảng mã tách được tối ưu, nhận biết được bảng mã như thế nào là bảng mã tối ưu và có thể vận dụng để viết các chương trình sinh mã, giải mã (hay viết chương trình nén và giải nén). Từ đây, các sinh viên có thể tự nghiên cứu các loại bảng mã khác để vận dụng cho việc mã hóa và bảo mật thông tin một cách hiệu quả.

Biết các khái niệm về kênh truyền tin rời rạc không nhớ, dung lượng kênh truyền và phân lớp kênh truyền. Hiểu định lý về dung lượng kênh truyền, phương pháp xây dựng lược đồ giải mã tối ưu và cách tính xác suất truyền sai trên kênh truyền.

Biết các khái niệm về khoảng cách Hamming, nguyên lý khoảng cách Hamming, các định lý về Cận Hamming, phương pháp kiểm tra chẵn lẻ, các lược đồ sửa lỗi, Bảng mã Hamming và Bảng mã xoay vòng.

Vận dụng các kiến thức học được để thiết kế một hệ thống truyền nhận dữ liệu với quy trình cơ bản: mã hóa, giải mã và bảo mật thông tin.

Lý thuyết thông tin cũng là một trong các môn học khó của ngành Công nghệ thông tin vì nó đòi hỏi người học phải có kiến thức cơ bản về toán và xác suất thống kê. Do đó, đòi hỏi người học phải tự bổ sung các kiến thức cơ bản về toán và xác suất thống kê cho mình (nếu thiếu), tham gia lớp học đầy đủ và làm các bài tập theo yêu cầu của môn học thì mới tiếp thu kiến thức môn học một cách hiệu quả.

nội dung cốt lõi

NỘI DUNG CỐT LÕI

Giáo trình gồm 5 chương được trình bày trong 45 tiết giảng cho sinh viên chuyên ngành Công nghệ thông tin, trong đó có khoảng 30 tiết lý thuyết và 15 tiết bài tập mà giáo viên sẽ hướng dẫn cho sinh viên trên lớp.

Chương 1: Giới thiệu. *Chương này trình bày các nội dung có tính tổng quan về môn học bao gồm: các đối tượng nghiên cứu, mô hình lý thuyết thông tin theo quan điểm của nhà toán học Shannon, khái niệm về lượng tin biết và chưa biết, định lý cơ bản của kỹ thuật truyền tin.*

Chương 2: Độ đo lượng tin. *Chương này trình bày các vấn đề cơ bản về entropy, các tính chất của entropy, entropy của nhiều biến, entropy có điều kiện, các định lý về quan hệ giữa các entropy và lượng tin của một sự kiện.*

Chương 3: Sinh mã tách được. *Nội dung chính của chương này bao gồm các khái niệm về mã tách được, quan hệ giữa mã tách được và độ dài mã, tính tối ưu của độ dài mã.*

Chương 4: Kênh truyền. *Các nội dung được trình bày trong chương này bao gồm khái niệm về kênh truyền tin rời rạc không nhớ, các mô hình truyền tin ở khía cạnh vật lý và toán học, dung lượng trên kênh truyền, phân lớp các kênh truyền. Phương pháp xây dựng lược đồ giải mã tối ưu và cách tính xác suất truyền sai cũng được giới thiệu trong chương này.*

Chương 5: Sửa lỗi. *Chương này trình bày các nội dung cốt lõi sau: khái niệm về khoảng cách Hamming, nguyên lý khoảng cách nhỏ nhất Hamming, bổ đề về tự sửa lỗi và định lý Cận Hamming. Chương này cũng giới thiệu về bộ mã kiểm tra chẵn lẻ, phương pháp kiểm tra chẵn lẻ, lược đồ sửa lỗi tối ưu, mã Hamming và mã xoay vòng.*

kiến thức tiên quyết

KIẾN THỨC TIÊN QUYẾT

Để học tốt môn học này, đòi hỏi sinh viên phải nắm vững các môn học có liên quan như: xác suất thống kê, đại số boole (phép toán Modulo 2 và đa thức nhị phân). Các môn học có liên quan và có thể tham khảo thêm như kỷ thuật số, hệ điều hành, mạng máy tính.

