Teoria da Informação

Prof. Leonardo Araújo

https://sites.google.com/site/leolca/teaching/information-theory

Teoria da Informação e Codificação

- Surgiu em 1948 com a publicação do trabalho "The Mathematical Theory of Communications" Shannon (1948).
- Teoria da Informação lida com as limitações teóricas e potencialidades de sistemas de comunicação.
- O que é informação? Como mensurar?
- Codificação. Como representar uma informação?
- Canal de Comunicação.

Teoria da Informação e Codificação

- Surgiu em 1948 com a publicação do trabalho "The Mathematical Theory of Communication" Shannon (1948).
- Teoria da Informação lida com a eficiência e confiabilidade de sistemas de comunicação.
- O que é informação? Como mensurá-la?
- Codificação. Como representar uma informação?
- Canal de Comunicação.

- A distinguibilidade entre as mensagens é fator importante para caracterizar informação. Pela definição de Shannon, "Informação é a habilidade de distinguir de forma confiável dentre um rol de alternativas possíveis".
- Shannon cunhou o termo 'auto-informação' de um evento ou mensagem aleatória definindo-o como "menos o logaritmo da probabilidade do evento aleatório". A 'entropia' da fonte estocástica que gera os eventos é o valor esperado da auto-informação.
- Shannon mostrou que a entropia de uma fonte estocástica possui um significado físico: em média, é o menor número de bits necessários para representar ou comunicar de forma fidedigna eventos gerados por uma fonte estocástica.
- O problema central em Teoria da Informação é a transmissão eficiente e confiável de dados, do transmissor a um receptor, através de um canal de comunicação.
- Eficiência (usar o mínimo de recursos possível).
- Confiabilidade (evitar erros, ser capaz de detectá-los e corrigi-los).

Teoria da Informação e Codificação

- Surgiu em 1948 com a publicação do trabalho "The Mathematical Theory of Communication" Shannon (1948).
- Teoria da Informação lida com a eficiência e confiança e potencialidades de sistemas de comunicação.
- O que é informação? Como mensurar?
- Codificação. Como representar uma informação?
- Canal de Comunicação.

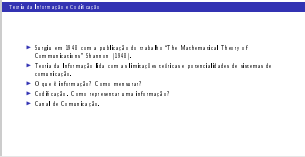
- A informação é um conceito paradoxal. Por um lado necessita de uma representação física, por outro lado é abstrata. Uma mesma informação pode ser representada em papel, em um meio magnético ou ótico, pode ser representada por ondas mecânicas ou elétricas.
- Linguagem. Comunicação falada e escrita. Código faz associação entre símbolo e mensagem e é 'arbitrário'.

Teoria da Informação e Codificação

- Surgiu em 1948 com a publicação do trabalho "The Mathematical Theory of Communication" Shannon (1948).
- Teoria da Informação lida com a eficiência e confiança e potencialidades de sistemas de comunicação.
- O que é informação? Como mensurar?
- Codificação. Como representar uma informação?
- Canal de Comunicação.

A área da ciência criada por Shannon ampliou-se ao longo do tempo e influência diversas outras áreas. Por exemplo: teoria da comunicação, criptografia, ciência da computação, física (mecânica estatística), matemática (probabilidade e estatística), filosofia da ciência, linguística e processamento de linguagem natural, reconhecimento de fala, reconhecimento de padrões e aprendizado de máquina, compressão de dados, economia, biologia e genética, psicologia, etc.

Shannon entrou para o Bell Labs para trabalhar com sistemas de controle de disparo e criptografia durante a Segunda Guerra Mundial, sob um contrato com o Comitê Nacional de Pesquisa para Defesa Em 1945, Shannon elaborou um memorando sigiloso, que posteriormente foi publicado sob o título "Communication Theory of Secrecy Systems". Este incorporava muitos dos conceitos e formulações matemáticas do artigo mais consagrado "A Mathematical Theory of Communication" (1948).



- Codificação - criar códigos com algoritmos práticos para codificação e decodificação para serem utilizados na comunicação no mundo real em canais ruidosos.
- Exemplos de codificações conhecidas para representar informação: código Morse, código ASCII, etc.

	Compressão (codificação de fonte) eficiência	Correção de Erros (codificação de canal) confiabilidade
Teoria da Informação (matemática)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmo)	Códigos Simbólicos: código de Huffman Códigos de Fluxo: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager

	Compressão (codificação de fonte) eficiência	Correção de Erros (codificação de canal) confiabilidade
Teoria da Informação (paramétrica)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmos)	Códigos Simbólicos: código de Huffman Código de Fluxo: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager

Huffman (1952) sem perdas; ótimo (dentre os códigos de símbolos); simples implementação; PNG, JPEG, MPEG, WinZip, GZip, MP3, AAC.

Codificação Aritmética (1978, patente IBM) Peter Elias (1963): trabalho não publicado; problema: precisão infinita (solucionado, Jorma Rissanen e Richard Pasco, 1976) patentes (IBM, todas já expiraram); JPEG, JBIG, Skype, PPM, PAQ, DjVu.

Lempel-Ziv (1978), Lempel–Ziv–Welch (1984) assintoticamente ótimo; eficiente e simples de ser implementado; PNG, GIF, PKZip, GZipg, PDF.

Códigos de Hamming verificação de paridade; primeiro código de correção de erros efetivamente bom (1950); RAM (podem ocorrer erros no processo de escrita e leitura), RAID 2.

Código de Reed-Solomon (1960) muitas vezes combinados com códigos convolucionais, por exemplo: RSV (algoritmo de vitterbi); robusto a erros em rajada; códigos de barra, CD, DVD, Blue-Ray, DSL (telefone), DVB; (Digital Video Broadcasting), RAID 6, comunicação satélite e sondas espaciais (Voyager 1977, Galileo 1989, Cassini-Huygens; 1997, Pathfinder 1996, MER 2003).

	Compressão (codificação de fonte) eficiência	Correção de Erros (codificação de canal) confiabilidade
Teoria da Informação (paramétrica)	Compressão sem perdas: teorema da codificação de fonte	Teorema da Codificação de Canal
Métodos de Codificação (algoritmos)	Códigos Simbólicos: código de Huffman Código de Fluxo: Codificação Aritmética, Lempel-Ziv	Códigos de Hamming, Códigos BCH Códigos Turbo Códigos de Gallager

códigos Turbo (1993) 3G, 4G, LTE, WiMax, Mars Reconnaissance Orbiter (MRO) 2005.

códigos Gallager: low-density parity-check (LDPC) Robert G. Gallager, MIT (1960), não eram práticos para os computadores da época; David J. C. MacKay e Radford M. Neal (1996); 10Gb/s Ethernet, WiFi 802.11 N, Internet por rede elétrica ITU-T G.hn, DVB-S2 (tv via satélite), China Mobile Multimedia Broadcasting (CMMB) transmissão multimídia via satélite para aparelhos móveis.

Código Morse

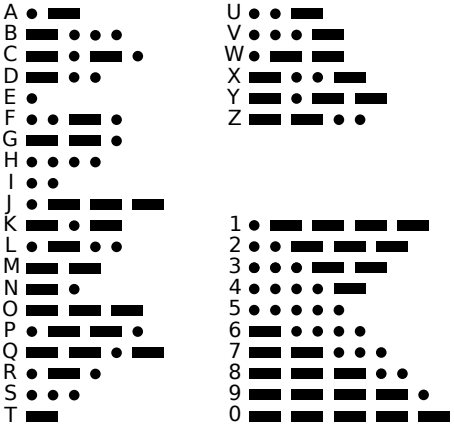


Figura 1: Código Morse internacional (Wikipedia (2020d)). Letras do alfabeto ordenadas por frequência de ocorrência no inglês: etaoin shrdlu cmfwyp vbgkjq xz (Wikipedia (2020a,c)).

Código Unário

Claude Mendibil utilizou o código unário (1, 01, 001, 0001, ...) para representar as letras do alfabeto ESARINTULOMDPCFBVHGJQZYXKW. Utilizando este código, Jean-Dominique Bauby ditou o livro *Le Scaphandre et le Papillon* (O Escafandro e a Borboleta).

Figura 2: Foto de Bauby em 1996 ditando suas memórias para Claude Mendibil (Wikipedia (2020b)).

Compressão

- Compressão é importante para utilizarmos melhor os recursos disponíveis.
- Shannon mostrou que o limite para a representação é a entropia.

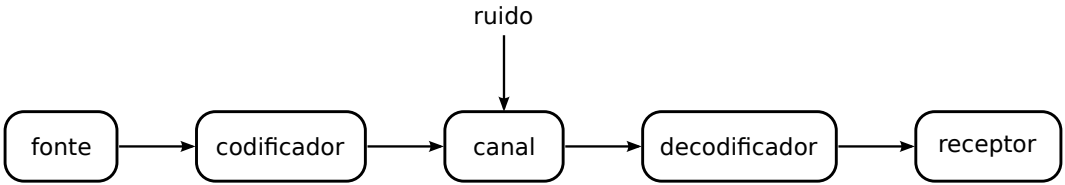
Compressão

- Compressão é importante para utilizarmos melhor os recursos disponíveis.
- Shannon mostrou que o limite para a representação é a entropia.

Compressão é importante para utilizarmos melhor os recursos disponíveis.

1. Armazenar mais dados em meio (disco rígido, memória, fita, etc).
2. Transmitir mais informação através de um canal (essencialmente, armazenar e transmitir são o mesmo problema).
3. Diminuir o desgaste do meio ao reduzir o número de vez que se faz leitura e escrita. Solid State Drives (SSDs) baseados em memórias flash NAND possuem um número finito de ciclos de programar/apagar. É importante reduzir a quantidade de bits que serão gravados para aumentar a vida útil dessas memórias/discos. (A compressão LZ4 vem sendo utilizada com esta finalidade, e também para que o S.O. tenha um boot mais rápido)

Modelo Geral de Comunicação



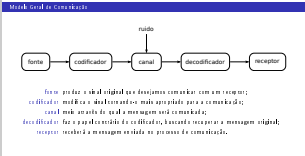
fonte produz o sinal original que desejamos comunicar com um receptor;

codificador modifica o sinal tornando-o mais apropriado para a comunicação;

canal meio através do qual a mensagem será comunicada;

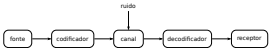
decodificador faz o papel contrário do codificador, buscando recuperar a mensagem original;

receptor receberá a mensagem enviada no processo de comunicação.



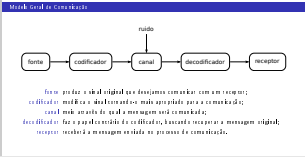
- Separação do codificador/decodificador em duas partes: codificador/decodificador de fonte e codificador/decodificador de canal.
- Remover redundância do sinal produzido pela fonte e acrescentar redundância por causa do ruído no canal de comunicação.
- Fontes e Canais de comunicação: discretos ou contínuos.

Modelo Geral de Comunicação



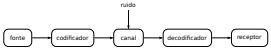
fonte: produz o sinal original que desejamos comunicar com um receptor;
codificador: modifica o sinal transmitido para o mais apropriado para a comunicação;
canal: meio através do qual a mensagem será comunicada;
decodificador: faz a operação inversa do codificador, buscando recuperar a mensagem original;
receptor: receberá a mensagem enviada no processo de comunicação.

No contexto de Teoria da Informação, fonte é qualquer coisa que produza uma mensagem, um sinal que carregue informação. Podemos considerar uma fonte que produz mensagens como: voz, sons, palavras, imagens, vídeo, sequência de bits de um programa de computador, etc.



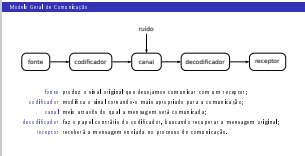
Canal é o meio através do qual o sinal produzido pela fonte será transmitido/propagado/armazenado. Por exemplo: espaço aberto (ar), linha telefônica, link de rádio, link em uma comunicação espacial, disco rígido, CD, DVD, fita magnética (armazenamento - transmissão no tempo ao invés de espaço pode sofrer deterioração ao longo do tempo); DNA de seres vivos ao longo de gerações, envio de mensagens por estímulos elétricos ou químicos em um organismo biológico.

Modelo Geral de Comunicação

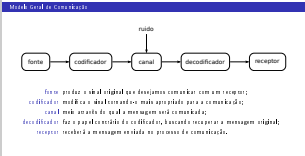


fonte: produz o sinal original que queremos comunicar com um receptor;
codificador: modifica o sinal transmitido para o mais apropriado para a comunicação;
canal: meio através do qual a mensagem será comunicada;
decodificador: faz a operação inversa do codificador, buscando recuperar a mensagem original;
receptor: receberá a mensagem enviada no processo de comunicação.

Receptor é aquele a quem é destinada a mensagem transmitida. Exemplos: computador ou equipamento, uma pessoa, rádio, tv, sistema de áudio, etc.

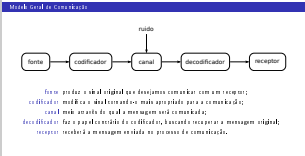


Ruído é qualquer sinal que interfere com aquele que está sendo transmitido. Exemplos: ruído térmico, ruído impulsivo, cross-talk, outro sinal qualquer indesejado. Ruído representa a nossa compreensão imperfeita do universo. Desta forma, tratamos ruído como algo aleatório e que usualmente obedece certas regras, tais como uma determinada distribuição probabilística.



O codificador processa o sinal antes de inseri-lo no canal de comunicação.

- Redução dos dados, removendo redundância do sinal.
- Inserção de redundâncias de acordo com as características do canal de comunicação, para garantir integridade aos dados transmitidos.
- Codificação para representar as informações de um sinal sob a forma de outro sinal.



O decodificador faz o papel inverso do codificador.

- Remove os erros de transmissão.
- Recupera a informação original enviada pela fonte.

Notação

X é uma variável aleatória

x é uma valor que a v.a. assume

$\mathcal{X}$ é o alfabeto de tamanho $|\mathcal{X}| = K$ dentro do qual a v.a. assume valores, $\mathcal{X} = \{a_1, \ldots, a_K\}$

$\mathcal{P}_X$ é o conjunto de probabilidades associadas aos valores, $\mathcal{P}_X = \{p_1, \ldots, p_K\}$, tais que $p_i \geq 0$ e $\sum_{i=1}^{K} p_i = 1$

p_i é a probabilidade da v.a. assumir um determinado valor, $p_i = \Pr(X = a_i)$

$\mathcal{P}_X = p$ é a distribuição da v.a., $\sum_{x \in \mathcal{X}} \Pr(X = x) = 1$

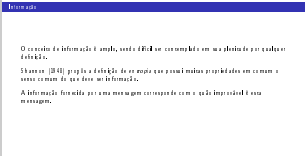
$X \sim p$, a v.a. X possui distribuição p

Informação

O conceito de informação é amplo, sendo difícil ser contemplado em sua plenitude por qualquer definição.

Shannon (1948) propôs a definição de *entropia* que possui muitas propriedades em comum o senso comum do que deve ser informação.

A informação fornecida por uma mensagem corresponde com o quão improvável é esta mensagem.



O que é previsível fornece pouca ou nenhuma informação.

Quanto mais incerto, mais informação há.

Informação

Hartley (1928) propõem uma medida de informação para uma variável aleatória X:

$$I(X) = \log_b L, \tag{1}$$

onde L é o número de possíveis valores que X pode assumir. Se $b = 2$, a informação será medida em 'bits' (nome sugerido por J.W. Tukey).

Introdução

Hartley (1928) propôs uma medida de informação para uma variável aleatória X:

$$I(X) = \log_b L, \tag{1}$$

onde L é o número de possíveis valores que X pode assumir. Se $b = 2$, a informação será medida em "bits" (como sugerido por J.W. Tukey).

A definição de Hartley é condizente com as seguintes intuições sobre informação:

- Dois cartões de memória devem possuir o dobro da capacidade de um cartão para armazenamento de informação.
- Dois canais de comunicação idênticos devem possuir o dobro da capacidade de transmitir informação que um único canal.
- Um dispositivo com duas posições estáveis, como um relé ou um flip-flop, armazena um bit de informação. N dispositivos deste tipo podem armazenar N bits de informação, já que o número total de estados é 2^N e $\log_2 2^N = N$.

Entretanto, isto é válido apenas quando as mensagens/eventos são equiprováveis. No caso extremo, note que se o cartão de memória armazena apenas zeros, ele não é capaz de armazenar informação alguma.

Entropia

Suponha que existam eventos E_k com probabilidade de ocorrência p_k.

- Shannon: informação associada ao evento E_k é dada por $I(E_k) = \log(1/p_k)$.
  - Se $p_k = 1 \rightarrow$ não há surpresa na ocorrência do evento E_k.
  - Se $p_k = 0 \rightarrow$ surpresa infinita, afinal o evento E_k é impossível.
  - $I(E_k) = -\log p(E_k)$ é a auto-informação do evento ou mensagem E_k.
- **Sempre** utilizaremos a base 2 para o cálculo do logaritmo, desta forma $\log \equiv \log_2$, a menos que seja especificado o contrário.
- $\ln$ é o logaritmo na base natural e.

Entropia

- Notação: $p(x) = P_X(X = x)$, a probabilidade do evento $\{X = x\}$, da v.a. X assumir o valor x.
- Valor esperado da v.a. X: $E[X] = EX = \sum_x xp(x)$.
- Dada uma função $g : \mathcal{X} \to \mathbb{R}$, o valor esperado da v.a. $g(X)$ é $Eg(X) = \sum_x g(x)p(x)$.
- Considere $g(x) = \log(1/p(x))$. Então $g(x)$ é a imprevisão (surpresa) de encontrar o evento $X = x$. Tomando o valor esperado de g teremos

$$\sum_x p(x) \log \frac{1}{p(x)}, \tag{2}$$

ou seja, a esperança da surpresa, ou o valor esperado da imprevisão na variável aleatória X. Esta é a definição de entropia.

Entropia

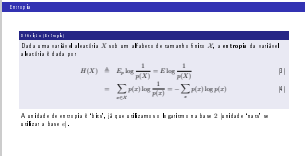
Definição (Entropia)

Dada uma variável aleatória X sob um alfabeto de tamanho finito $\mathcal{X}$, a **entropia** da variável aleatória é dada por

$$H(X) \triangleq E_p \log \frac{1}{p(X)} = E \log \frac{1}{p(X)} \tag{3}$$

$$= \sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)} = -\sum_{x} p(x) \log p(x) \tag{4}$$

A unidade de entropia é 'bits', já que utilizamos o logaritmo na base 2 (unidade 'nats' se utilizar a base e).