TÀI LIỆU THAM KHẢO

David J.C. Mackey, *Information Theory, Infernce, and Learning Algorithms,* CamBridge University Express-2003.

G.J.ChaiTin, *Algorithmic Information Theory*, CamBridge University Express-1992.

Sanford Goldman, *Information Theory.*

http://www.inference.phy.cam.ac.uk/mackay/info-theory/course.html.

http://en.wikipedia.org/wiki/Information_theory.

http://www-2.cs.cmu.edu/~dst/Tutorials/Info-Theory/.

http://cscs.umich.edu/~crshalizi/notebooks/information-theory.html.

http://www.lecb.ncifcrf.gov/~toms/paper/primer/primer.pdf.

http://www.cs.ucl.ac.uk/staff/S.Bhatti/D51-notes/node27.html.

http://guest.engelschall.com/~sb/hamming/.

http://www2.rad.com/networks/1994/err_con/hamming.htm

phương pháp học tập

PHƯƠNG PHÁP HỌC TẬP

Để phục vụ cho mục tiêu nâng cao khả năng tự học tập và tự nghiên cứu của sinh viên, giáo trình này được biên soạn cùng với các giáo trình khác thuộc chuyên ngành Công nghệ thông tin của Khoa Công nghệ thông tin và Truyền thông – Đại Học Cần Thơ theo dự án **ASVIET002CNTT "Tăng cường hiệu quả đào tạo và năng lực đào tạo của sinh viên khoa Công nghệ Thông tin-Đại học Cần Thơ"**. Chúng tôi đã cố gắng trình bày giáo trình này một cách có hệ thống các nội dung theo bố cục các chương ứng với các khối kiến thức nêu trên, mỗi chương được được trình bày theo bố cục của các bài học và mỗi bài học giới thiệu đến người học một vấn đề nào đó trong số các vấn đề của một khối kiến thức tương ứng với một chương. Khi học xong các bài học của một chương, người học sẽ có một khối kiến thức cần thiết tương ứng cho môn học. Nội dung của các bài học đều được đưa vào các ví dụ để người học dễ hiểu, tùy theo từng vấn đề mà người học cần phải học và nghiên cứu trong thời lượng từ 1 đến 2 tiết tự học cho một bài học trong một chương. Như vậy, để học tốt môn học này, trước hết sinh viên cần phải:

Học đầy đủ các môn học tiên quyết, bổ sung những kiến thức cơ bản về toán và xác suất thống kê (nếu thiếu).

Học và nghiên cứu kỹ từng chương theo trình tự các chương được trình bày trong giáo trình này. Trong từng chương, học các bài theo thứ tự được trình bày, sau mỗi bài phải làm bài tập đầy đủ (nếu có).

Tham gia lớp đầy đủ, thảo luận các vấn đề tồn tại chưa hiểu trong quá trình tự học.

Sau mỗi chương học, phải nắm vững các khái niệm, các định nghĩa, các công thức tính toán và vận dụng giải các bài toán có tính chất tổng hợp được giới thiệu ở cuối chương.

Vận dụng kiến thức có được sau khi học xong các chương để giải một số bài tập tổng hợp ở cuối giáo trình, từ đó giúp cho người học hiểu sâu hơn về môn học và có thể giải quyết các vấn đề tương tự trong thực tế.

Việc cho ra đời một giáo trình với những mục đích như trên là không đơn giản khi khả năng và kinh nghiệm của người soạn còn có hạn, nhiều khái niệm, thuật ngữ dùng trong giáo trình chưa được định nghĩa một cách chính thống. Vì vậy giáo trình này chắc không tránh khỏi những khiếm khuyết, rất mong nhận được sự góp ý của các đồng nghiệp và người đọc.

giới thiệu

GIỚI THIỆU

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể biết:

Đối tượng nghiên cứu,

Mô hình lý thuyết thông tin theo quan điểm Shannon,

Các khái niệm về Lượng tin biết và lượng tin chưa biết,

Định lý cơ sở của kỹ thuật truyền tin,

Khái niệm chung về dung lượng kênh truyền,

Vấn đề sinh mã và giải mã.