- Entropia mede a grau de incerteza associado a uma distribuição.
- Entropia mede a desordem ou o espalhamento de uma distribuição.
- Entropia mede a 'escolha' que a fonte tem na escolha de símbolos de acordo com uma densidade (maior entropia implica em mais escolha).
- Vamos utilizar a seguinte convenção: $0 \log 0 = 0$.

Entropia

Se uma v.a. $X \sim p(x)$, então o valor esperado de uma função desta v.a., $g(X)$, é dada por

$$E[g(X)] = \sum_{x \in \mathcal{X}} p(x) g(x). \tag{5}$$

A entropia de X pode ser interpretada como o valor esperado da v.a. $\log \frac{1}{p(X)}$, onde X é descrita pela função massa de probabilidade $p(x)$.

$$H(X) = E\left[\log \frac{1}{p(X)}\right]. \tag{6}$$

Entropia

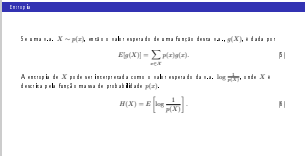
Se uma v.a. $X \sim p(x)$, então o valor esperado de uma função desta v.a., $g(X)$, é dado por:

$$E[g(X)] = \sum_{x \in \mathcal{X}} p(x) g(x). \quad \square$$

A entropia de X pode ser interpretada como o valor esperado da v.a. $\log \frac{1}{p(x)}$, onde X é denotada pela função massa de probabilidade $p(x)$:

$$H(X) = E\left[\log \frac{1}{p(X)}\right]. \quad \square$$

- Entropia é uma medida da real 'incerteza' média, o que é uma medida sobre toda a distribuição.
- Entropia mede o grau de incerteza médio ou esperado do resultado de uma distribuição de probabilidade.
- É uma medida de desordem ou espalhamento. Distribuições com alta entropia devem ser planas, mais uniformes, enquanto distribuições com baixa entropia devem possuir poucas modas (unimodal, bimodal).



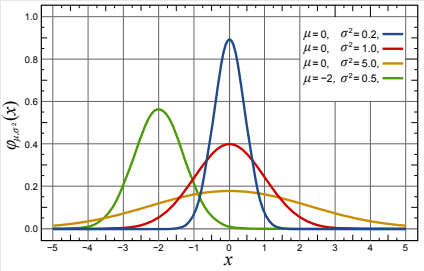
Entropia

Se uma v.a. $X \sim p(x)$, então o valor esperado de uma função desta v.a., $g(X)$, é dado por:

$$E[g(X)] = \sum_{x \in \mathcal{X}} p(x) g(x). \qquad (5)$$

A entropia de X pode ser interpretada como o valor esperado da v.a. $\log \frac{1}{p(x)}$, onde X é descrita pela função massa de probabilidade $p(x)$:

$$H(X) = E\left[\log \frac{1}{p(X)}\right]. \qquad (6)$$



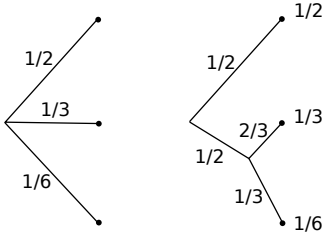
- mais concentrado: menor entropia
- mais espalhado: maior entropia
- os valores em x não importam, apenas os valores das probabilidades associadas $p(x)$ importam no cálculo da entropia

Escolha, Incerteza e Entropia I

Suponha um conjunto de eventos cujas probabilidades de ocorrências sejam dadas por $p_1, p_2, \ldots, p_n$. É possível encontrar uma medida de quanta 'escolha' está envolvida na seleção de um evento ou quão incertos estamos da saída?
Para tal medida $H(p_1, p_2, \ldots, p_n)$, é razoável requerermos as seguintes propriedades:

1) H deve ser contínuo em p_i;
2) Se todos os p_i são iguais, $p_i = \frac{1}{n}$, então H deve ser uma função monotonicamente crescente de n (quando temos eventos equiprováveis, teremos mais incerteza quão maior for o número de eventos possíveis);
3) Se for possível quebrar uma escolha em uma sequência de escolhas sucessivas, a medida H original deve ser a soma ponderada dos valores individuais das medidas H_i após a quebra.

Escolha, Incerteza e Entropia II



$$H(\frac{1}{2},\frac{1}{3},\frac{1}{6}) = H(\frac{1}{2},\frac{1}{2}) + \frac{1}{2}H(\frac{2}{3},\frac{1}{3}) \tag{7}$$

A única função H que satisfaz às suposições acima é da forma Shannon (1948):

$$H = -K\sum_{i=1}^{k} p(i)\log p(i) \ , \tag{8}$$

onde K é uma constante positiva.

Demonstração da Equação (8) I

Nesta secção iremos apresentar a demonstração de $H = -\sum p_i \log p_i$ (conforme Apêndice 2 de Shannon (1948)).
Vamos definir

$$A(n) = H\left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right). \tag{9}$$

Desejamos que uma escolha dentre s^m opções igualmente prováveis possa ser decomposta como uma sequência de m escolhas que se subdividem em s possibilidades igualmente prováveis.

Demonstração da Equação (8) II

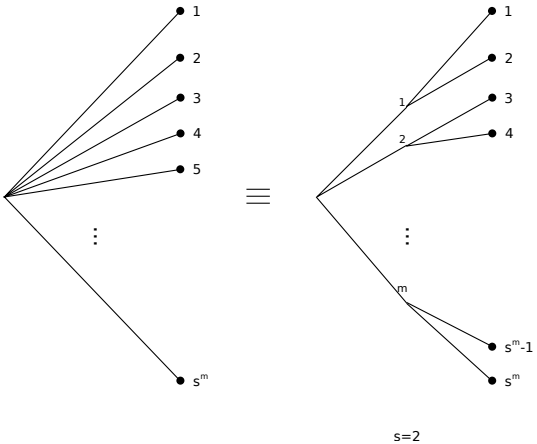


Figura 3: Exemplo de equivalência para $s = 2$.

Demonstração da Equação (8) III

Teremos então que

$$A(s^m) = mA(s). \tag{10}$$

Da mesma forma, para t e n, teremos $A(t^n) = nA(t)$. Podemos tomar n arbitrariamente grande e encontrar m que satisfaça

$$s^m \leq t^n \leq s^{(m+1)}. \tag{11}$$

Tomando o logaritmo[1] da expressão acima e dividindo por $n \log s$ todos os termos[2], teremos

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n}, \tag{12}$$

o que é equivalente a

$$\left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \epsilon, \tag{13}$$

onde ϵ é arbitrariamente pequeno, já que n é arbitrariamente grande.

Demonstração da Equação (8) IV

Usando agora a propriedade desejada de monotonicidade de $A(n)$, teremos

$$\begin{aligned} A(s^m) \leq A(t^n) \quad &\leq A(s^{(m+1)}) \\ mA(s) \leq nA(t) &\leq (m+1)A(s) \end{aligned} \tag{14}$$

Dividindo a expressão acima por $nA(s)$, teremos

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n}, \tag{15}$$

ou, de forma equivalente,

$$\left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \epsilon, \tag{16}$$

e assim, como as duas frações ($\log t / \log s$ e $A(t)/A(s)$) estão ϵ próximas de m/n, podemos concluir que

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| < 2\epsilon. \tag{17}$$

Demonstração da Equação (8) V

Como ϵ é arbitrariamente pequeno, no limite teremos

$$\begin{aligned} \frac{A(t)}{A(s)} &= \frac{\log t}{\log s} \\ A(t) &= \frac{A(s)}{\log s}\log t = K\log t, \end{aligned} \tag{18}$$

onde K deve ser positivo, de forma que $A(n)$ seja monótona crescente.
Suponha uma escolha com n possibilidades em que as probabilidades são comensuráveis, $p_i = {}^{n_i}/\sum n_i$, onde n_i são inteiros. De forma equivalente, uma escolha entre $\sum n_i$ opções pode ser expressa como uma escolha dentre n opções com probabilidades $p_1, \ldots, p_n$, e para uma

Demonstração da Equação (8) VI

i-ésima dada escolha, realizar uma nova escolha dentre n_i opções igualmente prováveis. Teremos então:

$$\begin{aligned} \overbrace{K \log \left(\sum n_i \right)}^{A(\sum n_i)} &= H(p_1, \ldots, p_n) + \overbrace{K \log n_i}^{A(n_i)} \\ K \underbrace{\left(\sum p_i \right)}_{=1} \log \left(\sum n_i \right) &= H(p_1, \ldots, p_n) + K \underbrace{\left(\sum p_i \right)}_{=1} \log n_i. \end{aligned} \tag{19}$$

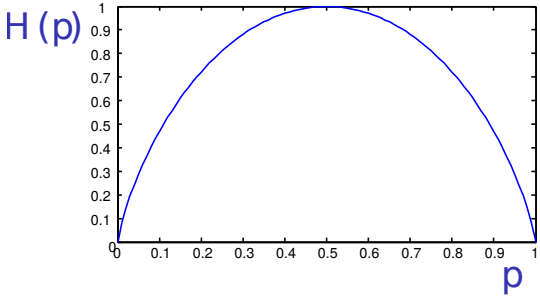
E assim,

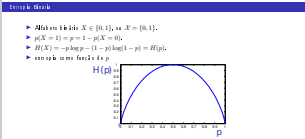
$$\begin{aligned} H(p_1, \ldots, p_n) &= K \left[\left(\sum p_i \right) \log \left(\sum n_i \right) - \left(\sum p_i \right) \log n_i \right] \\ &= -K \sum p_i \log \frac{n_i}{\sum n_i} = -K \sum p_i \log p_i. \quad \square \end{aligned} \tag{20}$$

[1] Logaritmo é uma função monótona crescente.
[2] $n \log s$ é positivo para $n \geq 0$ e $s \geq 1$.

Entropia Binária

- Alfabeto binário $X \in \{0, 1\}$, ou $\mathcal{X} = \{0, 1\}$.
- $p(X = 1) = p = 1 - p(X = 0)$.
- $H(X) = -p \log p - (1 - p) \log(1 - p) = H(p)$.
- entropia como função de p





- maior incerteza ($H = 1$) quando $p = 0.5$ e menor incerteza ($H = 0$) quando $p = 0$ ou $p = 1$.
- note que a entropia $H(p)$ é concava em p.

Entropia - GNU Octave

```
function H = entropy(p,b)

  if (nargin == 0 || nargin > 2) print_usage (); endif;
  if any(p < 0) | any(p > 1) | abs(sum(p)-1) > 1E-10, error('not a
      ↪ valid pmf!'); endif;

  id = find(p!=0);
  p = p(id);
  H = sum( - p .* log2(p) );

  if nargin > 1, H *= log(2)/log(b); endif;

endfunction
```

[download do código]

Entropia - GNU Octave - demo

```
%!demo
%! p = [0.5 0.5];
%! H = entropy(p);
%! printf('The pmf p has entropy = %.2f bits.\n',H);
%! He = entropy(p,e);
%! printf('The pmf p has entropy = %.2f nats.\n',He);
%! p = [0:0.02:1];
%! for i=1:length(p), H(i) = entropy([p(i), (1-p(i))]); endfor;
%! figure; plot(p,H); xlabel('p'); ylabel('H(p) (bits)'); title('
    ↪ binary entropy');
```

Entropia - Exemplo

Suponha uma v.a. $X \in \mathcal{X} = \{a, b, c, d\}$ com distribuição dada por

$$X = \begin{cases} a, & \text{com probabilidade } \frac{1}{2}, \\ b, & \text{com probabilidade } \frac{1}{4}, \\ c, & \text{com probabilidade } \frac{1}{8}, \\ d, & \text{com probabilidade } \frac{1}{8}. \end{cases} \tag{21}$$

A entropia associada será dada por

$$\begin{aligned} H(X) &= H\left(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{8}\right) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \\ &= -\frac{1}{2}\log\frac{1}{2} - \frac{1}{4}\log\frac{1}{4} - \frac{1}{8}\log\frac{1}{8} - \frac{1}{8}\log\frac{1}{8} \\ &= \frac{1}{2} + \frac{2}{4} + \frac{3}{8} + \frac{3}{8} = \frac{14}{8} = \frac{7}{4} \end{aligned} \tag{22}$$

Entropia Conjunta

Duas variáveis aleatórias X e Y possuem **entropia conjunta**

$$H(X,Y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = E\log\frac{1}{p(X,Y)}. \tag{23}$$

Generalizando para vetores $X_{1:N} = (X_1, X_2, \ldots, X_N)$

$$\begin{aligned} H(X_{1:N}) &= H(X_1, X_2, \ldots, X_N) \\ &= \sum_{x_1,x_2,\ldots,x_N} p(x_1,\ldots,x_N)\log\frac{1}{p(x_1,\ldots,x_N)} \\ &= E\log\frac{1}{p(X_1,\ldots,X_N)} \end{aligned} \tag{24}$$

Entropia Condicional I

Dadas duas v.a. X e Y relacionadas por $p(x, y)$, conhecer o evento $X = x$ pode alterar a entropia de Y.

- Entropia condicionada a um evento $H(Y|X = x)$

$$\begin{aligned} H(Y|X = x) &= E \log \frac{1}{p(Y|X = x)} \\ &= -\sum_{y \in \mathcal{Y}} p(y|x) \log p(y|x) \end{aligned} \tag{25}$$

- $H(Y|X = x)$ é uma função de X. Podemos então tomar seu valor esperado $E[H(Y|X = x)]$ e obter a entropia condicional $H(Y|X)$.

Entropia Condicional II

- Realizando a média sobre todos os x, obteremos a entropia condicional $H(Y|X)$.

$$\begin{aligned} H(Y|X) &= \sum_x p(x) H(Y|X=x) \\ &= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\ &= -\sum_{x,y} p(x,y) \log p(y|x) \\ &= E \log \frac{1}{p(Y|X)} \end{aligned} \tag{26}$$

Regra da Cadeia

Teorema (Regra da Cadeia para a Entropia)

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{27}$$

Demonstração.

$$-\log p(x,y) = -\log p(x) - \log p(y|x) \tag{28}$$

tomando o valor esperado de ambos os lados, obtemos o resultado desejado. □

Corolário

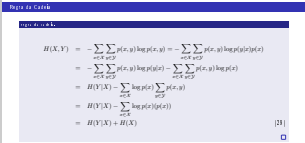
Se $X \perp\!\!\!\perp Y$ *então* $H(X,Y) = H(X) + H(Y)$.

Regra da Cadeia

regra da cadeia.

$$
\begin{aligned}
H(X,Y) &= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x,y) = -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x)p(x) \\
&= -\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(y|x) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log p(x) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}} \log p(x) \sum_{y\in\mathcal{Y}} p(x,y) \\
&= H(Y|X) - \sum_{x\in\mathcal{X}} \log p(x)(p(x)) \\
&= H(Y|X) + H(X) \qquad (29)
\end{aligned}
$$

□



Exemplo Canal de Comunicação

Suponha um canal de comunicação com entrada X e saída Y.
$H(X|Y)$ pode ser visto como a incerteza sobre X (a mensagem enviada) quando Y (a mensagem recebida) for conhecido.
Sem nenhuma observação no processo de comunicação através deste canal, o receptor não sabe nada sobre X nem Y, assim a incerteza inicial é $H(X,Y)$. Quando o receptor recebe a mensagem Y, ele ganha uma quantidade de informação $H(Y)$. Assim a informação que falta sobre X mesmo conhecendo Y é dada por $H(X|Y) = H(X,Y) - H(Y)$. Esta pode ser tido como uma medida do erro na comunicação.
A quantidade de informação que o receptor de fato ganha é $I(X;Y) = H(X) - H(X,Y)$.

Regra da Cadeia Generalizada I

Teorema (Regra da Cadeia para a Entropia)

$$H(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, X_2, \ldots, X_{i-1}) \tag{30}$$

$$\begin{aligned} H(X_1, X_2, \ldots, X_N) &= H(X_1) + H(X_2|X_1) + \\ &\quad H(X_3|X_1, X_2) + H(X_4|X_1, X_2, X_3) + \cdots \end{aligned} \tag{31}$$

Regra da Cadeia Generalizada II

Demonstração.

Utilizando a regra da cadeia da probabilidade condicional, teremos

$$p(x_1, x_2, \ldots, x_N) = \prod_{i=1}^{N} p(x_i | x_1, \ldots, x_{i-1}), \tag{32}$$

então

$$-\log p(x_1, x_2, \ldots, x_N) = -\sum_{i=1}^{N} \log p(x_i | x_1, x_2, \ldots, x_{i-1}) \tag{33}$$

tomando o valor esperado de ambos os lados, obtemos o resultado desejado. □

Propriedades da Entropia

1) H é uma função estritamente côncava de X, i.e., para $0 \leq \lambda \leq 1$ e variáveis aleatórias X e Y

$$H(\lambda X + (1-\lambda)Y) \geq \lambda H(X) + (1-\lambda)H(Y) \tag{34}$$

com igualdade sse (se e somente se) $\lambda = 0$ ou $\lambda = 1$ ou $X = Y$.

2) $H(X) \geq 0$ com igualdade sse $p(X)$ for não nulo apenas em um ponto $x_0 \in \mathcal{X}$.
3) $H(X) \leq \log |\mathcal{X}|$ com igualdade sse $p(X)$ for uniforme ($p \sim \frac{1}{n}$).
4) $H(X)$ é uma função apenas das probabilidades $p(x_i)$, independente da ordem ou rótulo.
5) $H_b(X) = (\log_b a) H_a(X)$.

Continuidade da Entropia I

Todas as medidas de informação de Shannon são funções contínuas das distribuições conjuntas das variáveis aleatórias envolvidas.

Definição (distância das variações)

Seja p e q duas distribuições probabilísticas em um alfabeto comum $\mathcal{X}$. A distância das variações entre p e q é definida por

$$V(p,q) = \sum_{x \in \mathcal{X}} |p(x) - q(x)|. \tag{35}$$

Dado um alfabeto finito fixo $\mathcal{X}$, considere $\mathcal{P}_{\mathcal{X}}$ o conjunto de todas as distribuições em $\mathcal{X}$. A entropia para uma dada distribuição p sobre o alfabeto $\mathcal{X}$ é definida por

$$H(p) = -\sum_{x \in S_p} p(x) \log p(x), \tag{36}$$

Continuidade da Entropia II

onde S_p denota o suporte de p, ou seja, $S_p \subset \mathcal{X}$.
Para que $H(p)$ seja contínuo com respeito à convergência em distância das variações, em uma determinada distribuição $p \in \mathcal{P}_{\mathcal{X}}$, devemos ter que, para qualquer $\epsilon > 0$, existe $\delta > 0$ tal que

$$|H(p) - H(q)| < \epsilon, \tag{37}$$

para todo $q \in \mathcal{P}_{\mathcal{X}}$ satisfazendo

$$V(p, q) < \delta, \tag{38}$$

ou, de forma equivalente,

$$\lim_{p' \to p} H(p') = H\left(\lim_{p' \to p} p'\right) = H(p), \tag{39}$$

onde a convergência $p' \to p$ é em distância das variações.

Continuidade da Entropia III

Como $a \log a \to 0$ quando $a \to 0$, definimos uma função $l : [0, \infty) \to \mathbb{R}$ da forma

$$l(a) = \begin{cases} a \log a & \text{se } a > 0, \\ 0 & \text{se } a = 0, \end{cases} \tag{40}$$

ou seja, $l(a)$ é uma extensão contínua de $a \log a$. Podemos reescrever a entropia da seguinte forma

$$H(p) = -\sum_{x \in \mathcal{X}} l(p(x)), \tag{41}$$

onde o somatório é tomado em todo $x \in \mathcal{X}$ ao invés de S_p. Defininido uma função $l_x : \mathcal{P}_{\mathcal{X}} \to \mathbb{R}$, para todo $x \in \mathcal{X}$, da forma

$$l_x(p) = l(p(x)), \tag{42}$$

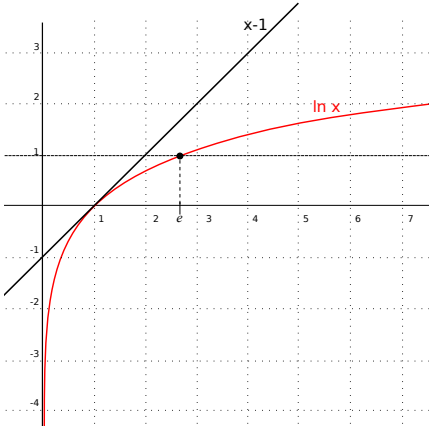
teremos

$$H(p) = -\sum_{x \in \mathcal{X}} l_x(p). \tag{43}$$

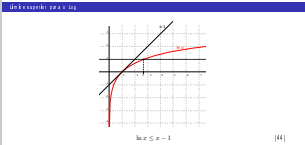
Continuidade da Entropia IV

Evidentemente $l_x(p)$ é contínua em p (com relação à convergência em distância das variações). Como o somatório na Equação 43 possui apenas um número finito de termos, podemos concluir que $H(p)$ é uma função contínua de p.

Limite superior para o Log



$$\ln x \leq x - 1 \tag{44}$$



$\ln x \leq x - 1$, para $x \geq 1$

- Sabemos que para $x = 1$ é verdadeiro, $0 = \ln 1 \leq 1 - 1 = 0$.
- Vamos demonstrar que $\ln x \leq x - 1$, para $x \geq 1$ por contradição.

Suponha que existe $b > 1$ tal que $\ln x > x - 1$, para $x = b$. Vamos definir $f(x) = \ln x - x + 1$, logo $f(1) = 0$ (conforme visto acima) e $f(b) > 0$ (por hipótese). Pelo teorema do valor médio $\exists c$, $1 < c < b$, tal que

$$f'(c) = \frac{f(b) - f(1)}{b - 1} = \frac{\overbrace{f(b)}^{>0}}{\underbrace{b - 1}_{>0}} > 0. \tag{45}$$

Mas, $f'(x) = 1/x - 1$, e assim $f'(x) < 0$ para $x > 1$. Logo há uma contradição e nossa hipótese é falsa. Teremos assim $\ln x \leq x - 1$, para $x \geq 1$.
Para $x \in (0, 1)$, basta seguir os mesmos passos, escolhendo um ponto $x = b$, tal que $0 < b < 1$. Vamos encontrar um ponto c tal que $0 < b < c < 1$. E usar o teorema do valor médio para mostrar uma contradição na hipótese.

Valor Máximo da Entropia (discreta) I

Teorema (Limite Superior da Entropia)

Seja $X \in \{x_1, x_2, \ldots, x_n\}$. *Então* $H(X) \leq \log n$, *sendo a igualdade alcançada se e somente se* $p(X = x_i) = \frac{1}{n}$ *para todo* i.

Valor Máximo da Entropia (discreta) II

Demonstração.

Vamos mostrar que $H(X) - \log n \leq 0$.

$$
\begin{aligned}
H(X) - \log n &= -\sum_x p(x) \log p(x) - \log n \overbrace{\sum_x p(x)}^{=1} \\
&= -\sum_x p(x) \log p(x) - \sum_x p(x) \log n \\
&= -\sum_x p(x) \log p(x) n = \log_2 e \sum_x p(x) \ln \frac{1}{p(x)n} \\
&\leq \log_2 e \sum_x p(x) \left[\frac{1}{p(x)n} - 1 \right]
\end{aligned}
$$

...

Valor Máximo da Entropia (discreta) III

Demonstração.

continuação...

$$
\begin{aligned}
H(X) - \log n &\leq \ldots \\
&= \log_2 e \left[\underbrace{\sum_x \frac{1}{n}}_{\sum_{x\in\mathcal{X}} \frac{1}{n} = n\frac{1}{n} = 1} - \underbrace{\sum_x p(x)}_{=1} \right] = 0
\end{aligned}
\tag{46}
$$

□

Valor Máximo da Entropia

Na demonstração acima utilizamos $\ln z \leq z - 1$. A igualdade $\ln z = z - 1$ se dará no ponto estacionário $z = 1$, isto é, quando $\frac{1}{p(x)n} = 1$, ou seja, quando $p(x) = 1/n$, teremos assim uma distribuição uniforme.

Se tivermos $p_i = 1/n$, então

$$-\sum_i p_i \log p_i = -\sum_i \frac{1}{n} \log \frac{1}{n} = -\log \frac{1}{n} = \log n. \tag{47}$$

Podemos mostrar (através da concavidade da entropia) que este é o único conjunto de valores com esta propriedade.

- Entropia aumenta quando a distribuição se torna mais uniforme.

Valor Máximo da Entropia I

Outra demonstração...

Demonstração.

Considere $X \in \mathcal{X} = \{x_1, x_2, \ldots, x_n\}$ com probabilidades $p = \{p_1, p_2, \ldots, p_n\}$, respectivamente. A entropia de X é dada por

$$\begin{aligned} H(X) &= -\sum_{i=1}^{n} p_i \log p_i \\ &= -\sum_{i=1}^{n-1} p_i \log p_i - p_n \log p_n \\ &= -\left(\frac{1}{\ln 2}\right)\left[\sum_{i=1}^{n-1} p_i \ln p_i + p_n \ln p_n\right]. \end{aligned} \tag{48}$$

...

Valor Máximo da Entropia II

Demonstração.

continuação...

Da mesma forma, podemos expressa p_n da seguinte maneira

$$p_n = 1 - \sum_{i=1}^{n-1} p_i. \tag{49}$$

Utilizando 49 em 48, podemos expressar a entropia $H(X)$ como uma função de $n-1$ probabilidades p_i. O máximo será dado quando a seguinte condição ocorrer

$$\frac{\partial H(X)}{\partial p_k} = 0 \quad \text{for} \quad k = 1, \ldots, n-1 \ . \tag{50}$$

...

Valor Máximo da Entropia III

Demonstração.

continuação...

Teremos então

$$
\begin{aligned}
0 = \frac{\partial H(X)}{\partial p_k} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}\left[\sum_{i=1}^{n-1} p_i \ln p_i + p_n \ln p_n\right] \\
&= -\left(\frac{1}{\ln 2}\right)\left[\ln p_k + 1 + (\ln p_n + 1)\frac{\partial p_n}{\partial p_k}\right] \\
&= -\left(\frac{1}{\ln 2}\right)\left[\ln p_k + 1 - (\ln p_n + 1)\right], \qquad (51)
\end{aligned}
$$

onde utilizamos a Equação 49, que nos fornece $\partial p_n / \partial p_k = -1$.

...

Valor Máximo da Entropia IV

Demonstração.

continuação...

A Equação 51 mostra que devemos encontrar $\ln p_k = \ln p_n$ para cada $k = 1, \ldots, n-1$. Todas as $n-1$ equações serão satisfeitas quando todas as probabilidades p_k forem iguais a $1/n$.

...

Valor Máximo da Entropia V

Demonstração.

continuação...

Devemos agora calcular a derivada segunda para mostrar que o extremo que achamos é de fato um máximo.

$$\begin{aligned} \frac{\partial^2 H(X)}{\partial p_k^2} &= -\left(\frac{1}{\ln 2}\right)\frac{\partial}{\partial p_k}\left[\ln p_k - \ln p_n\right] \\ &= -\left(\frac{1}{\ln 2}\right)\left[\frac{1}{p_k} + \frac{1}{p_n}\right] \leq 0 \ , \end{aligned} \tag{52}$$

já que as probabilidades são valores positivos. Escolhendo então $p_k = 1/n$, teremos a entropia máxima.

□

Subdividindo a entropia em partes I

A entropia deve permanecer inalterada, mesmo quando subdividimos as escolhas em partes.

Exemplo (Exemplo simples)

Suponha uma v.a. X com alfabeto $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$ e distribuição $q = (q_1.q_2, q_3, q_4)$. A entropia associada a esta variável aleatória é dada por

$$\begin{aligned} H(X) &= H(q_1, q_2, q_3, q_4) \\ &= -\sum_{i=1}^{4} q_i \log q_i \\ &= -q_1 \log q_1 - q_2 \log q_2 - q_3 \log q_3 - q_4 \log q_4. \end{aligned} \tag{53}$$

...

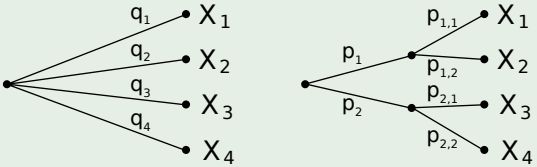
Subdividindo a entropia em partes II

Exemplo (Exemplo simples)

continuação...

Se dividirmos a escolha na determinação de X em duas escolhas sucessivas, conforme ilustrado na figura abaixo, poderemos então escrever

$$H(q_1, q_2, q_3, q_4) = H(p_1, p_2) + p_1 H(p_{1,1}, p_{1,2}) + p_2 H(p_{2,1}, p_{2,2}). \tag{54}$$



...

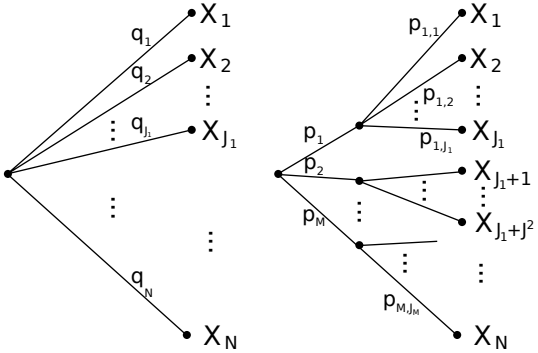
Subdividindo a entropia em partes III

Exemplo (Exemplo simples)

continuação...

$$
\begin{aligned}
H(X) &= -q_1 \log q_1 - q_2 \log q_2 - q_3 \log q_3 - q_4 \log q_4 \\
&= -p_1 p_{1,1} \log p_1 p_{1,1} - p_1 p_{1,2} \log p_1 p_{1,2} - p_2 p_{2,1} \log p_2 p_{2,1} - p_2 p_{2,2} \log p_2 p_{2,2} \\
&= -p_1 p_{1,1} \log p_1 - p_1 p_{1,1} \log p_{1,1} - p_1 p_{1,2} \log p_1 - p_1 p_{1,2} \log p_{1,2} \ldots \\
&\quad -p_2 p_{2,1} \log p_2 - p_2 p_{2,1} \log p_{2,1} - p_2 p_{2,2} \log p_2 - p_2 p_{2,2} \log p_{2,2} \\
&= -p_1 \log p_1 \left(p_{1,1} + p_{1,2}\right) - p_2 \log p_2 \left(p_{2,1} + p_{2,2}\right) \ldots \\
&\quad +p_1 \left(-p_{1,1} \log p_{1,1} - p_{1,2} \log p_{1,2}\right) \ldots \\
&\quad +p_2 \left(-p_{2,1} \log p_{2,1} - p_{2,2} \log p_{2,2}\right) \\
&= H(p_1, p_2) + p_1 H(p_{1,1}, p_{1,2}) + p_2 H(p_{2,1}, p_{2,2}) \qquad (55)
\end{aligned}
$$

Subdividindo a entropia em partes IV



Subdividindo a entropia em partes V

De forma geral, como $q_n = p_m p_{m,j}$, teremos

$$\begin{aligned}
H(X) &= -\sum_{n=1}^{N} q_n \log q_n = -\sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_m p_{m,j} \\
&= -\sum_{m=1}^{M}\sum_{j=1}^{J_m} (p_m p_{m,j} \log p_m + p_m p_{m,j} \log p_{m,j}) \\
&= -\sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_m - \sum_{m=1}^{M}\sum_{j=1}^{J_m} p_m p_{m,j} \log p_{m,j} \\
&= -\sum_{m=1}^{M} p_m \log p_m \left(\sum_{j=1}^{J_m} p_{m,j}\right) - \sum_{m=1}^{M} p_m \sum_{j=1}^{J_m} p_{m,j} \log p_{m,j} \\
&= H(p_1, \ldots, p_M) + \sum_{m=1}^{M} p_m H(p_{m,1}, \ldots, p_{m,J_m}). \qquad (56)
\end{aligned}$$

Embaralhar

- Suponha que X seja uma v.a. indicando as posições de cartas (i.e. $X = x$ representa um conjunto de posições, uma determinada configuração).
- Seja T uma operação de embaralhamento independente, i.e. $T \perp\!\!\!\perp X$.
- Então $H(TX) \geq H(X)$.

$$
\begin{aligned}
H(TX) &\geq H(TX|T) \\
&\quad \text{onde utilizamos que condicionar não pode aumentar a entropia, como veremos adiante} \\
&= H(T^{-1}TX|T) \\
&\quad \text{como } T \text{ é conhecido, aplicá-lo novamente ou seu inverso não altera a entropia} \\
&= H(X|T) = H(X) \qquad (57) \\
&\quad \text{onde utilizamos que } T \perp\!\!\!\perp X
\end{aligned}
$$

Permutação

O que ocorre se permutarmos as probabilidades?
Seja $p = (p_1, p_2, \ldots, p_n)$, uma distribuição discreta de probabilidade e $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_n)$ uma permutação de $1, 2, \ldots, n$.
Considere $p_\sigma = (p_{\sigma_1}, p_{\sigma_2}, \ldots, p_{\sigma_n})$ uma permutação da distribuição p.
Quem será maior? $H(p)$ ou $H(p_\sigma)$?

$$H(p) = -\sum_i p_i \log p_i = -\sum_j p_{\sigma_j} \log p_{\sigma_j} = H(p_\sigma). \tag{58}$$

Sumário

Definição de Entropia

$$H(X) = -\sum_{x} p(x) \log p(x) \tag{59}$$

Entropia Conjunta

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \tag{60}$$

Entropia Condicional

$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \tag{61}$$

Regra da Cadeia

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{62}$$

Limites da Entropia

$$0 \leq H(X) \leq \log n, \text{onde } n \text{ é o tamanho do alfabeto de } X. \tag{63}$$

Entropia do Jogo de Adivinhação

Qual é a melhor estratégia para adivinhar o valor de uma variável aleatória com perguntas sim/não do tipo "$X \in S$?", para algum conjunto $S \subseteq D_X$ (domínio da v.a. X).

Exemplo

Seja $X \in D_X = \{x_1, x_2, x_3, x_4, x_5\}$ com probabilidades

x	x_1	x_2	x_3	x_4	x_5
$p(x)$	0.3	0.2	0.2	0.15	0.15

Considere a seguinte estratégia: 1) $X = x_5$? 2) $X = x_4$? 3) $X = x_3$? 4) $X = x_2$? 5) $X = x_1$?

Desta forma faremos 5 perguntas 30% das vezes, 4 perguntas 20% das vezes, etc.

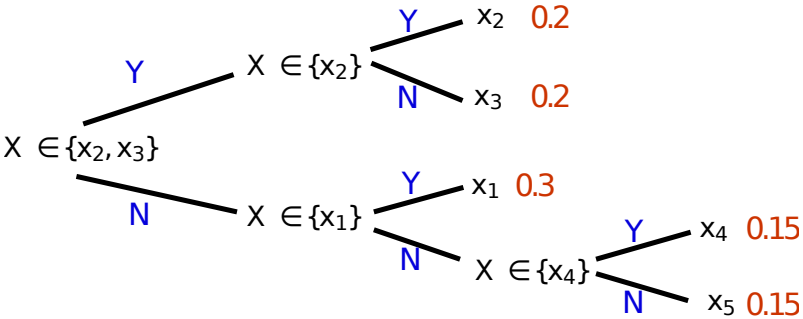
O número médio de perguntas é: $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (5, 4, 3, 2, 1)^\intercal = 3.35$.

Se invertermos a ordem das perguntas teremos: $(0.3, 0.2, 0.2, 0.15, 0.15) \cdot (1, 2, 3, 4, 5)^\intercal = 2.65$.

Existe uma estratégia melhor?

Entropia do Jogo de Adivinhação

Considere a estratégia ilustrada abaixo.



O número médio de perguntas será: $2(0.2 + 0.2 + 0.3) + 3(0.15 + 0.15) = 2.3$
Note que $H(X) = 2.271$.
O número médio de perguntas é sempre $\geq H(X)$.

Entropia do Jogo de Adivinhação

Vamos analisar a melhor e a pior estratégias, vistas anteriormente, com relação à forma como elas dividem a distribuição.

- pior estratégia: $X = x_5$?
  Divide a distribuição em dois grupos ($X = x_5$ e $X \neq x_5$), com probabilidades $p(X = x_5) = 0.15$, $p(X \neq x_5) = 0.85$, e a entropia será $H(0.15, 0.85) = 0.6098$.
- melhor estratégia: $X \in \{x_2, x_3\}$?
  $p(X \in \{x_2, x_3\}) = 0.4$, $p(X \notin \{x_2, x_3\}) = 0.6$, $H(0.4, 0.6) = 0.971$.
- De forma geral, é melhor realizar primeiro perguntas que, analisadas como variáveis aleatórias, possuem maior entropia (algoritmo guloso).
- Note a relação com $H(Y|X) + H(X) = H(X, Y)$. Se fizermos uma pergunta com $H(X)$ grande, a entropia residual $H(Y|X)$ fica menor.

Entropia do Jogo de Adivinhação

Vamos analisar a melhor e a pior estratégia, vistas anteriormente, com relação à forma como elas dividem a distribuição.