Đối tượng nghiên cứu

Lý thuyết thống kê về thông tin được xây dựng trên hai hướng khác nhau bởi hai nhà toán học Shannon (1948) và Wiener (1949). Lý thuyết thông tin nghiên cứu quá trình xử lý tín hiệu như sau:

Đầu vào (input): nhận tín hiệu từ một lĩnh vực cụ thể, tức là tín hiệu xuất hiện theo các ký hiệu (symbol) từ một tập hợp cho trước và theo phân phối xác suất đã biết.

Tín hiệu được truyền đi trên kênh truyền (channel) và có thể bị nhiễu cũng theo một phân phối xác suất nào đó. Kênh truyền có thể được hiểu dưới hai nghĩa:

Dưới nghĩa vật lý: kênh truyền là một hệ thống truyền tín hiệu (dây dẫn, mạch, sóng, ...) và gây nhiễu tùy thao chất lượng của hệ thống.

Dưới nghĩa toán học: kênh truyền là các phân phối xác suất xác định trên lớp các tín hiệu đang xét ở đầu nhận tín hiệu (output).

Ở đầu ra (output): dựng lại tín hiệu chân thật nhất có thể có so với tín hiệu ở đầu vào.

Shannon xây dựng mô hình lý thuyết thông tin trên cơ sở giải quyết bài toán: sinh mã độ dài tối ưu khi nhận tín hiệu đầu vào. Tín tối ưu được xét trên 3 yếu tố sau:

Phân phối xác suất của sự xuất hiện của các tín hiệu.

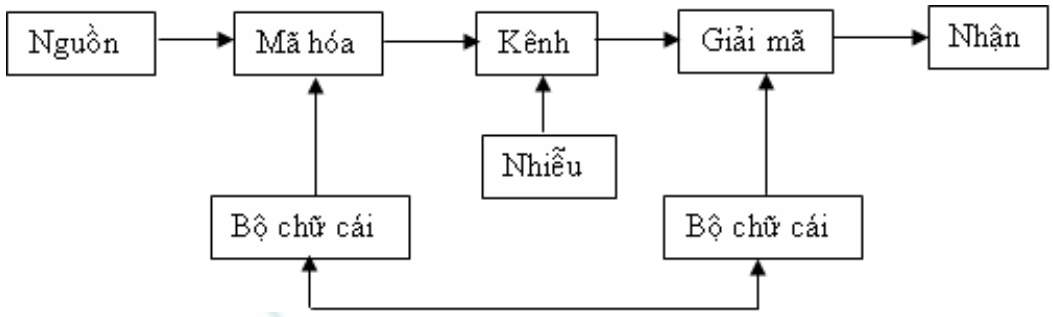
Tính duy nhất của mã và cho phép tự điều chỉnh mã sai nếu có với độ chính xác cao nhất. Giải mã đồng thời tự động điều chỉnh mã hoặc xác định đoạn mã truyền sai.

Trong khí đó, Wiener lại nghiên cứu phương pháp xử lý tín hiệu ở đầu ra: ước lượng tối ưu chuỗi tín hiệu so với chính nó khi nhận ở đầu vào không qua quá trình sinh mã. Như vậy phương pháp Wiener được áp dụng trong những trường hợp con người không kiểm soát được quá trình truyền tín hiệu. Môn "xử lý tín hiệu" đã đề cập đến vấn đề này.

mô hình lý thuyết thông tin theo quan điểm Shannon

Mô hình lý thuyết thông tin theo quan điểm Shannon

Lý thuyết thông tin được xét ở đây theo quan điểm của Shannon. Đối tượng nghiên cứu là một hệ thống liên lạc truyền tin (communication system) như sơ đồ dưới đây:



Diễn giải:

- Nguồn (source) thông tin còn gọi là thông báo cần được truyền ở đầu vào (Input).
- Mã hóa (encode) là bộ sinh mã. Ứng với một thông báo, bộ sinh mã sẽ gán cho một đối tượng (object) phù hợp với kỹ thuật truyền tin. Đối tượng có thể là:

Dãy số nghị phân (Digital) dạng: 01010101, cũng giống như mã máy tính.

Sóng liên tục (Analog) cũng giống như truyền radio.