- pior estratégia: $X = x_3$?
  Dividir a distribuição em dois grupos $\{X = x_3, X \neq x_3\}$, com probabilidades $p(X = x_3) = 0.15$, $p(X \neq x_3) = 0.85$, a entropia será $H(0.15, 0.85) = 0.6098$.
- melhor estratégia: $X \in \{x_2, x_3\}$?
  $p(X \in \{x_2, x_3\}) = 0.4$, $p(X \notin \{x_2, x_3\}) = 0.6$, $H(0.4, 0.6) = 0.971$.
- De forma geral, melhores perguntas, perguntas que, avaliadas como variáveis aleatórias, possuem maior entropia (algoritmo guloso).
- Nota a relação com $H(Y|X) + H(X) = H(X, Y)$. Se fizermos uma pergunta com $H(X)$ grande, a entropia residual $H(Y|X)$ é menor.

Veremos adiante que o algoritmo guloso não é ótimo (entropia mínima).

Intuição sobre Informação Mútua

- Dadas duas variáveis aleatórias X e Y, quanta informação uma possui sobre a outra?
- Conhecendo X, quanto sabemos sobre Y? Conhecendo Y, quanto sabemos sobre X?
- Se as v.a.s são independentes, $X \perp\!\!\!\perp Y$, então conhecer X não nos diz nada sobre Y e vice-versa.
- Como temos uma medida de informação em uma fonte aleatória, $H(X)$, podemos quantificar quanta informação variáveis aleatórias possuem uma sobre as outras. Isto é chamado de <u>informação mútua</u>.

Informação Mútua de Evento

Dado o evento $\{X = x, Y = y\}$, podemos nos perguntar sobre qual é a informação fornecida pelo evento x dado o fato de que o evento y ocorreu. Isto pode ser quantificado da seguinte forma:

$$I(x;y) = \log \frac{p(x|y)}{p(x)} = \underbrace{\log \frac{1}{p(x)}}_{A} - \underbrace{\log \frac{1}{p(x|y)}}_{B} \tag{64}$$

- Primeiro termo A: surpresa de que x ocorreu.
- Segundo termo B: surpresa de que x ocorreu dado que y ocorreu.
- Diferença: diferença entre as duas surpresas, quanto mudou na surpresa de quando não sabíamos y para quando passamos a saber y.

Note que $p(x|x) = 1$, então $I(x;x) = \log 1/p(x) - \log 1 = \log 1/p(x) = I(x)$, então $I(x)$ pode ser visto como uma forma de 'auto-informação'.

Informação Mútua

Informação Mútua é a quantidade média de informação que uma variável aleatória X possui sobre outra v.a. Y e vice-versa.

Definição (Informação Mútua)

$$\begin{aligned} I(X;Y) &= E_{p(x,y)} \log \frac{p(x|y)}{p(x)} = E_{p(x,y)} \log \frac{p(x|y)p(y)}{p(x)p(y)} \\ &= E_{p(x,y)} \log \frac{p(x,y)}{p(x)p(y)} = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \end{aligned} \tag{65}$$

Informação Mútua e Entropia

Proposição

$$I(X;Y) = H(X) - H(X|Y) \tag{66}$$

Demonstração.

$$\begin{aligned} I(X;Y) &= E\log\frac{p(x|y)}{p(x)} \\ &= E\log\frac{1}{p(x)} - E\log\frac{1}{p(x|y)} \\ &= H(X) - H(X|Y) \end{aligned} \tag{67}$$

□

- Por simetria, temos que $I(X;Y) = H(Y) - H(Y|X)$.
- Como $H(X) \geq 0$ e $H(X|Y) \geq 0$, teremos $I(X;Y) \leq \min(H(X), H(Y))$.

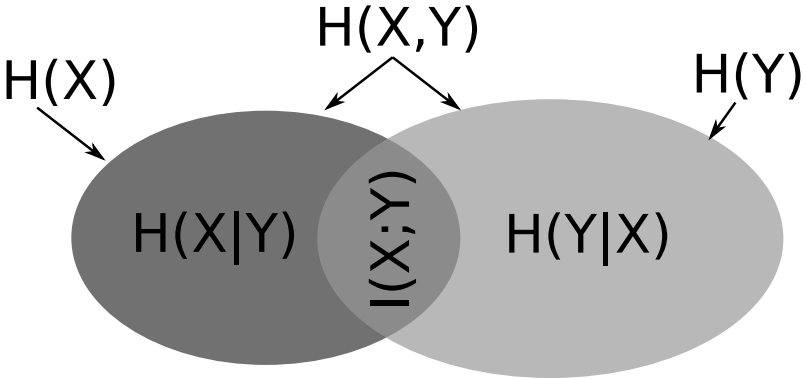
Informação Mútua e Entropia

- Regra da Cadeia da Entropia: $H(X,Y) = H(X) + H(Y|X)$.
- Informação Mútua: $I(X;Y) = H(X) - H(X|Y)$.
- Teremos então:

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{68}$$

No próximo slide representamos estas grandezas através de um diagrama. As áreas utilizadas não representam conjuntos no sentido comum, mas representam 'grau de informação' e as intersecções correspondem a sobreposição de informação. Isto é, a interseção consiste em informação fornecida por X e Y.

Informação Mútua e Entropia - Diagrama



Divergência de Kullbach-Leibler

A divergência de Kullbach-Leibler é uma relação fundamental entre duas distribuições probabilísticas sobre um mesmo alfabeto, $p = (p_1, \ldots, p_n)$ e $q = (q_1, \ldots, q_n)$. Esta divergência possui relação importante com a entropia e a informação mútua.

Como podemos medir a 'distância' entre duas distribuições p e q de forma útil? Poderíamos utilizar $D(p, q) = \sum_{i=1}^{n} (p_i - q_i)^2$, mas gostaríamos de ter uma medida de 'distância de informação', isto é, uma distância que nos dê o custo incorrido pelo erro de considerar que uma distribuição é q sendo que na realidade ela é p. Veremos que isto está ligado à insuficiência na compressão. A Divergência de Kullbach-Leibler, definida a seguir, satisfaz estas ideias.

Distância I

Definição (distância)

Seja S um conjunto. Uma função $d : S \times S \to \mathbb{R}$ é chamada **distância** em S se, para todo $x, y \in S$, tivermos:

- $d(x, y) \geq 0$ (não-negatividade)
- $d(x, y) = d(y, x)$ (simetria)
- $d(x, x) = 0$ (reflexividade)

Distância II

Definição (métrica)

Seja S um conjunto. Uma função $d : S \times S \to \mathbb{R}$ é chamada **métrica** em S se, para todo $x, y \in S$, tivermos:

- $d(x, y) \geq 0$ (não-negatividade)
- $d(x, y) = 0$ se e somente se $x = y$ (identidade dos indiscerníveis)
- $d(x, y) = d(y, x)$ (simetria)
- $d(x, y) + d(y, z) \geq d(x, z)$ (desigualdade triangular)

Distância II

Definição

Seja S um conjunto. Uma função $d : S \times S \to \mathbb{R}$ é chamada **métrica** em S se, para todo $x, y \in S$, tivermos:

- $d(x, y) \geq 0$ (não-negatividade)
- $d(x, y) = 0$ se e somente se $x = y$ (identidade dos indiscerníveis)
- $d(x, y) = d(y, x)$ (simetria)
- $d(x, y) + d(y, z) \geq d(x, z)$ (desigualdade triangular)

Teremos uma semi-métrica se substituirmos a identidade dos indiscerníveis pela reflexividade.

Diferentes formas de medir 'distância' entre distribuições

divergência de Kullback-Leibler: $D_{\mathrm{KL}}(p \parallel q) = \sum p(x) \ln\left(\frac{p(x)}{q(x)}\right)$;

distância de Hellinger: $H^2(p, q) = 2\sum\left(\sqrt{p(x)} - \sqrt{q(x)}\right)^2$;

divergência de Jeffreys: $D_J(p \parallel q) = \sum(p(x) - q(x))\big(\ln p(x) - \ln q(x)\big)$;

divergência α de Chernoff: $D^{(\alpha)}(p \parallel q) = \frac{4}{1-\alpha^2}\left(1 - \sum p(x)^{\frac{1-\alpha}{2}} q(x)^{\frac{1+\alpha}{2}}\right)$;

divergência exponencial: $D_e(p \parallel q) = \sum p(x)\big(\ln p(x) - \ln q(x)\big)^2$;

divergência de Kagan: $D_{\chi^2}(p \parallel q) = \frac{1}{2}\sum \frac{(p(x)-q(x))^2}{p(x)}$;

divergência K: $D_{\mathrm{K}}(p \parallel q) = \sum(p(x) - q(x)) \log(p(x)/q(x))$;

divergência de Jensen-Shannon: $D_{\mathrm{JS}}(p \parallel q) = \frac{1}{2}D_{\mathrm{KL}}(p \parallel m) + \frac{1}{2}D_{\mathrm{KL}}(q \parallel m)$, onde $m = \frac{1}{2}(p+q)$.

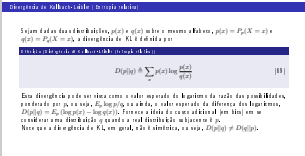
Divergência de Kullbach-Leibler (Entropia relativa)

Sejam dadas duas distribuições, $p(x)$ e $q(x)$ sobre o mesmo alfabeto, $p(x) = P_p(X = x)$ e $q(x) = P_q(X = x)$, a divergência de KL é definida por

Definição (Divergência de Kullbach-Leibler (entropia relativa))

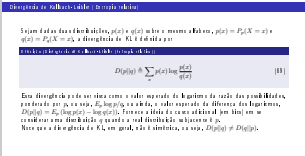
$$D(p||q) \triangleq \sum_x p(x) \log \frac{p(x)}{q(x)} \tag{69}$$

Esta divergência pode ser vista como o valor esperado do logaritmo da razão das possibilidades, ponderado por p, ou seja, $E_p \log p/q$, ou ainda, o valor esperado da diferença dos logaritmos, $D(p||q) = E_p (\log p(x) - \log q(x))$. Fornece a ideia do custo adicional (em bits) em se considerar uma distribuição q quando a real distribuição subjacente é p.
Note que a divergência de KL, em geral, não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.

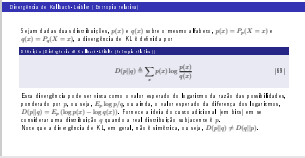


Utilizando argumentos de limite e continuidade, mostra-se que $0 \log 0 = 0$ e $p \log(p/0) = \infty$. Fazendo estas suposições, teremos $D(p||q) \leq \infty$.
A divergência de KL é uma função dos valores de probabilidade e não dos valores que a variável aleatória assume (assim como a entropia e a informação mútua).

A razão de chances ou razão de possibilidades (em inglês: *odds ratio*) é definida como a razão entre a chance de um evento ocorrer em um grupo e a chance de ocorrer em outro grupo.



(Wikipedia) In statistics, the odds ratio (usually abbreviated "OR") is one of three main ways to quantify how strongly the presence or absence of property A is associated with the presence or absence of property B in a given population. If each individual in a population either does or does not have a property "A", (e.g. "high blood pressure"), and also either does or does not have a property "B"(e.g. "moderate alcohol consumption") where both properties are appropriately defined, then a ratio can be formed which quantitatively describes the association between the presence/absence of "A"(high blood pressure) and the presence/absence of "B"(moderate alcohol consumption) for individuals in the population.



Divergência de Kullbach-Leibler | Entropia relativa

Sejam duas distribuições, $p(x)$ e $q(x)$ sobre o mesmo alfabeto, $p(x) = P_p(X = x)$ e $q(x) = P_q(X = x)$, a divergência de K.L é dada por

Divergência de Kullbach-Leibler (Entropia relativa)

$$D(p||q) \triangleq \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)} \qquad (15)$$

Essa divergência pode ser vista como o valor esperado do logaritmo da razão das probabilidades, ponderado por p, ou seja, $E_p \log p/q$, ou ainda, o valor esperado da diferença dos logaritmos, $D(p||q) = E_p(\log p(x) - \log q(x))$. Tomemos a média da como adicional (em bits) em se considerar uma distribuição q quando a real distribuição subjacente é p.
Note que a divergência de K.L, em geral, não é simétrica, ou seja, $D(p||q) \neq D(q||p)$.

This ratio is the odds ratio (OR) and can be computed following these steps:

1. For a given individual that has "B"compute the odds that the same individual has "A"
2. For a given individual that does not have "B"compute the odds that the same individual has "A"
3. Divide the odds from step 1 by the odds from step 2 to obtain the odds ratio (OR). The term "individual"in this usage does not have to refer to a human being, as a statistical population can measure any set of entities, whether living or inanimate.

`http://en.wikipedia.org/wiki/Odds_ratio`

Exemplo

Seja $\mathcal{X} = \{1, 0\}$ e considere duas distribuições p e q em $\mathcal{X}$. Seja $p(0) = 1 - r$, $p(1) = r$, e seja $q(0) = 1 - s$ e $q(1) = s$. Então

$$D(p||q) = (1 - r) \log \frac{1 - r}{1 - s} + r \log \frac{r}{s} \tag{70}$$

e

$$D(q||p) = (1 - s) \log \frac{1 - s}{1 - r} + s \log \frac{s}{r}. \tag{71}$$

Se $r = s$, então $D(p||q) = D(q||p) = 0$. Se $r = \frac{1}{2}$ e $s = \frac{1}{4}$,

$$D(p||q) = \frac{1}{2} \log \frac{1/2}{3/4} + \frac{1}{2} \log \frac{1/2}{1/4} = 1 - \frac{1}{2} \log 3 = 0.2075\text{bits}. \tag{72}$$

$$D(q||p) = \frac{3}{4} \log \frac{3/4}{1/2} + \frac{1}{4} \log \frac{1/4}{1/2} = \frac{3}{4} \log 3 - 1 = 0.1887\text{bits}. \tag{73}$$

Note que, em geral, $D(p||q) \neq D(q||p)$.

Generalização da divergência de KL

A divergência de KL pode ser generalizada para vetores de variáveis aleatórias.
Seja $p(x_1, \ldots, x_N)$ e $q(x_1, \ldots, x_N)$ duas distribuições sobre o vetor $(x_1, x_2, \ldots, x_N)$. A divergência de KL entre p e q é definida por

$$D(p||q) = \sum_{x_1,\ldots,x_N} p(x_1, \ldots, x_N) \log \frac{p(x_1, \ldots, x_N)}{q(x_1, \ldots, x_N)} \tag{74}$$

Divergência de KL e Informação Mútua I

Seja $\mu_1(x,y) = p(x,y)$ (distribuição conjunta) e $\mu_2(x,y) = p(x)p(y)$ (produto das marginais) com $p(x) = \sum_y p(x,y)$ e $p(y) = \sum_x p(x,y)$, então

$$\begin{aligned} D(\mu_1||\mu_2) &= \sum_{x,y} \mu_1(x,y) \log \frac{\mu_1(x,y)}{\mu_2(x,y)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = I(X;Y). \end{aligned} \tag{75}$$

A informação mútua é a distância entre a distribuição conjunta em X e Y e o produto das distribuições marginais em X e Y.
Se as v.a.s são independentes, teremos $p(x,y) = p(x)p(y)$ e por conseguinte a divergência será nula, a informação mútua entre X e Y será zero.
A informação mútua é o erro em se assumir independência entre as v.a.s.

Divergência de KL e Informação Mútua II

O produto das distribuições marginais $p(x)p(y)$, onde $p(x) = \sum_y p(x,y)$ e $p(y) = \sum_x p(x,y)$, é uma projeção da distribuição conjunta $p(x,y)$ sobre o conjunto das distribuições independentes. I.e.,

$$p(x)p(y) = \underset{p'(x,y)\setminus p'(x,y)=p'(x)p'(y)}{\operatorname{argmin}} D(p(x,y)||p'(x,y)) \tag{76}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança I

Suponha que tenhamos uma v.a. $\mathbf{X} = (X_1, \ldots, X_N)$ com uma distribuição subjacente p que depende de um parâmetro θ (modelo hipotético). Queremos definir um estimador $\hat{\theta} = T(X_1, \ldots, X_N)$ para este parâmetro θ, dadas as observações $x_1, \ldots, x_N$. Um bom estimador para o parâmetro desconhecido θ é aquela que maximiza a verossimilhança $L(\theta)$ do parâmetro, dada a observação dos dados,

$$L(\theta) = \Pr(X_1 = x_1, \ldots, X_N = x_N) = p(x_1|\theta) \ldots p(x_N|\theta) = \prod_{n=1}^{N} p(x_n|\theta). \tag{77}$$

Como a função logaritmo é monotônica crescente, maximizar $L(\theta)$ é equivalente a maximiza $l(\theta) = \log L(\theta)$,

$$\ell(\theta) = \sum_{n=1}^{N} \log p(x_n|\theta). \tag{78}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança II

A estimativa de máxima verossimilhança (MLE, *maximum likelihood estimator*) de θ é dada por

$$\hat{\theta}_{\text{mle}} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, \ell(\theta\,;\, x_1, \ldots, x_N). \tag{79}$$

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança III

Chamaremos de $\hat{p}$ a distribuição empírica. Seja $x_1, \ldots, x_N \in \mathcal{X}$, N observações i.i.d. de uma variável aleatória X. A distribuição empírica será dada por

$$\hat{p}(x) = \frac{1}{N} \sum_{n=1}^{N} \delta(x - x_n), \tag{80}$$

onde δ é a função de Dirac.

Seja p_θ uma distribuição em $\mathcal{X}$ parametrizada por θ. Maximizar a verossimilhança de $p_\theta(x)$ é equivalente a minimizar a divergência de KL $D_{\mathrm{KL}}(\hat{p} \parallel p_\theta)$.

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança IV

$$
\begin{aligned}
D_{\mathrm{KL}}(\hat{p} \parallel p_\theta) &= \sum_{x \in \mathcal{X}} \hat{p}(x) \log \frac{\hat{p}(x)}{p_\theta(x)} \\
&= -H(\hat{p}) - \sum_{x \in \mathcal{X}} \hat{p}(x) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{x \in \mathcal{X}} \sum_{n=1}^{N} \delta(x - x_n) \log p_\theta(x) \\
&= -H(\hat{p}) - \frac{1}{N} \sum_{n=1}^{N} \log p_\theta(x_n). \qquad (81)
\end{aligned}
$$

O segundo termo é o oposto do logaritmo da verossimilhança de $p_\theta(x)$.

Minimizar a Divergência de KL equivale a maximizar o logaritmo da verossimilhança V

A estimativa máxima verossimilhança de θ a partir das N observações é dada por

$$\begin{aligned}\hat{\theta}_n &= \underset{\theta\in\Theta}{\operatorname{argmax}} \prod_{n=1}^{N} p_\theta(x_n) \\ &= \underset{\theta\in\Theta}{\operatorname{argmax}} \sum_{n=1}^{N} \log p_\theta(x_n) \\ &= \underset{\theta\in\Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{n=1}^{N} -\log p_\theta(x_n). \end{aligned} \tag{82}$$

Desta forma, podemos constatar que a distribuição que minimiza a divergência de KL para a distribuição empírica é aquela que maximiza a verossimilhança (ou logaritmo desta).

Informação Mútua Condicionada a um evento

A informação pode se alterar se for condicionada a um evento de uma terceira variável aleatória $\{Z = z\}$, e isto é denotado por $I(X;Y|Z = z)$, onde X, Y, Z são variáveis aleatórias. Dada a distribuição $p(x, y, z)$, a informação mútua condicionada ao evento específico $\{Z = z\}$ é dada por

$$I(X;Y|Z = z) = \sum_{x,y} p(x, y|z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)}. \tag{83}$$

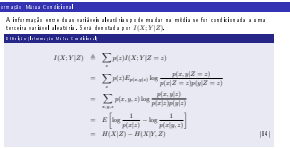
Obs. Fazemos as seguintes alterações sobre a informação mútua padrão: $p(x, y) \to p(x, y|z)$, $p(x) \to p(x|z)$ e $p(y) \to p(y|z)$.

Informação Mútua Condicional

A informação entre duas variáveis aleatórias pode mudar na média se for condicionada a uma terceira variável aleatória. Será denotada por $I(X;Y|Z)$.

Definição (Informação Mútua Condicional)

$$\begin{aligned} I(X;Y|Z) &\triangleq \sum_z p(z) I(X;Y|Z=z) \\ &= \sum_z p(z) E_{p(x,y|z)} \log \frac{p(x,y|Z=z)}{p(x|Z=z)p(y|Z=z)} \\ &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} \\ &= E\left[\log \frac{1}{p(x|z)} - \log \frac{1}{p(x|y,z)}\right] \\ &= H(X|Z) - H(X|Y,Z) \end{aligned} \tag{84}$$



$$I(X;Y) = H(X) - H(X|Y) \tag{85}$$

$$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \tag{86}$$

Regra da Cadeia para Informação Mútua

Proposição

$$I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y | X_1, X_2, \ldots, X_{i-1}) \tag{87}$$

Exemplo: $I(X_1, X_2; Y) = I(X_1; Y) + I(X_2; Y | X_1)$

Demonstração.

$$\begin{aligned} I(X_1, \ldots, X_N; Y) &= H(X_1, \ldots, X_N) - H(X_1, \ldots, X_N | Y) \\ &= \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}) - \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}, Y) \\ &= \sum_{i=1}^{N} I(X_i; Y | X_1, \ldots, X_{i-1}) \end{aligned} \tag{88}$$

□

Entropia Relativa Condicional - divergência de KL

Definição

Para pmf conjuntas $p(x,y)$ e $q(x,y)$, a entropia relativa condicional é definida como

$$\begin{aligned} D(p(y|x)||q(y|x)) &\triangleq \sum_x p(x) \sum_y p(y|x) \log \frac{p(y|x)}{q(y|x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} \\ &= \mathrm{E}_{p(x,y)} \log \frac{p(Y|X)}{q(Y|X)}, \end{aligned} \tag{89}$$

é o valor esperado das entropias relativas entre as pmfs condicionais $p(y|x)$ e $q(y|x)$, tomando o valor esperado sobre a distribuição de massa $p(x)$.

Regra da Cadeia para divergência de KL

Proposição

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x)) \tag{90}$$

Demonstração.

$$\begin{aligned} D(p(x,y)||q(x,y)) &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{q(x,y)} = \sum_{x,y} p(x,y) \log \frac{p(y|x)p(x)}{q(y|x)q(x)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)} + \sum_{x,y} p(x,y) \log \frac{p(x)}{q(x)} \end{aligned} \tag{91}$$

□

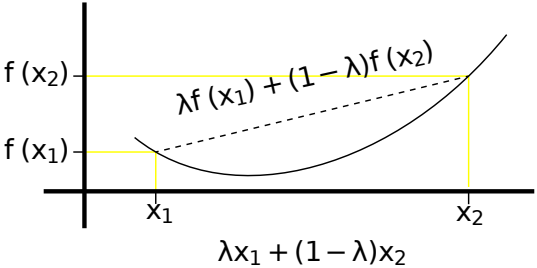
Funções Convexas

Definição

Dizemos que f é convexa em (a, b) se para todo $x_1, x_2 \in (a, b)$, $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \tag{92}$$

Exemplos: 1) $f(x) = x^2$ 2) $f(x) = e^x$ 3) $x \log x$, $x \geq 0$.



- f é estritamente convexa se a igualdade for verdadeira apenas para $\lambda = 0$ ou $\lambda = 1$.

Derivada Segunda e Convexidade I

Teorema (derivada segunda e convexidade)

Se uma função f possui derivada segunda não-negativa (positiva) em um intervalo, a função é convexa (estritamente convexa) no intervalo.

Demonstração.

A expansão de Taylor de uma função f em torno do ponto x_0 é dada por

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{93}$$

onde $x^* \in (x_0, x)$. Por hipótese, $f''(x^*) \geq 0$, e desta forma, o último termo é não-negativo.

...

Derivada Segunda e Convexidade II

Demonstração.

continuação...

Seja $x_0 = \lambda x_1 + (1-\lambda)x_2$. Analisando em $x = x_1$, teremos

$$\begin{aligned} f(x_1) &\geq f(x_0) + f'(x_0)(x_1 - \lambda x_1 - (1-\lambda)x_2) \\ &= f(x_0) + f'(x_0)((1-\lambda)(x_1 - x_2)) \end{aligned} \tag{94}$$

Da mesma forma, em $x = x_2$, teremos

$$\begin{aligned} f(x_2) &\geq f(x_0) + f'(x_0)(x_2 - \lambda x_1 - (1-\lambda)x_2) \\ &= f(x_0) + f'(x_0)(\lambda(x_2 - x_1)) \end{aligned} \tag{95}$$

...

Derivada Segunda e Convexidade III

Demonstração.

continuação...

Somando λ 94 com $(1-\lambda)$ 95, teremos

$$\begin{aligned} \lambda f(x_1) + (1-\lambda) f(x_2) &\geq \lambda f(x_0) + \lambda f'(x_0)((1-\lambda)(x_1 - x_2)) + \\ &\quad (1-\lambda) f(x_0) + (1-\lambda) f'(x_0)(\lambda(x_2 - x_1)) \\ &\geq f(x_0) = f(\lambda x_1 + (1-\lambda) x_2) \end{aligned} \tag{96}$$

□

Desigualdade de Jensen

Teorema (Jensen)

Seja f uma função convexa e X uma variável aleatória, então

$$Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f\left(\sum_x xp(x)\right) \tag{97}$$

Se f é estritamente convexa, então $\{Ef(X) = f(E(X))\} \Rightarrow \{X = EX\}$, o que significa que X é uma v.a. constante.

Desigualdade de Jensen - demonstração I

- Para uma distribuição de massa com apenas dois pontos.

$$E[f(X)] = p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2) = f(EX) \tag{98}$$

  já que f é convexa e $p_1 + p_2 = 1$.
- Para uma distribuição com mais de dois ponto, iremos fazer uma demonstração por indução.

Demonstração.

Suponha que o teorema seja verdadeiro para uma distribuição com $k-1$ pontos de massa. Para uma distribuição com k pontos de massa podemos escrever cada $p_i' = p_i/(1-p_k)$ para $i = 1, 2, \ldots, k-1$. ...

Desigualdade de Jensen - demonstração II

Demonstração.

continuação...

Desta forma, teremos

$$
\begin{aligned}
E[f(X)] &= \sum_{i=1}^{k} p_i f(x_i) \\
&= \sum_{i=1}^{k-1} (1-p_k) p'_i f(x_i) + p_k f(x_k) \\
&= (1-p_k) \sum_{i=1}^{k-1} p'_i f(x_i) + p_k f(x_k)
\end{aligned}
$$

...

Desigualdade de Jensen - demonstração III

Demonstração.

continuação...

Poderemos utilizar a hipótese de indução, já que

$$\sum_{i=1}^{k-1} p'_i = \sum_{i=1}^{k-1} \frac{p_i}{1-p_k} = \frac{1-p_k}{1-p_k} = 1. \tag{99}$$

Então

$$\begin{aligned} E[f(X)] &= \dots \\ &\geq (1-p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right) + p_k f(x_k) \end{aligned}$$

...

Desigualdade de Jensen - demonstração IV

Demonstração.

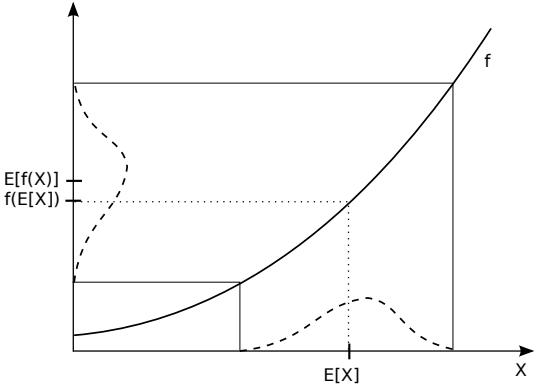
continuação...

Pela definição de convexidade, teremos

$$\begin{aligned} E[f(X)] &= \dots \\ &\geq f\left((1-p_k)\sum_{i=1}^{k-1} p_i' x_i + p_k x_k\right) \\ &= f\left(\sum_{i=1}^{k} p_i x_i\right) \end{aligned} \tag{100}$$

Desta forma, sendo o teorema válido para uma distribuição de massa com $k-1$ pontos, também será verdadeiro para uma distribuição de massa com k pontos. Como mostramos que para $k=2$ é verdadeiro, logo o teorema é verdadeiro para qualquer k. □

Desigualdade de Jensen - demonstração gráfica



O mapeamento feito pela função convexa f aumenta gradativamente o estiramento da distribuição mapeada por f com o aumento dos valores de X. Desta forma, o valor esperado da distribuição mapeada por f tende a possuir um valor maior que o mapeamento por f do valor esperado da distribuição.

A divergência de KL é não-negativa

Lema

$$D(p||q) \geq 0 \text{ com igualdade se e somente se } p(x) = q(x) \forall x. \tag{101}$$

A divergência de KL é não-negativa

Demonstração.

Mostre que $-D(p||q) \leq 0$. Seja $S_p = \{x : p(x) > 0\} = \sup(p)$, então

$$\begin{aligned} -D(p||q) &= -\sum_x p(x) \log \frac{p(x)}{q(x)} = -\sum_{x \in S_p} p(x) \log \frac{p(x)}{q(x)} = \sum_{x \in S_p} p(x) \log \frac{q(x)}{p(x)} = E \log \frac{q(X)}{p(X)} \\ &\text{utilizando a desigualdade de Jensen} \\ &\leq \log \left(E \frac{q(X)}{p(X)} \right) = \log \left(\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} \right) \\ &= \log \left(\sum_{x \in S_p} q(x) \right) \leq \log \left(\sum_x q(x) \right) = \log 1 = 0 \end{aligned} \tag{102}$$

□

A divergência de KL é não-negativa

- Note que $\log x$ é estritamente côncavo.
- Então, a igualdade $\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} = \log \left(\sum_{x \in S_p} p(x) \frac{q(x)}{p(x)} \right)$ significa $Z = EZ$ com $Z = q(X)/p(X)$, então Z é uma variável aleatória constante.
- A única constante válida, com p e q sendo distribuições de probabilidade é $Z = 1$ ou $p(x) = q(x)$.
- Então, se $p(x) = q(x)$ teremos $D(p||q) = 0$ e vice-versa.

A informação mútua é não-negativa

Proposição

$$I(X;Y) \geq 0 \text{ e } I(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y. \tag{103}$$

Demonstração.

$$I(X;Y) = D(p(x,y)||p(x)p(y)) \geq 0 \tag{104}$$

□

teremos igualdade se $p(x,y) = p(x)p(y)$, o que é também condição para independência.

Informação Mútua

- $I(X;Y)$ mede o 'grau de dependência' entre X e Y.
- Temos $0 \leq I(X;Y) \leq \min(H(X), H(Y))$.
- $I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$.
- Se $X \perp\!\!\!\perp Y$, então $I(X;Y) = 0$, já que em tal caso $H(X|Y) = H(X)$ e $H(Y|X) = H(Y)$.
- Se $X = Y$, então $I(X;Y) = H(X) = H(Y)$ já que em tal caso $H(Y|X) = H(X|Y) = 0$.

Limite Superior para a Entropia I

Teorema

$H(X) \leq \log |\mathcal{X}|$, onde $|\mathcal{X}|$ denota o número de elementos da extensão de X (a cardinalidade do domínio), com igualdade se e somente se X possuir distribuição uniforme.

Limite Superior para a Entropia II

Demonstração.

Seja $u(x) = \frac{1}{|\mathcal{X}|}$ a função probabilidade de massa uniforme em $\mathcal{X}$, e seja $p(x)$ a função probabilidade de massa para X. Então

$$\begin{aligned} D(p||u) &= \sum_{x\in\mathcal{X}} p(x)\log\frac{p(x)}{u(x)} \\ &= \sum_{x\in\mathcal{X}} p(x)\log p(x) + \sum_{x\in\mathcal{X}} p(x)\log\frac{1}{u(x)} \\ &= -H(X) + \log|\mathcal{X}|\sum_{x\in\mathcal{X}} p(x) \\ &= \log|\mathcal{X}| - H(X) \end{aligned} \tag{105}$$

...

Limite Superior para a Entropia III

Demonstração.

continuação...

Como a entropia relativa é não negativa, $D(p||u) \geq 0$, teremos

$$D(p||u) = \log|\mathcal{X}| - H(X) \geq 0 \tag{106}$$

e assim

$$H(X) \leq \log|\mathcal{X}| \tag{107}$$

□

Condicionar Reduz Entropia

Comparando $H(X)$ com $H(X|Y)$, conhecendo Y, na média, pode nos dizer algo sobre X reduzindo a entropia.

Proposição

$$H(X|Y) \leq H(X) \text{ e } H(X|Y) = H(X) \text{ se e somente se } X \perp\!\!\!\perp Y. \tag{108}$$

Demonstração.

$$0 \leq I(X;Y) = H(X) - H(X|Y) \tag{109}$$

□

Poderíamos ter $H(X|Y=y) > H(X)$, mas, na média, $\sum_y p(y)H(X|Y=y) \leq H(X)$.

Limite da Entropia para um Conjunto de V.A.

A entropia de um conjunto de variáveis aleatórias é maior quando as variáveis aleatórias são independentes, há menor redundância entre elas.

Proposição

$$H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \tag{110}$$

Demonstração.

$$H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i | X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \tag{111}$$

□

Limite da Entropia para um Conjunto de V.A.

A entropia de um conjunto de variáveis aleatórias é maior quando as variáveis aleatórias são independentes, há menor redundância entre elas.

Proposição

$$H(X_1, X_2, \ldots, X_N) \leq \sum_{i=1}^{N} H(X_i) \tag{110}$$

Demonstração

$$H(X_1, \ldots, X_N) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}) \leq \sum_{i=1}^{N} H(X_i) \tag{111}$$

$$\sum_{i=1}^{N} H(X_i|X_{\neg i}) = \sum_{i=1}^{N} H(X_i|X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N) \leq H(X_1, \ldots, X_N) \tag{112}$$

Limites da Independência na Entropia

A proposição 6 para duas variáveis é da forma

$$H(X_1, X_2) \leq H(X_1) + H(X_2) \tag{113}$$

Note que a igualdade na Equação 110 é alcançada quando todas as variáveis são mutuamente independentes, isto é, quando $X_i \perp\!\!\!\perp X_j \forall i, j$.

Condicionamento e Informação Mútua

- Se $X \perp\!\!\!\perp Y|Z$ então $I(X;Y|Z) = 0$. Por exemplo, $X \perp\!\!\!\perp Y|Z$ quando $X \to Z \to Y$.
- Alternativamente, se $Z = Y$, então $I(X;Y|Z) = 0$.
- Podemos ter $I(X;Y) > I(X;Y|Z)$.
- Por outro lado, se $Z = X + Y$ e $X \perp\!\!\!\perp Y$, então $I(X;Y) = 0$ mas $I(X;Y|Z) > 0$.
- Não existe uma relação genérica entre informação mútua e informação mútua condicional.

Relações de H

$$H(X) = EI(X) = -\sum_{x} p(x) \log p(x) \tag{114}$$

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) \tag{115}$$

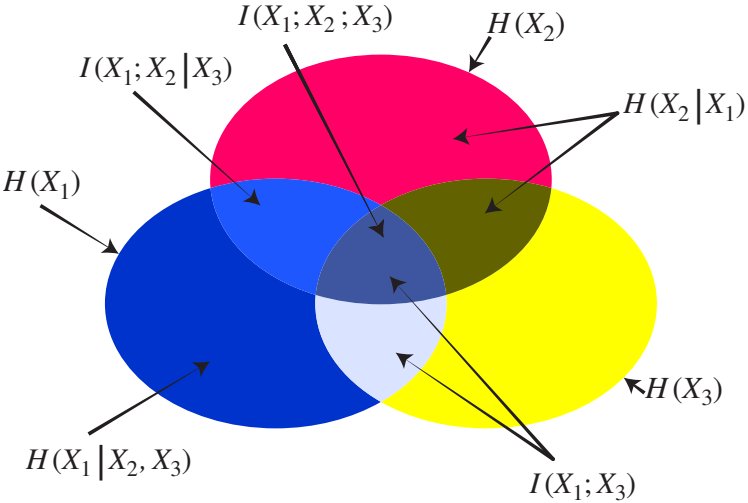
$$H(Y|X) = -\sum_{x,y} p(x,y) \log p(y|x) \tag{116}$$

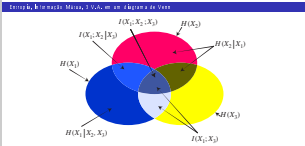
$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y) \tag{117}$$

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{118}$$

$$0 \leq H(X) \leq \log n, \text{ onde } n \text{ é o tamanho do alfabeto de } X. \tag{119}$$

Entropia, Informação Mútua, 3 V.A. em um diagrama de Venn





- $I(X_1;X_2) = I(X_1;X_2|X_3) + I(X_1;X_2;X_3)$.
- $I(X_1;X_2) \gtreqless I(X_1;X_2|X_3)$, mas isto nunca será negativo.
- Então, $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3)$ pode ser negativo.
- $I(X_1;X_2;X_3) = I(X_1;X_2) - I(X_1;X_2|X_3) = I(X_2;X_3) - I(X_2;X_2|X_1) = I(X_3;X_1) - I(X_3;X_1|X_2)$
- $I(X_1;X_2;X_3) = H(X_1) + H(X_2) + H(X_3) - H(X_1;X_2) - H(X_2;X_3) - H(X_3;X_1) + H(X_1,X_2,X_3)$

Revisão I

- divergência de KL: $D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$
- informação mútua: $I(X;Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = D(p(x,y)||p(x)p(y))$
- informação mútua condicional: $I(X;Y|Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)} = H(X|Z) - H(X|Y,Z)$
- regra da cadeia da informação mútua: $I(X_1, X_2, \ldots, X_N; Y) = \sum_i I(X_i; Y|X_1, X_2, \ldots, X_{i-1})$
- entropia relativa condicional: $D(p(y|x)||q(y|x)) \triangleq \sum_{x,y} p(x,y) \log \frac{p(y|x)}{q(y|x)}$
- regra da cadeia da divergência de KL: $D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$
- Jensen: f convexa $\Rightarrow Ef(X) = \sum_x p(x)f(x) \geq f(EX) = f\left(\sum_x px(x)\right)$
- não negatividade da divergência de KL: $D(p||q) \geq 0$, $D(p||q) = 0 \Leftrightarrow p = q$.
- não negatividade da informação mútua: $I(X;Y) \geq 0$, $I(X;Y) = 0 \Leftrightarrow X \perp\!\!\!\perp Y$.