- Kênh (channel) là phương tiện truyền mã của thông tin.
- Nhiễu (noise) được sinh ra do kênh truyền tin. Tùy vào chất lượng của kênh truyền mà nhiễu nhiều hay ít.
- Giải mã (decode) ở đầu ra (output) đưa dãy mã trở về dạng thông báo ban đầu với xác suất cao nhất. Sau đó thông báo sẽ được chuyển cho nới nhận. Trong sơ đồ trên, chúng ta quan tâm đến 2 khối mã hóa và giải mã trong toàn bộ môn học.

Lượng tin biết và chưa biết

Một biến ngẫu nhiên (BNN) X luôn mang một lượng tin nào đó. Nếu X chưa xảy ra (hay ta chưa biết cụ thể thông tin về X) thì lượng tin của nó là chưa biết, trong trường hợp này X có một lượng tin chưa biết. Ngược lại nếu X đã xảy ra (hay ta biết cụ thể thông

tin về X) thì lượng tin về biến ngẫu nhiên X coi như đã biết hoàn toàn, trong trường hợp này X có một lượng tin đã biết.

Nếu biết thông tin của một BNN X thông qua BNN Y đã xảy ra thì ta có thể nói: chúng ta chỉ biết một phần lượng thông tin của X đó trên cơ sở biết Y.

Ví dụ về lượng tin biết và chưa biết

Ta xét ví dụ về một người tổ chức trò chơi may rủi khách quan với việc tung một đồng tiền "có đầu hình – không có đầu hình". Nếu người chơi chọn mặt không có đầu hình thì thắng khi kết quả tung đồng tiền là không có đầu hình, ngược lại thì thua. Tuy nhiên người tổ chức chơi có thể "ăn gian" bằng cách sử dụng 2 đồng tiền "Thật- Giả" khác nhau sau:

+ Đồng tiền loại 1 (hay đồng tiền thật): đồng chất có 1 mặt có đầu hình.

+ Đồng tiền loại 2 (hay đồng tiền giả): đồng chất, mỗi mặt đều có 1 đầu hình.

Mặc dù người tổ chức chơi có thể "ăn gian" nhưng quá trình trao đổi 2 đồng tiền cho nhau là ngẫu nhiêu, vậy liệu người tổ chức chơi có thể "ăn gian" hoàn toàn được không? Hay lượng tin biết và chưa biết của sự kiện lấy một đồng tiền từ 2 đồng tiền nói trên được hiểu như thế nào?

Ta thử xét một trường hợp sau: nếu người chơi lấy ngẫu nhiên 1 đồng tiền và sau đó thực hiện việc tung đồng tiền lấy được 2 lần. Qua 2 lần tung đồng tiền, ta đếm được số đầu hình xuất hiện. Dựa vào số đầu hình xuất hiện, ta có thể phán đoán được người tổ chức chơi đã lấy được đồng tiền nào.

Chẳng hạn: Nếu số đầu hình đếm được sau 2 lần tưng là 1 thì đồng tiền đã lấy được là đồng tiền thật. Ngược lại nếu số đầu hình đếm được là 2 thì đồng tiền đã lấy được có thể là thật hay cũng có thể là giả. Như vậy, ta đã nhận được một phần thông tin về loại đồng tiền qua số đầu hình đếm được sau 2 lần tung. Ta có thể tính được lượng tin đó bằng bao nhiêu? *(Việc tính lượng tin này sẽ được thảo luận sau)*. Dưới đây là một số bảng phân phối của bài toán trên:

Gọi BNN X về loại đồng tiền (X=1 nếu lấy được đồng tiền loại 1 và X=1 nếu lấy được đồng tiền loại 2 được lấy).

Khi đó phân phối của X có dạng:

X	1	2
P	0.5	0.5

Đặt BNN Y là BNN về số đầu hình đếm được sau 2 lần tung. Khi đó ta có thể xác định được phân phối của Y với điều kiện xảy ra của X trong 2 trường hợp sau.