Revisão II

- condicionar reduz a entropia: $H(X) \geq H(X|Y)$, $H(X) = H(X|Y) \Leftrightarrow X \perp\!\!\!\perp Y$.
- limite da independência em H: $H(X_1, \ldots, X_N) \leq \sum_i H(X_i)$, com igualdade sse todos X_i forem independentes

Medida de Informação

Veremos as correspondências entre a medida de informação de Shannon (e suas manipulações) com a teoria de conjuntos. A utilização dos diagramas de informação podem ser utilizadas para simplificar várias demonstrações em teoria da informação.

Medida de Informação

- Temos um conjunto de variáveis aleatórias: $X_1, X_2, \ldots, X_n$.
- Para cada variável aleatória associamos um conjunto $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

Definição (campo (field))

Um campo $\mathcal{F}_n$ gerado pelos conjuntos $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$ é a coleção de conjuntos que podem ser obtidos através de qualquer sequência de operações usuais de conjuntos (união, interseção, complemento, e diferença) sobre os conjuntos $\tilde{X}_1, \tilde{X}_2, \ldots, \tilde{X}_n$.

Definição (átomo)

Os átomos de $\mathcal{F}_n$ são os conjuntos da forma $\cap_{i=1}^{n} Y_i$, onde Y_i é $\tilde{X}_i$ ou $\tilde{X}_i^c$, o complemento de $\tilde{X}_i$.

Átomos

átomos - $n = 2$

Para $n = 2$, teremos os conjuntos $\tilde{X}_1, \tilde{X}_2$ e seus complementos, respectivamente, $\tilde{X}_1^c, \tilde{X}_2^c$. Existirão 4 átomos:

1) $\tilde{X}_1 \cap \tilde{X}_2$,
2) $\tilde{X}_1 \cap \tilde{X}_2^c$,
3) $\tilde{X}_1^c \cap \tilde{X}_2$, e
4) $\tilde{X}_1^c \cap \tilde{X}_2^c$

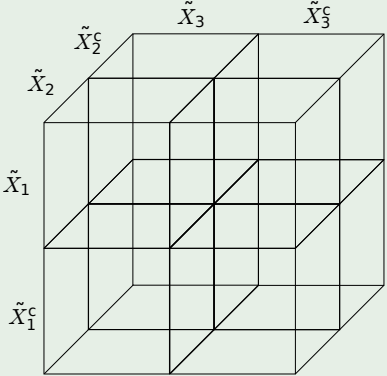
	$\tilde{X}_2$	$\tilde{X}_2^c$
$\tilde{X}_1$	$\tilde{X}_1 \cap \tilde{X}_2$	$\tilde{X}_1 \cap \tilde{X}_2^c$
$\tilde{X}_1^c$	$\tilde{X}_1^c \cap \tilde{X}_2$	$\tilde{X}_1^c \cap \tilde{X}_2^c$

Átomos

átomos - $n = 3$

Para $n = 3$, teremos os conjuntos $\tilde{X}_1, \tilde{X}_2, \tilde{X}_3$ e seus complementos, respectivamente, $\tilde{X}_1^c, \tilde{X}_2^c, \tilde{X}_3^c$. Existirão 8 átomos:

1) $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3$,
2) $\tilde{X}_1 \cap \tilde{X}_2 \cap \tilde{X}_3^c$,
3) $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3$,
4) $\tilde{X}_1 \cap \tilde{X}_2^c \cap \tilde{X}_3^c$,
5) $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3$,
6) $\tilde{X}_1^c \cap \tilde{X}_2 \cap \tilde{X}_3^c$,
7) $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3$, e
8) $\tilde{X}_1^c \cap \tilde{X}_2^c \cap \tilde{X}_3^c$



Campo e Átomos

- existem 2^n átomos;
- existem 2^{2^n} conjuntos no campo $\mathcal{F}_n$;
- todos os átomos em $\mathcal{F}_n$ são disjuntos;
- todo conjunto em $\mathcal{F}_n$ pode ser expresso de forma única como uma união de um subconjunto dos átomos em $\mathcal{F}_n$.

Medida com sinal

Em análise matemática, uma medida em um conjunto S é uma forma sistemática de atribuir números a todo subconjunto de S, sendo intuitivamente interpretada como o seu tamanho. Medida com sinal é uma generalização do conceito de medida permitindo que esta assuma valores negativos.

Definição (medida com sinal)

Uma função real μ definida em $\mathcal{F}_n$ é chamada medida com sinal se for aditiva no conjunto, i.e., para A e B disjuntos em $\mathcal{F}_n$,

$$\mu(A \cup B) = \mu(A) + \mu(B). \tag{120}$$

Para uma medida com sinal μ teremos $\mu(\emptyset) = 0$, já que $\mu(A) = \mu(A \cup \emptyset) = \mu(A) + \mu(\emptyset)$.

Medida com sinal

Uma medida com sinal μ em $\mathcal{F}_n$ é completamente especificada por seus valores nos átomos de $\mathcal{F}_n$. Os valores de μ em outros conjuntos de $\mathcal{F}_n$ podem ser obtidos pela aditividade de conjuntos, já que qualquer $\tilde{X} \in \mathcal{F}_n$ pode ser representado como $\tilde{X} = \cup_{i=1} Y_i$, onde Y_i são átomos escolhidos apropriadamente.

Medida com sinal

$n = 2$

Uma medida com sinal μ em $\mathcal{F}_2$ é completamente especificada pelos valores $\mu(\tilde{X}_1 \cap \tilde{X}_2)$, $\mu(\tilde{X}_1 \cap \tilde{X}_2^c)$, $\mu(\tilde{X}_1^c \cap \tilde{X}_2)$, e $\mu(\tilde{X}_1^c \cap \tilde{X}_2^c)$.
O valor de μ em $\tilde{X}_1$ pode ser obtido da seguinte forma

$$\begin{aligned} \mu(\tilde{X}_1) &= \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \\ &= \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c). \end{aligned} \tag{121}$$

O valor de μ em $\tilde{X}_1 \setminus \tilde{X}_2$ é dado por

$$\mu(\tilde{X}_1 \setminus \tilde{X}_2) = \mu(\tilde{X}_1 \cap \tilde{X}_2^c). \tag{122}$$

O valor de μ em $\tilde{X}_1 \cup \tilde{X}_2$ pode ser obtido através de

$$\begin{aligned} \mu(\tilde{X}_1 \cup \tilde{X}_2) &= \mu((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c) \cup (\tilde{X}_1^c \cap \tilde{X}_2)) \\ &= \mu(\tilde{X}_1 \cap \tilde{X}_2) + \mu(\tilde{X}_1 \cap \tilde{X}_2^c) + \mu(\tilde{X}_1^c \cap \tilde{X}_2) \end{aligned} \tag{123}$$

Correspondência com a informação de Shannon I

Os conjuntos $\tilde{X}_1$ e $\tilde{X}_2$ estão associados às variáveis aleatória X_1 e X_2. O campo $\mathcal{F}_2$ é gerado por $\tilde{X}_1$ e $\tilde{X}_2$, através dos átomos $(\tilde{X}_1 \cap \tilde{X}_2)$, $(\tilde{X}_1 \cap \tilde{X}_2^c)$, $(\tilde{X}_1^c \cap \tilde{X}_2)$, e $(\tilde{X}_1^c \cap \tilde{X}_2^c)$. O diagrama de informação é apresentado na Figura 4.

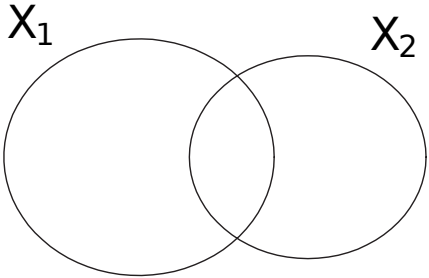


Figura 4: Diagrama de informação para X_1 e X_2.

Correspondência com a informação de Shannon II

O conjunto universo será considerada como sendo $\Omega = \tilde{X}_1 \cup \tilde{X}_2$. Desta forma, o átomo $\tilde{X}_1^c \cap \tilde{X}_2^c$ se degenera ao conjunto vazio,

$$\tilde{X}_1^c \cap \tilde{X}_2^c = (\tilde{X}_1 \cup \tilde{X}_2)^c = \Omega^c = \emptyset. \tag{124}$$

Para as v.a.s X_1 e X_2, as medidas de informação de Shannon são

$$H(X_1), H(X_2), H(X_1|X_2), H(X_2|X_1), H(X_1, X_2), I(X_1; X_2). \tag{125}$$

Utilizando a notação $A \cap B^c \equiv A \setminus B$, definimos uma medida com sinal μ^*

$$\mu^*(\tilde{X}_1 \setminus \tilde{X}_2) = H(X_1|X_2), \tag{126}$$
$$\mu^*(\tilde{X}_2 \setminus \tilde{X}_1) = H(X_2|X_1), \tag{127}$$
$$\mu^*(\tilde{X}_1 \cap \tilde{X}_2) = I(X_1; X_2). \tag{128}$$

Correspondência com a informação de Shannon III

Estes são os valores de μ^* nos átomos não vazios de $\mathcal{F}_2$. Os valores de μ^* nos demais conjuntos de $\mathcal{F}_2$ podem ser obtidos por adição de conjuntos. Em particular, temos as relações

$$\begin{aligned} \mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= H(X_1, X_2), && (129) \\ \mu^*(\tilde{X}_1) &= H(X_1), && (130) \\ \mu^*(\tilde{X}_2) &= H(X_2). && (131) \end{aligned}$$

Por exemplo, a Equação 129 pode ser verificada

$$\begin{aligned} \mu^*(\tilde{X}_1 \cup \tilde{X}_2) &= \mu^*((\tilde{X}_1 \setminus \tilde{X}_2) \cup (\tilde{X}_2 \setminus \tilde{X}_1) \cup (\tilde{X}_1 \cap \tilde{X}_2)) \\ &= \mu^*(\tilde{X}_1 \setminus \tilde{X}_2) + \mu^*(\tilde{X}_2 \setminus \tilde{X}_1) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2) \\ &= H(X_1|X_2) + H(X_2|X_1) + I(X_1; X_2) \\ &= H(X_1, X_2). && (132) \end{aligned}$$

Correspondência com a informação de Shannon IV

A Equação 130 também pode ser facilmente verificada

$$\begin{aligned}\mu^*(\tilde{X}_1) &= \mu^*((\tilde{X}_1 \cap \tilde{X}_2) \cup (\tilde{X}_1 \cap \tilde{X}_2^c)) \\ &= \mu^*(\tilde{X}_1 \cap \tilde{X}_2) + \mu^*(\tilde{X}_1 \cap \tilde{X}_2^c) \\ &= I(X_1; X_2) + H(X_1|X_2) = H(X_1). \end{aligned} \tag{133}$$

É possível então verificar a seguinte correspondência com as medidas de informação de Shannon

$$H/I \leftrightarrow \mu^* \tag{134}$$

$$, \leftrightarrow \cup \tag{135}$$

$$; \leftrightarrow \cap \tag{136}$$

$$| \leftrightarrow \setminus \tag{137}$$

- obs.: com a notação de medida, não existe distinção entre H e I, podemos escrever $H(X;Y) = I(X;Y)$, utilizando a notação do ponto-e-vírgula.

Desigualdade da soma de logaritmos

Teorema (desigualdade da soma de logaritmos)

Dados $(a_1, \ldots, a_n)$ *e* $(b_1, \ldots, b_n)$*, com* $a_i \geq 0$ *e* $b_i \geq 0$*, temos*

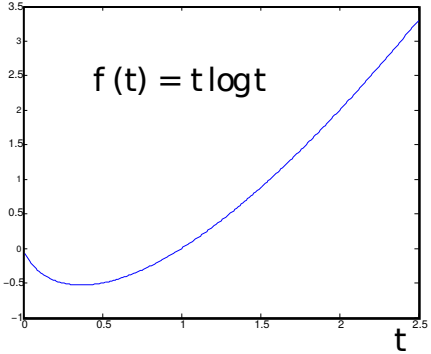
$$\sum_{i=1}^{n} a_i \log \frac{a_i}{b_i} \geq \left(\sum_{i=1}^{n} a_i \right) \log \frac{\sum_{i=1}^{n} a_i}{\sum_{i=1}^{n} b_i} \tag{138}$$

e teremos igualdade sse $a_i/b_i = c = const.$.

- Relembrando: $0 \log 0 = 0$, $a \log a/0 = \infty$ para $a > 0$, e $0 \log 0/0 = 0$.
- A desigualdade da soma de logaritmos é utilizada para demonstrar algumas propriedades importantes.

Desigualdade da soma de logaritmos

Considere $f(t) = t \log t = t(\ln t)(\log e)$, que é estritamente convexa, pois $f''(t) = 1/t \log e > 0$, $\forall t > 0$.



Desigualdade da soma de logaritmos I

Demonstração.

- Dada f convexa, a desigualdade de Jensen diz que

$$\sum_i \alpha_i f(t_i) \geq f\left(\sum_i \alpha_i t_i\right) \text{ com } \alpha_i \geq 0 \text{ e } \sum_i \alpha_i = 1 \tag{139}$$

- $f(x) = x \log x$ é estritamente convexa para $x > 0$, já que $f''(x) = \frac{1}{x} \log e > 0$ para $x > 0$.

...

Desigualdade da soma de logaritmos II

Demonstração.

continuação...

- Vamos fazer $\alpha_i = {}^{b_i}/_{\sum_{j=1}^{n} b_j}$ e $t_i = {}^{a_i}/_{b_i}$, então teremos

$$\begin{aligned} \sum_i \alpha_i f(t_i) &\geq f\left(\sum_i \alpha_i t_i\right) \\ \sum_i \left(\frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right)\right) &\geq f\left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i}\right) \end{aligned}$$

...

Desigualdade da soma de logaritmos III

Demonstração.

continuação...

$$
\begin{aligned}
\sum_i \left(\frac{b_i}{\sum_j b_j} f\left(\frac{a_i}{b_i}\right) \right) &\geq f\left(\sum_i \frac{b_i}{\sum_j b_j} \frac{a_i}{b_i} \right) \\
\frac{1}{\sum_j b_j} \sum_i \left(b_i \frac{a_i}{b_i} \log \frac{a_i}{b_i} \right) &\geq \left(\sum_i \frac{a_i}{\sum_j b_j} \right) \log \sum_i \frac{a_i}{\sum_j b_j} \\
\sum_i a_i \log \frac{a_i}{b_i} &\geq \left(\sum_i a_i \right) \log \sum_i \frac{a_i}{\sum_j b_j} \\
\sum_i a_i \log \frac{a_i}{b_i} &\geq \sum_i a_i \log \frac{\sum_i a_i}{\sum_j b_j} \qquad (140)
\end{aligned}
$$

□

Divergência é não negativa

A desigualdade da soma de logaritmos pode ser utilizada para mostrar que $D(p||q) \geq 0$.

Demonstração.

$$\begin{aligned} D(p||q) &= \sum_x p(x) \log \frac{p(x)}{q(x)} \\ &\geq \left(\sum_x p(x) \right) \log \frac{\sum_x p(x)}{\sum_x q(x)} \\ &= 1 \log \frac{1}{1} = 0 \end{aligned} \tag{141}$$

□

A Entropia Relativa é Convexa no Par I

Teorema

Seja (p_1, q_1) *e* (p_2, q_2) *dois pares de função massa probabilidade, então*

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) \leq \lambda D(p_1||q_1) + (1-\lambda)D(p_2||q_2), \tag{142}$$

para todo $0 \leq \lambda \leq 1$.

A Entropia Relativa é Convexa no Par II

Demonstração.

Pela definição da divergência de KL, temos

$$D(\lambda p_1 + (1-\lambda)p_2 || \lambda q_1 + (1-\lambda)q_2) = \sum_x (\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} \tag{143}$$

cada termo do somatório é da forma

$$(\lambda p_1(x) + (1-\lambda)p_2(x)) \log \frac{\lambda p_1(x) + (1-\lambda)p_2(x)}{\lambda q_1(x) + (1-\lambda)q_2(x)} = \left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i} \tag{144}$$

...

A Entropia Relativa é Convexa no Par III

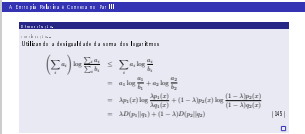
Demonstração.

continuação...

Utilizando a desigualdade da soma dos logaritmos

$$\begin{aligned}
\left(\sum_i a_i\right) \log \frac{\sum_i a_i}{\sum_i b_i} &\leq \sum_i a_i \log \frac{a_i}{b_i} \\
&= a_1 \log \frac{a_1}{b_1} + a_2 \log \frac{a_2}{b_2} \\
&= \lambda p_1(x) \log \frac{\lambda p_1(x)}{\lambda q_1(x)} + (1-\lambda) p_2(x) \log \frac{(1-\lambda) p_2(x)}{(1-\lambda) q_2(x)} \\
&= \lambda D(p_1 || q_1) + (1-\lambda) D(p_2 || q_2) \qquad (145)
\end{aligned}$$

□



- Note que podemos fazer $q_1 = q_2$ e desta forma obteremos convexidade apenas em p.
- Este é o fundamento para o procedimento de minimização alternada, que é um caso especial do algoritmo de EM (maximização da esperança), para o cálculo da função de taxa de distorção, e para o cálculo da função geral de capacidade de canal.

A Entropia é Concava I

Teorema

$H(p)$ é uma função concava de p.

A Entropia é Concava II

Demonstração.

$$
\begin{aligned}
H(p) &= -\sum_i p_i \log p_i = -\sum_i p_i \log p_i + \log|\mathcal{X}| - \log|\mathcal{X}| \\
&= \log|\mathcal{X}| - \sum_i p_i \log p_i - \log|\mathcal{X}| \underbrace{\sum_i p_i}_{=1} \\
&= \log|\mathcal{X}| - \sum_i (p_i \log p_i + p_i \log|\mathcal{X}|) \\
&= \log|\mathcal{X}| - \sum_i p_i (\log p_i - \log {}^{1}/_{|\mathcal{X}|}) \\
&= \underbrace{\log|\mathcal{X}|}_{\text{constante}} - \underbrace{D(p||u)}_{\text{convexo}} \qquad (146)
\end{aligned}
$$

onde u é a distribuição uniforme. □

A Entropia é Concava II

Demonstração.

$$
\begin{aligned}
H(p) &= -\sum_{i} p_i \log p_i = -\sum_{i} p_i \log p_i + \log|\mathcal{X}| - \log|\mathcal{X}| \\
&= \log|\mathcal{X}| - \sum_{i} p_i \log p_i - \log|\mathcal{X}| \underbrace{\sum_{i} p_i}_{=1} \\
&= \log|\mathcal{X}| - \sum_{i} \left(p_i \log p_i + p_i \log|\mathcal{X}|\right) \\
&= \log|\mathcal{X}| - \sum_{i} p_i \left(\log p_i - \log \mathcal{V}(x)\right) \\
&= \underbrace{\log|\mathcal{X}|}_{\text{entropia máxima}} - \underbrace{D(p\|u)}_{\text{entropia relativa}} \qquad (24)
\end{aligned}
$$

onde u é a **distribuição uniforme**. □

Podemos ver a entropia como a similaridade com a distribuição uniforme. Quanto maior a entropia, mais próximo estaremos da distribuição uniforme.

Consequências para a Informação Mútua I

Seja $(X, Y) \sim p(x, y) = p(x)p(y|x)$, a informação mútua $I(X; Y)$ é uma função côncava de $p(x)$ para $p(y|x)$ fixo e uma função convexa de $p(y|x)$ para $p(x)$ fixo.

Consequências para a Informação Mútua II

Demonstração.

- $I(X;Y)$ é uma função côncava de $p(x)$ para $p(y|x)$ fixo

$$\begin{aligned} I(X;Y) &= D(p(x,y)||p(x)p(y)) \text{ (definição)} \\ &= \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \text{ (definição)} \\ &= \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \end{aligned} \tag{147}$$

Se $p(y|x)$ é constante, então a informação mútua é função de $p(x)$

$$I_{p(x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{148}$$

...

Consequências para a Informação Mútua III

Demonstração.

continuação...

$$I_{p(x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{149}$$

Utilizando a propriedade da convexidade da divergência de Kullback-Leibler

$$I_{\lambda p_1(x)+(1-\lambda)p_2(x)}(X;Y) \geq \lambda I_{p_1(x)}(X;Y) + (1-\lambda) I_{p_2(x)}(X;Y) \tag{150}$$

então a informação mútua é uma função concava de $p(x)$ para $p(y|x)$ fixo.

...

Consequências para a Informação Mútua IV

Demonstração.

continuação...

- $I(X;Y)$ é uma função convexa de $p(y|x)$ para $p(x)$ fixo

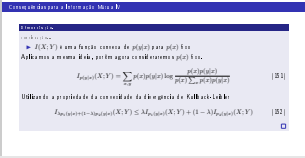
Aplicamos a mesma ideia, porém agora consideraremos $p(x)$ fixo.

$$I_{p(y|x)}(X;Y) = \sum_{x,y} p(x)p(y|x) \log \frac{p(x)p(y|x)}{p(x)\sum_x p(x)p(y|x)} \tag{151}$$

Utilizando a propriedade da convexidade da divergência de Kullback-Leibler

$$I_{\lambda p_1(y|x)+(1-\lambda)p_2(y|x)}(X;Y) \leq \lambda I_{p_1(y|x)}(X;Y) + (1-\lambda) I_{p_2(y|x)}(X;Y) \tag{152}$$

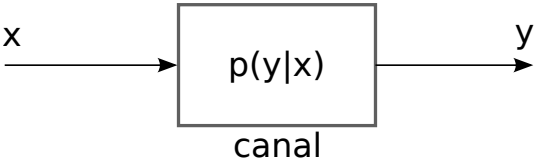
□



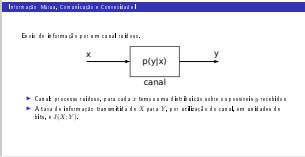
Estes resultados serão importantes para a capacidade de canal e vários outras otimizações envolvendo informação mútua e distribuições.

Informação Mútua, Comunicação e Convexidade I

Envio de informação por um canal ruidoso.



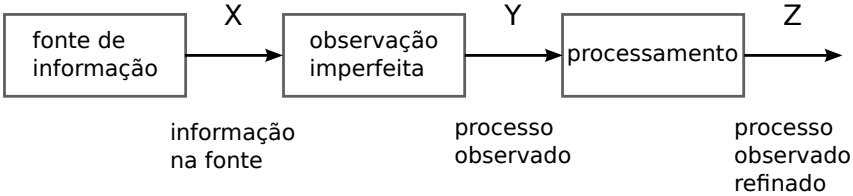
- Canal: processo ruidoso, para cada x temos uma distribuição sobre os possíveis y recebidos
- A taxa de informação transmitida de X para Y, por utilização do canal, em unidades de bits, é $I(X;Y)$.



- Embaralhando $p(x)$ não pode diminuir (pode aumentar ou não alterar) a transmissão de informação para um canal fixo, com relação à original mistura de taxas (ver Equação 150).
- Embaralhando $p(y|x)$ para um canal ruidoso e uma fonte fixa, podemos apenas não aumentar (pode reduzir ou manter constante) a taxa de transmissão, em relação à original mistura de taxas (ver Equação 152).

Desigualdade do Processamento de Dados

Dada uma fonte de informação, é possível utilizar alguma forma de processamento de dados de forma a obter mais informação sobre esta fonte?



Desigualdade do Processamento de Dados

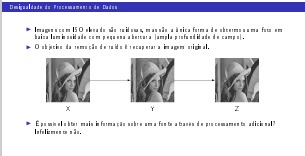
- Imagens com ISO elevado são ruidosas, mas são a única forma de obtermos uma foto em baixa luminosidade com pequena abertura (ampla profundidade de campo).
- O objetivo da remoção de ruído é recuperar a imagem original.

- É possível obter mais informação sobre uma fonte através de processamento adicional? Infelizmente não.

Desigualdade do Processamento de Dados

- Imagens com ISO elevado são ruidosas, mas são a única forma de obtermos uma foto em baixa luminosidade com pequena abertura (ampla profundidade de campo).
- O objetivo da remoção de ruído é recuperar a imagem original.

- É possível obter mais informação sobre uma fonte através de processamento adicional? Infelizmente não.



Profundidade de campo descreve até que ponto objetos que estão mais ou menos perto do plano de foco aparentam estar nítidos.

Regra geral, quanto menor for a abertura do diafragma/íris (maior o valor f/x), para uma mesma distância do objecto fotografado, maior será a distância do plano de foco a que os objetos podem estar enquanto permanecem nítidos.

Cadeia de Markov I

Definição (Cadeia de Markov)

As variáveis aleatórias X, Y e Z formam uma cadeia de Markov nesta ordem (denotado $X \to Y \to Z$) se a distribuição condicional de Z depende apenas de Y e é condicionalmente independente de X. Especificamente, X, Y e Z formam uma cadeia de Markov $X \to Y \to Z$ se a função massa de probabilidade conjunta pode ser escrita como

$$p(x, y, z) = p(x)p(y|x)p(z|y) \tag{153}$$

- $X \to Y \to Z$ sse X e Z são condicionalmente independentes dado Y ($X \perp\!\!\!\perp Z|Y$). Isto é,

$$p(x, z|y) = \frac{p(x, y, z)}{p(y)} = \frac{p(x)p(y|x)p(z|y)}{p(y)} = p(x|y)p(z|y) \quad \forall x, y, z \tag{154}$$

Cadeia de Markov II

- $X \to Y \to Z$ implica em $Z \to Y \to X$

Demonstração.

$$\begin{aligned} p(x,y,z) &= p(x)p(y|x)p(z|y) = p(x,y)p(z|y) \\ &= \frac{p(x,y)p(z|y)p(y)}{p(y)} = p(x|y)p(y,z) \\ &= p(x|y)p(y|z)p(z) \end{aligned} \tag{155}$$

□

- Se $Z = f(Y)$, então $X \to Y \to Z$ (i.e., X, Y e Z formam uma cadeia de Markov). $f(\cdot)$ pode ser aleatória ou determinística. X é irrelevante para determinar Z quando Y é dado.

Desigualdade do Processamento de Dados I

Teorema (Desigualdade do Processamento de Dados)

Se $X \to Y \to Z$ *então*

$$I(X;Y) \geq I(X;Z) \tag{156}$$

- Na cadeia de Markov, as setas correspondem ao processamento e as variáveis aleatórias correspondem aos dados.
- O processamento pode ser aleatório ou determinístico.
- A desigualdade de processamento de dados diz que ao efetuar mais processamento dos dados, só é possível perder informação sobre a fonte original, quando medida pela informação mútua.

Desigualdade do Processamento de Dados II

Demonstração.

Utilizando a regra da cadeia da informação mútua, teremos

$$\begin{aligned} I(X;Y,Z) &= I(X;Z) + I(X;Y|Z) \\ &= I(X;Y) + I(X;Z|Y) \end{aligned} \tag{157}$$

Como $X \perp\!\!\!\perp Z|Y$ (X e Z são condicionalmente independentes, dado Y), temos que $I(X;Z|Y) = 0$. Como $I(X;Y|Z) \geq 0$, teremos

$$I(X;Z) + \underbrace{I(X;Y|Z)}_{\geq 0} = I(X;Y) + \cancelto{0}{I(X;Z|Y)} \tag{158}$$

Então

$$I(X;Z) \leq I(X;Y) \tag{159}$$

□

Desigualdade do Processamento de Dados III

- Teremos igualdade sse $I(X;Y|Z) = 0$ (i.e. $X \to Z \to Y$).
- De forma similar, podemos mostrar que $I(Y;Z) \geq I(X;Z)$.

Corolário

Se $Z = g(Y)$, então $I(X;Y) \geq I(X;g(Y))$.
Demonstração: $X \to Y \to g(Y)$ forma uma cadeia de Markov.

Corolário

Se $X \to Y \to Z$ então $I(X;Y|Z) \leq I(X;Y)$.

$$\underbrace{I(X;Z)}_{\geq 0} + I(X;Y|Z) = I(X;Y) + \cancelto{0}{I(X;Z|Y)} \tag{160}$$

então

$$I(X;Y|Z) \leq I(X;Y) \tag{161}$$

Desigualdade do Processamento de Dados IV

- Processamento pode apenas perder informação sobre X. Quando X é a fonte e Y o receptor, nenhum processamento irá aumentar a informação sobre X.
- Considere o reconhecimento de padrões: X é um objeto, Y é uma lista de características e $f(Y)$ é processamento subsequente. Então, qualquer processamento subsequente poderá apenas reduzir a informação sobre o objeto.
- Como funciona então as técnicas de remoção de ruído em imagens ou áudio?
  - As técnicas supõem o conhecimento de algumas informações sobre a imagem original, ou seja, utilizam conhecimento *a priori* para processar a imagem.

Corolário

Se $X \to Y \to Z$, *então* $I(X;Y|Z) \leq I(X;Y)$. *I.e.,*
$I(X;Y|Z) = H(X|Z) - H(X|Y,Z) \leq H(X) - H(X|Y)$.

Estatística T I

Exemplo

Seja $X_1, X_2, \ldots, X_N$, $X_i \in \{0, 1\}$ uma sequência i.i.d. de arremessos de moeda, $p(X = 1) = \theta = 1 - P(X = 0)$.
Faça $T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i$, contagem do número de *caras*.
Dizemos que T é uma **estatística** da amostra.

- De forma geral, uma estatística é uma função de uma coleção de variáveis aleatórias (e.g., uma média empírica, uma variância empírica, ou um máximo empírico, etc)
- Uma estatística é por sua vez uma v.a.
- Uma boa estatística possui informação útil sobre as amostras, enquanto uma estatística ruim não (por exemplo $T(X_1, \ldots, X_N) = X_1$).
- As estatísticas costumam ser chamadas de 'características' no contexto de reconhecimento de padrões e aprendizado de máquina.

Ensaio de Bernoulli I

- Considere a estatística de contagem citada anteriormente.
- Uma vez que sabemos a estatística, a probabilidade de uma sequência pode ser expressa sem fazer referência à θ (parâmetro que caracteriza a distribuição).

$$p(x_1, \ldots, x_N | T(x_1, \ldots, x_N), \theta) = p(x_1, \ldots, x_N | T(x_1, \ldots, x_N)) = \begin{cases} \frac{1}{\binom{N}{k}} & \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \tag{162}$$

- Em outras palavras: $X_{1:N} \perp\!\!\!\perp \theta | T(X_{1:N})$.
- Isto implica na cadeia de Markov: $\theta \to T(X_{1:N}) \to X_{1:N}$
- Por outro lado, sabemos que $T(X_{1:N})$ é uma função de $X_{1:N}$.
- Desta forma, também temos a seguinte cadeia de Markov: $\theta \to X_{1:N} \to T(X_{1:N})$

Desigualdade de Processamento de Dados e Estatística I

- cadeia de Markov (A): $\theta \to T(X_{1:N}) \to X_{1:N}$.
- pela desigualdade de processamento de dados em (A) teremos: $I(\theta; T(X_{1:N})) \geq I(\theta; X_{1:N})$.
- cadeia de Markov (B): $\theta \to X_{1:N} \to T(X_{1:N})$.
- pela desigualdade de processamento de dados em (B) teremos: $I(\theta; X_{1:N}) \geq I(\theta; T(X_{1:N}))$.
- então, (A) e (B) $\Rightarrow I(\theta; X_{1:N}) = I(\theta; T(X_{1:N}))$, e nenhuma informação é perdida sobre θ indo de $X_{1:N}$ para $T(X_{1:N})$.

Desigualdade de Processamento de Dados e Estatística II

Definição (Estatística Suficiente)

Uma função $T(\cdot)$ é dita ser uma estatística suficiente em relação à família $\{f_\theta(x)\}$ se X é independente de θ dado $T(X)$ para qualquer distribuição em θ (i.e. $\theta \to T(X) \to X$ forma uma cadeia de Markov). Então

$$I(\theta; X) = I(\theta; T(X)) \ \forall \theta \tag{163}$$

Uma estatística suficiente preserva a informação mútua e reciprocamente

$$X \perp\!\!\!\perp \theta | T(X) \tag{164}$$

- i.e., uma cadeira de Markov (A) é condição suficiente para suficiência de uma estatística.
- uma estatística suficiente é utilizada para estimar os parâmetros a partir dos dados: no limite em que temos infinitos dados, teremos uma estimativa exata (consistência assintótica).

Estatística Suficiente I

Exemplo

Seja $X_1, \ldots, X_N$, $X_i \in \{0, 1\}$, uma sequência i.i.d. de lances de uma moeda com parâmetro $\theta = Pr(X_i = 1)$. Dado N, o número de 1's é uma estatística suficiente para θ.

$$T(X_1, \ldots, X_N) = \sum_{i=1}^{N} X_i \tag{165}$$

Dado T, todas as sequências com o mesmo número de 1's são igualmente prováveis e independentes do parâmetro θ.

...

Estatística Suficiente II

Exemplo

continuação...

Existem $\binom{N}{k}$ sequências de comprimento N com k 1's e são todas equiprováveis. $Pr(X_{1:N} = x_{1:N}) = \theta^k(1-\theta)^{N-k}$. Então

$$Pr\{(X_1, \ldots, X_N) = (x_1, \ldots, x_N) | \sum_{i=1}^{N} X_i = k\} = \begin{cases} \frac{1}{\binom{N}{k}} & \text{se } \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \tag{166}$$

Temos então que $\theta \to \sum X_i \to (X_1, \ldots, X_N)$ forma uma cadeia de Markov e T é uma estatística suficiente para θ (dado $\sum X_i$, a sequência $(X_1, \ldots, X_N)$ é estatisticamente independente de θ).

Teorema da Fatoração de Fisher-Neyman I

Teorema (Teorema da Fatoração de Fisher-Neyman)

Se a função densidade de probabilidade é $f_\theta(x)$, então T é suficiente para θ se e somente se podemos encontrar funções não-negativas g e h tais que

$$f_\theta(x) = h(x) g_\theta(T(x)), \tag{167}$$

i.e., a densidade f pode ser fatorada em um produto tal que um fator h não depende de θ e o outro fator, que depende de θ, dependerá de x apenas por meios de $T(X)$.

Estatística Suficiente I

Teorema (Estatística Suficiente)

$T(\cdot)$ é suficiente para θ sse a probabilidade $p(x_{1:N}|\theta)$ pode ser escrita como o produto

$$p(x_{1:N}|\theta) = g(T,\Theta)h(x_{1:N}) \tag{168}$$

$$\begin{aligned} p(x_{1:N}|\theta) &= g(T,\Theta)h(x_{1:N}) \\ &= g(T,\Theta)h(x_{1:N},T(x_{1:N})) \end{aligned} \tag{169}$$

Estatística Suficiente II

Definição (Independencia Condicional)

Dadas três variáveis aleatórias A, B, C, temos que $A \perp\!\!\!\perp B|C$ sse existem funções g e h tais que $p(a, b, c)$ possa ser reescrita na forma

$$p(a, b, c) = g(a, c)h(b, c) \tag{170}$$

Estatística Suficiente I

Exemplo

Se X possui distribuição normal com média θ e variância 1

$$f_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-(x-\theta)^2/2} = N(\theta, 1) \tag{171}$$

e $X_1, \ldots, X_n$ são tiradas de forma independente de acordo com esta distribuição. Uma estatística suficiente para θ é a média amostral

$$\overline{X_n} = \frac{1}{n} \sum_{i=1}^{n} X_i \tag{172}$$

...