Phân phối của Y khi biết X=1 có dạng:

Y/X=1	0	1	2
P	0.25	0.5	0.25

Phân phối của Y khi biết X=2 có dạng:

Y/X=2	0	1	2
P	0	0	1

định lý cơ sở của kĩ thuật truyền tin

Định lý cơ sở của kỹ thuật truyền tin

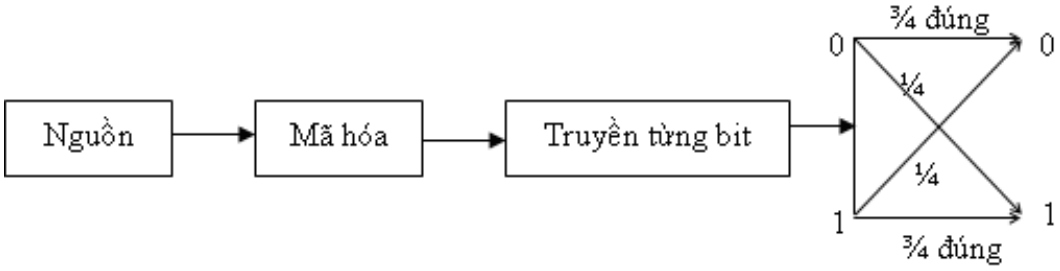
Trong" New Basic of Information Theory (1954)" Feinstein đã đưa ra định lý sau:"rên một kênh truyền có nhiễu, người ta luôn có thể thực hiện một phương pháp truyền sao cho đạt được sai số nhỏ hơn sai số cho phép (nhỏ bất kỳ) cho trước đối với kênh truyền."

Chúng ta sẽ không chứng minh định lý, thay vào đó, chúng ta sẽ tham khảo đến các minh họa giảm nhiễu trong các nội dung tiếp theo của bài học.

Mô tả trạng thái truyền tin có nhiễu

Giả sử, một thông báo được truyền đi trên một kênh truyền nhị phân rời rạc. Thông báo cần truyền được mã hóa thành dãy số nhị phân (0,1) và có độ dài được tính theo đơn vị bit. Giả sử 1 bit truyền trên kênh nhiễu với xác suất 1/4 (hay tính trung bình cứ truyền 4 bit thì có thể nhiễu 1 bit).

Ta có sơ đồ trạng thái truyền tin sau:



Minh họa kỹ thuật giảm nhiễu

Trong kỹ thuật truyền tin, người ta có thể làm giảm sai lầm khi nhận tin bằng cách truyền lặp lại 1 bit với số lẻ lần.

Ví dụ: truyền lặp lại 3 cho 1 bit cần truyền (xác suất nhiễu 1 bit bằng 1/4). Khi nhận 3 bit liền nhau ở cuối kếnh được xem như là 1 bit. Giá trị của bit này được hiểu là 0 (hay 1) nếu bit 0 (bit 1) có số lần xuất hiện nhiều hơn trong dãy 3 bit nhận được liền nhau (hay giải mã theo nguyên tắc đa số). Ta cần chứng minh với phương pháp truyền này thì xác suất truyền sai thật sự < 1/4 (xác suất nhiễu cho trước của kênh truyền).

Sơ đồ truyền tin:

Bit truyền	Tuyền lặp 3 lần	Nhận 3 bit	Giải mã
0 →	000 →	000	0
	000	001	0
	000	010	0
	000	100	0
	000	101	1
	000	011	1
	000	110	1
	000	111	1
1 →	111 →	000	0
	111	001	0
	111	010	0
	111	100	0
	111	011	1
	111	110	1
	111	111	1

Thật vậy:

Giả sử X_i xác định giá trị đúng hay sai của bit thứ i nhận được ở cuối kênh truyền với X_i =1 nếu bit thứ i nhận được là sai và X_i =0 nếu bit thứ i nhận được là đúng. Theo giả thiết ban đầu của kênh truyền thì phân phối xác suất của X_i có dạng Bernoulli b(1/4):

X_i	1	0
P	3/4	1/4

Gọi Y ={$X_1 + X_2 + X_3$ } là tổng số bit nhận sai sau 3 lần truyền lặp cho 1 bit. Trong trường hợp này Y tuân theo phân phối Nhị thức B(p,n), với p=1/4 (xác suất truyền sai một bit) và q =3/4 (xác suất truyền đúng 1 bit):

Y ~ B(i,n) hay

$$p(Y=i) = C_n^i . p^i q^{n-i}$$

Trong đó:

$$C_n^i = \frac{n!}{i!(n-i)!}$$

Vậy truyền sai khi Y thuộc {2, 3} có xác xuất là:

$$P_{sai} = P(y \geq 2) = P(Y=2) + P(Y=3) = B(2,3) + B(2,3)$$

Hay

$$Psai = (C_3^2(\frac{1}{4})^2.(\frac{3}{4})^1) + (C_3^3(\frac{1}{4})^3(\frac{3}{4})^0) = \frac{10}{64} < \frac{1}{4}$$

(đpcm).

Chi phí phải trả cho kỹ thuật giảm nhiễu

Theo cách thức lặp lại như trên, ta có thể giảm sai lầm bao nhiêu cũng được (lặp càng nhiều thì sai càng ít), nhưng thời gian truyền cũng tăng lên và chi phí truyền cũng sẽ tăng theo.

Hay ta có thể hiểu như sau:

Lặp càng nhiều lần 1 bit=>thời gian truyền càng nhiều=>chi phí càng tăng.

khái niệm về dung lượng kênh truyền

Khái niệm về dung lượng kênh truyền

Ví dụ trên cho chúng ta thấy cần phải xác định một thông số cho truyền tin để đảm bảo sai số chấp nhận được và đồng thời tốc độ truyền cũng không quá chậm.

Khái niệm "dung lượng" kênh truyền là khái niệm rất cơ bản của lý thuyết truyền tin và là một đại lượng vật lý đồng thời cũng là đại lượng toán học (có đơn vị là bit). Nó cho phép xác định tốc độ truyền tối đa của mỗi kênh truyền. Do đó, dựa vào dung lượng kênh truyền, người ta có thể chỉ ra tốc độ truyền tin đồng thời với một phương pháp truyền có sai số cho phép.

Vấn đề sinh mã

Từ kỹ thuật truyền tin trên cho ta thấy quá trình sinh mã và giải mã được mô tả như sau: một đơn vị thông tin nhận được ở đầu vào sẽ được gán cho một ký hiệu trong bộ ký hiệu sinh mã. Một ký hiệu mã được gán n lần lặp lại (dựa vào dung lượng của kênh truyền, ta có thể xác định được n). Thiết bị sinh mã (Coding device/ Encoder) sẽ thực hiện quá trình sinh mã.

Như vậy, một đơn vị thông tin từ nguồn phát tin sẽ được thiết bị sinh mã gán cho một dãy n ký hiệu mã. Dãy ký hiệu mã của 1 đơn vị thông tin được gọi là một từ mã (Code word). Trong trường hợp tổng quát, người ta có thể gán một khối ký tự mã cho một khối thông tin nào đó và được gọi là một từ mã.

Vấn đề giải mã

Ở cuối kênh truyền, một thiết bị giải mã (Decoding device/ Decoder) sẽ thực hiện quá trình ngược lại như sau: kiểm tra dãy ký hiệu mã để quyết định giải mã về một từ mã và đưa nó về dạng khối tin ban đầu.

Ví dụ:

Khối tin ban đầu : 01010101

Khối ký hiệu mã ở đầu truyền (lặp 3 lần): 000111000111000111000111.

Khối ký hiệu mã ở đầu nhận : 001110100111011001000111

Khối tin nhận được cuối cùng : 01011001 (sai 2 bit so với khối tin ban đầu)

Do đó làm sao để đua khối tin nhận được về khối tin ban đầu 01010101, đây chính là công việc của bộ giải mã (Decoder).

Một vấn đề quan trọng cần lưu ý là phải đồng bộ giữa tốc độ nạp thông tin (phát tín hiệu) với tốc độ truyền tin. Nếu tốc độ nạp thông tin bằng hoặc lớn hơn so với tốc độ truyền tin của kênh, thì cần phải giảm tốc độ nạp thông tin sao cho nhỏ hơn tốc độ truyền tin.

độ đo lượng tin

ĐỘ ĐO LƯỢNG TIN

Mục tiêu: trình bày các khái niệm về độ đo lượng tin chưa biết và đã biết về một biến ngẫu nhiên X. Tính toán các lượng tin này thông qua định nghĩa và các tính chất của Entropy từ một hay nhiều biến ngẫu nhiên.

ENTROPY

Mục tiêu

Sau khi hoàn tất bài học này bạn có thể:

- Hiểu được các khái niệm Entropy,
- Biết Entropy của một sự kiện và Entropy của một phân phối,
- Hiểu Định lý dạng giải tích của Entropy,
- Biết Bài toán về cây tìm kiếm nhị phân và
- Làm kiến thức cơ sở để hiểu và học tốt các bài học tiếp theo.

Ví dụ về entropy

Trước hết, ta cần tìm hiểu một ví dụ về khái niệm độ do của một lượng tin dựa vào các sự kiện hay các phân phối xác suất ngẫu nhiên như sau:

Xét 2 BNN X và Y có phân phối sau:

X={1, 2, 3, 4, 5} có phân phối đều hay p(X=i) = 1/5.

Y={1, 2} cũng có phân phối đều hay p(Y=i) = 1/2.

Bản thân X và Y đều mang một lượng tin và thông tin về X và Y chưa biết do chúng là ngẫu nhiên. Do đó, X hay Y đều có một lượng tin không chắc chắn và lượng tin chắc chắn, tổng của 2 lượng tin này là không đổi và thực tế nó bằng bao nhiêu thì ta chưa thể biết. Lượng tin không chắc chắn của X (hay Y) được gọi là Entropy.

Tuy nhiên, nếu X và Y có tương quan nhau thì X cũng có một phần lượng tin không chắc chắn thông qua lượng tin đã biết của Y (hay thông tin về Y đã được biết). Trong trường hợp này, một phần lượng tin không chắc chắn của thông qua lượng tin đã biết của Y được gọi là Entropy có điều kiện.

Nhận xét về độ đo lượng tin

Rõ ràng, ta cần phải xây dựng một đại lượng toán học rất cụ thể để có thể đo được lượng tin chưa biết từ một biến ngẫu nhiên. Một cách trực quan, lượng tin đó phải thể hiện được các vấn đề sau:

Một sự kiện có xác suất càng nhỏ thì sự kiện đó ít xảy ra, cũng có nghĩa là tính không chắc chắn càng lớn. Nếu đo lượng tin của nó thì nó cho một lượng tin không biết càng lớn.

Một tập hợp các sự kiện ngẫu nhiên (hay Biến ngẫu nhiên) càng nhiều sự kiện có phân phối càng đều thì tính không chắc chắn càng lớn. Nếu đo lượng tin của nó thì sẽ được lượng tin không biết càng lớn. Hay lượng tin chắc chắn càng nhỏ.

Một phân phối xác suất càng lệch nhiều (có xác xuất rất nhỏ và rất lớn) thì tính không chắc chắn càng ít và do đó sẽ có một lượng tin chưa biết càng nhỏ so với phân phối xác suất đều hay lượng tin chắc chắn của nó càng cao.

Khái niệm entropy

Trong tiếng việt ta chưa có từ tương đương với từ Entropy, tuy nhiên chúng ta có thể tạm hiểu hiểu thoáng qua trước khi đi vào định nghĩa chặc chẽ về mặt toán học của Entropy như sau:

Entropy là một đại lượng toán học dùng để đo lượng tin không chắc (hay lượng ngẫu nhiên) của một sự kiện hay của phân phối ngẫu nhiên cho trước. Hay một số tài liệu tiếng anh gọi là Uncertainty Measure.

Entropy của một sự kiện

Giả sử có một sự kiện A có xác suất xuất hiện là p. Khi đó, ta nói A có một lượng không chắc chắn được đo bởi hàm số h(p) với p [0,1]. Hàm h(p) được gọi là Entropy nếu nó thoả 2 tiêu đề toán học sau:

Tiên đề 01: h(p) là hàm liên tục không âm và đơn điệu giảm.

Tiên đề 02: nếu A và B là hai sự kiện độc lập nhau, có xác suất xuất hiện lần lượt là p_A và p_B. Khi đó, $p(A,B) = p_A.p_B$ nhưng $h(A,B) = h(p_A) + h(p_B)$.

Entropy của một phân phối

Xét biến ngẫu nhiên X có phân phối:

X	x_1	x_2	x_3	...	x_M
P	p_1	p_2	p_3	...	p_M