Estatística Suficiente II

Exemplo

continuação...

$$
\begin{aligned}
f_\theta(x_1,\ldots,x_n) &= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{\sum_{i=1}^n (x_i\theta-\theta^2/2)} \\
&= \left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2} e^{\theta n\left(\frac{1}{n}\sum_{i=1}^n x_i-\frac{\theta}{2}\right)} \\
&= \underbrace{\left(\frac{1}{\sqrt{2\pi}}\right)^n e^{-\frac{1}{2}\sum_{i=1}^n x_i^2}}_{h(x_1,\ldots,x_n)} \underbrace{e^{\theta n\left(T(X_{1:n})-\frac{\theta}{2}\right)}}_{g_\theta(T(x_{1:n}))}
\end{aligned}
\tag{173}
$$

Então, pelo teorema de Fisher-Neyman, podemos concluir que a média amostral é uma estatística suficiente para θ quando X possui distribuição normal.

Tipo da Amostra I

Exemplo

Seja $X_1, \ldots, X_N \equiv X_{1:N}$ uma amostra de comprimento N de uma variável aleatória discreta D-ária. Então $x_i \in \mathcal{X}$, o tamanho do alfabeto é $D = |\mathcal{X}|$, e $\mathcal{X} = \{a_1, \ldots, a_D\}$.
Define-se uma estatística: o histograma empírico da amostra.

$$P_{x_{1:N}} \triangleq \left(\frac{N(a_1|x_{1:N})}{N}, \frac{N(a_2|x_{1:N})}{N}, \ldots, \frac{N(a_D|x_{1:N})}{N} \right), \tag{174}$$

onde $N(a_i|x_{1:N})$ é a contagem do número de ocorrências do símbolo a_i na amostra $x_{1:N}$. O histograma é uma estatística, já que é uma função da amostra. É uma estatística suficiente?

...

Tipo da Amostra II

Exemplo

continuação...

Para o caso em que $D = 2$, temos o teste de Bernoulli visto anteriormente. Para D qualquer, temos

$$\begin{aligned} p(x_{1:N}|P_{x_{1:N}}, \theta) &= \begin{cases} \frac{1}{\binom{N}{N_1, N_2, \ldots, N_D}} & \text{se } \forall i, N_i = NP_{x_{1:N}}(a_i) \\ 0 & \text{caso contrário,} \end{cases} \\ &= p(x_{1:N}|P_{x_{1:N}}) \end{aligned} \tag{175}$$

onde temos o coeficiente multinomial $\binom{n}{k_1, k_2, \ldots, k_m} = \frac{n!}{k_1! k_2! \ldots k_m!}$. Podemos observar que $p(x_{1:N}|P_{x_{1:N}}, \theta) = p(x_{1:N}|P_{x_{1:N}})$, ou seja, é independente de θ.
Então $X_{1:N} \perp\!\!\!\perp \theta | P_{x_{1:N}}$, então $P_{x_{1:N}}$ é uma estatística suficiente.

Teorema Multinomial

$$(x_1 + x_2 + \ldots + x_m)^n = \sum_{k_1+k_2+\ldots+k_m=n} \binom{n}{k_1, k_2, \ldots, k_m} \prod_{1\leq t\leq m} x_t^{k_t} \tag{176}$$

Caso Binário - Suficiência do Tipo I

Exemplo

- $X_i \in \{0,1\}$, $T(x_{1:N})$ = número de 1s em $x_{1:N}$.
- A probabilidade conjunta:

$$p(x_{1:N}, T(x_{1:N}), \theta) = \prod_{a \in \mathcal{X}} p(a)^{N(a|x_{1:N})} = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{177}$$

- Evento $\{x_{1:N}, T(x_{1:n}) = k\}$ quando k é o verdadeiro número de 1s em $x_{1:N}$ e é o mesmo que o evento $\{x_{1:n}\}$. Quando k não é o número de 1s, temos probabilidade zero (impossível).

...

Caso Binário - Suficiência do Tipo II

Exemplo

continuação...

- Marginal $p(\theta, T(x_{1:N}) = k)$:

$$\begin{aligned} p(\theta, T(x_{1:N}) = k) &= \sum_{x_{1:N}} p(x_{1:N}, T(x_{1:N}) = k, \theta) \\ &= \sum_{x_{1:N}:T(x_{1:N})=k} p(x_{1:N}, T(x_{1:N}) = k, \theta) \\ &= \binom{N}{k} p(0)^{N-k} p(1)^k \end{aligned} \tag{178}$$

...

Caso Binário - Suficiência do Tipo III

Exemplo

continuação...

- A probabilidade conjunta

$$p(x_{1:N}, T(x_{1:N}), \theta) = p(0)^{N(0|x_{1:N})} p(1)^{N(1|x_{1:N})} \tag{179}$$

- A marginal

$$p(\theta, T(x_{1:N}) = k) = \binom{N}{k} p(0)^{N-k} p(1)^k \tag{180}$$

- Então

$$p(x_{1:N}|T, \Theta) = \frac{p(x_{1:N}, T, \Theta)}{p(T, \Theta)} = \begin{cases} \frac{1}{\binom{N}{k}} & \text{se } \sum_i x_i = k \\ 0 & \text{caso contrário} \end{cases} \tag{181}$$

Estatística Mínima Suficiente I

Definição

Uma estatística $T(X)$ é uma estatística mínima suficiente em relação a $\{p_\theta(x)\}$ se ela for uma função de todas as demais estatísticas suficientes U.

- Sabemos pela definição de T mínima e qualquer outra estatística suficiente U que $\theta \to X_{1:N} \to U(X_{1:N}) \to T(X_{1:N})$
- Interpretando com relação à desigualdade do processamento de dados, temos

$$\theta \to T(X) \to U(X) \to X \tag{182}$$

- A estatística mínima suficiente T fornece qualquer outra estatística U independente do parâmetro θ.
- O fato de que é uma estatística significa que $p(X|T,U,\theta) = p(X|T,U) = p(X|T)$, o que significa que T é, para todos propósitos, uma substituto estatístico mínimo para θ no cálculo da probabilidade.

Estatística Suficiente I

Exemplo (Entropia Condicional Nula)

Mostre que se $H(Y|X) = 0$, então Y é uma função de X, i.e., para todo x com $p(x) > 0$, existe apenas um possível valor de y com $p(x, y) > 0$.

solução

Assuma que existe x, digamos x_0, e dois valores diferentes de y, digamos y_1 e y_2, tal que $p(x_0, y_1) > 0$ e $p(x_0, y_2) > 0$. Então a marginal é $p(x_0) \geq p(x_0, y_1) + p(x_0, y_2) > 0$. Temos também

$$p(y_1|x_0) = \frac{p(x_0, y_1)}{p(x_0)} \text{ e } p(y_2|x_0) = \frac{p(x_0, y_2)}{p(x_0)} \tag{183}$$

então ambos $p(y_1|x_0)$ e $p(y_2|x_0)$ não são iguais a 0 (zero) ou 1 (um). ...

Estatística Suficiente II

Exemplo (Entropia Condicional Nula)

continuação...

$$
\begin{aligned}
H(Y|X) &= E[H(Y|X)] \\
&= -\sum_x p(x) H(Y|X=x) \\
&= -\sum_x p(x) \sum_y p(y|x) \log p(y|x) \\
&\geq -p(x_0) \sum_y p(y|x_0) \log p(y|x_0) \\
&\geq -\underbrace{p(x_0)}_{>0}[\underbrace{p(y_1|x_0)}_{>0} \underbrace{\log p(y_1|x_0)}_{<0} + \underbrace{p(y_2|x_0)}_{>0} \underbrace{\log p(y_2|x_0)}_{<0}] \\
&> 0
\end{aligned}
\tag{184}
$$

...

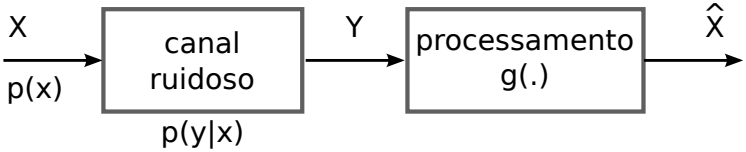
Estatística Suficiente III

Exemplo (Entropia Condicional Nula)

continuação...

Então, a entropia condicional $H(Y|X)$ é nula se e somente se Y for uma função de X. Se Y for uma função de X, teremos $p(y_1|x_0) = 0$ ou 1, ou seja, a probabilidade $p(y_i|x_0)$ será igual a 1 apenas para um y_i e zero para os demais.

Erro nas Comunicações I



- $\hat{X}$ é uma estimativa de X.
- a estimativa é errada quando $X \neq \hat{X}$
- probabilidade de erro: $P_e \triangleq p(X \neq \hat{X})$
- podemos relacionar a entropia condicional $H(X|Y)$ com a probabilidade de erro P_e?
- sabemos (exercício anterior) que a entropia condicional $H(X|Y)$ é nula se e somente se X for uma função de Y
- esperamos ser capazes de estimar X com baixa probabilidade de erro apenas quando a entropia condicional $H(X|Y)$ for pequena

Desigualdade de Fano

Teorema (Desigualdade de Fano)

Para qualquer estimador $\hat{X}$ tal que $X \to Y \to \hat{X}$, com $P_e = Pr(X \neq \hat{X})$, temos

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{185}$$

Esta desigualdade pode ser simplificada (menos rígida) na forma

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} \tag{186}$$

onde utilizamos $H(P_e) \leq 1$.

Note que $P_e = 0 \Rightarrow H(X|Y) = 0$ pois $H(P_e) = 0$ e $H(X|Y) \geq 0$.

Desigualdade de Fano

Teorema (Desigualdade de Fano)

Para qualquer estimador $\hat{X}$ tal que $X \to Y \to \hat{X}$, com $P_e = Pr(X \neq \hat{X})$, temos

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{185}$$

Esta desigualdade pode ser simplificada (menos rígida) na forma

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} \tag{186}$$

onde utilizamos $H(P_e) \leq 1$.

Note que $P_e = 0 \Rightarrow H(X|Y) = 0$ pois $H(P_e) = 0$ e $H(X|Y) \geq 0$.

Esta desigualdade será utilizada para provar o reverso no teorema de codificação de Shannon, i.e., que qualquer código com probabilidade de erro $\to 0$, à medida que o comprimento do bloco cresce, devemos ter uma taxa $R < C$ (a capacidade do canal, a ser definida).

Para o caso de um alfabeto binário ($|\mathcal{X}| = 2$), a desigualdade de Fano na forma da Equação 186 não poderá ser aplicada. Devemos então utilizar a forma mais relaxada:

$$P_e \geq \frac{H(X|Y) - 1}{\log(|\mathcal{X}| - 1)} > \frac{H(X|Y) - 1}{\log|\mathcal{X}|} \tag{187}$$

Desigualdade de Fano I

Demonstração.

Definir uma função de erro:

$$E = \begin{cases} 1 & \text{, se } \hat{X} \neq X \text{(erro)} \\ 0 & \text{, se } \hat{X} = X \text{(sem erro)} \end{cases} \tag{188}$$

...

Desigualdade de Fano II

Demonstração.

continuação...

Utilizando a regra da cadeia temos:

$$H(E,X|\hat{X}) = H(X|\hat{X}) + \underbrace{H(E|X,\hat{X})}_{=0}$$

ou

$$= \underbrace{H(E|\hat{X})}_{\leq H(E)=H(P_e)} + \underbrace{H(X|E,\hat{X})}_{\leq P_e \log(|\mathcal{X}|-1)} \tag{189}$$

- O erro é uma função determinística de X e $\hat{X}$, então, sabendo X e $\hat{X}$, determinamos E. Desta forma: $H(E|X,\hat{X}) = 0$.
- Condicionar só pode reduzir a entropia: $H(E|\hat{X}) \leq H(E) = H(P_e)$.
- Veremos abaixo que $H(X|E,\hat{X}) \leq P_e \log(|\mathcal{X}|-1)$.

...

Desigualdade de Fano III

Demonstração.

continuação...

$$
\begin{aligned}
H(X|\hat{X},E) &= p(E=0)\underbrace{H(X|\hat{X},E=0)}_{=0} + p(E=1)H(X|\hat{X},E=1) \\
&= (1-P_e)0 + P_e H(X|\hat{X},E=1) \leq P_e \log(|\mathcal{X}|-1)
\end{aligned}
$$

- Se não há erro e conhecemos $\hat{X}$, então determinamos X. Não existe entropia residual em X quando é dado $\hat{X}$ e $E=0$. Logo $H(X|\hat{X},E=0)=0$.
- Se conhecemos $\hat{X}$ e existe um erro $(E=1)$, então sabemos que X é diferente de $\hat{X}$, logo isto nos deixa com $(|\mathcal{X}|-1)$ alternativas.

...

Desigualdade de Fano IV

Demonstração.

continuação...

Temos então

$$\begin{aligned} H(X|\hat{X}) &= H(E|\hat{X}) + H(X|E,\hat{X}) \\ &\leq H(P_e) + P_e \log(|\mathcal{X}| - 1) \end{aligned} \tag{190}$$

Como $X \to Y \to \hat{X}$ é uma cadeia de Markov, podemos utilizar a desigualdade de processamento de dados.

$$\begin{aligned} I(X;Y) &\geq I(X;\hat{X}) \\ H(X) - H(X|Y) &\geq H(X) - H(X|\hat{X}) \\ H(X|\hat{X}) &\geq H(X|Y) \end{aligned} \tag{191}$$

...

Desigualdade de Fano V

Demonstração.

continuação...

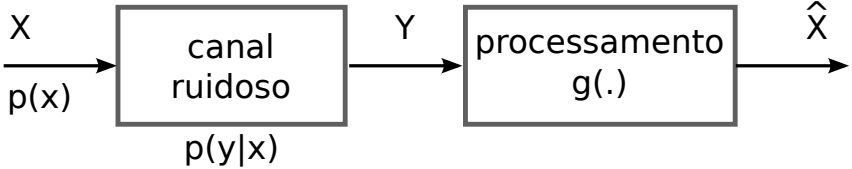
Então, utilizando as Equações 190 e 191, obtemos

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{192}$$

□

Desigualdade de Fano - Sumário

Considere a seguinte situação: enviamos X através de um canal ruidoso, recebemos Y e realizamos algum pós-processamento.



$\hat{X}$ é uma estimativa de X.

- Erro: $X \neq \hat{X}$; com probabilidade $P_e \triangleq p(X \neq \hat{X})$.
- Intuitivamente, a entropia condicional deveria nos dizer algo sobre a probabilidade de erro. Na verdade temos o seguinte:

Teorema (Desigualdade de Fano)

$$H(P_e) + P_e \log(|\mathcal{X}| - 1) \geq H(X|\hat{X}) \geq H(X|Y) \tag{193}$$

Desigualdade de Fano I

Exemplo

Considere uma v.a. discreta $X \in \mathcal{X} = \{1, 2, \ldots, 5\}$ com função massa de probabilidade $p(x) = (0.35, 0.35, 0.1, 0.1, 0.1)$. Seja $Y \in \mathcal{Y} = \{1, 2\}$, de forma que, se $x \leq 2$ teremos $y = x$ com probabilidade $6/7$ e, se $x > 2$, teremos $y = 1$ ou 2 com igual probabilidade. A melhor estratégia é utilizar o estimador $\hat{x} = y$. Calcule a probabilidade de erro e o limite dado pela desigualdade de Fano.

...

Desigualdade de Fano II

Exemplo

continuação...

solução

A distribuição condicional $p(y|x)$ é apresentada na tabela abaixo:

X \ Y	1	2
1	$6/7$	$1/7$
2	$1/7$	$6/7$
3	$1/2$	$1/2$
4	$1/2$	$1/2$
5	$1/2$	$1/2$

...

Desigualdade de Fano III

Exemplo

continuação...

A efetiva probabilidade de erro é dada por

$$\begin{aligned} P_e &= 1 - P_a \text{ (prob. de acerto)} \\ &= 1 - \sum_{i=1}^{5} P(x_i = y_i) \\ &= 1 - (p(y=1|x=1)p(x=1) + p(y=2|x=2)p(x=2) + 0 + 0 + 0) \\ &= 1 - \left(\frac{6}{7}0.35 + \frac{6}{7}0.35\right) = 0.4 = \frac{2}{5} \end{aligned} \tag{194}$$

...

Desigualdade de Fano IV

Exemplo

continuação...

A desigualdade de Fano fornece um limite inferior pra a probabilidade de erro (predição incorreta do valor de X baseado em Y). Este limite inferior é determinado pela incerteza remanescente $H(X|Y)$ sobre X quando Y é conhecido.

Pelo teorema da desigualdade de Fano temos que

$$P_e \geq \frac{H(X|Y) - 1}{\log\left(|\mathcal{X}| - 1\right)} \tag{195}$$

...

Desigualdade de Fano V

Exemplo

continuação...

Precisaremos calcular

$$\begin{aligned} H(X|Y) &= -\sum_{x,y} p(x,y)\log p(x|y) \\ &= -\sum_{x,y} p(y|x)p(x)\log p(x|y) \end{aligned} \tag{196}$$

onde $p(y|x)$ e $p(x)$ são dados do problema e ainda será necessário calcular $p(x|y)$ para encontrar $H(X|Y)$.

...

Desigualdade de Fano VI

Exemplo

continuação...

$$
\begin{aligned}
P(X|Y=1) &= \frac{P(X,Y=1)}{P(Y=1)} = \frac{P(Y=1|X)P(X)}{P(Y=1)} \\
&= \frac{(\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{6}{7},\frac{1}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)\right)} \\
&= \frac{(0.3,0.05,0.05,0.05,0.05)}{1/2} \\
&= (0.6,0.1,0.1,0.1,0.1) \qquad (197)
\end{aligned}
$$

...

Desigualdade de Fano VII

Exemplo

continuação...

$$
\begin{aligned}
P(X|Y=2) &= \frac{(\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)}{\sum\left((\frac{1}{7},\frac{6}{7},\frac{1}{2},\frac{1}{2},\frac{1}{2})\cdot(0.35,0.35,0.1,0.1,0.1)\right)} \\
&= (0.1,0.6,0.1,0.1,0.1) \qquad (198)
\end{aligned}
$$

...

Desigualdade de Fano VIII

Exemplo

continuação...

Desta forma teremos

$$\begin{aligned} H(X|Y) &= H(X|Y=1)P(Y=1) + H(X|Y=2)P(Y=2) \\ &= -\frac{1}{2}\left(0.6\log 0.6 + 4\times 0.1\log 0.1\right) - \frac{1}{2}\left(4\times 0.1\log 0.1 + 0.6\log 0.6\right) \\ &= 1.771 \text{ bits.} \end{aligned} \tag{199}$$

Utilizando a desigualdade de Fano

$$P_e \geq \frac{H(X|Y-1)}{\log\left(|\mathcal{X}|-1\right)} = \frac{1.771-1}{\log(5-1)} = 0.3855 \tag{200}$$

Desigualdade de Fano I

Lema

Se X e X' são i.i.d. com entropia $H(X)$,

$$Pr(X = X') \geq 2^{-H(X)} \tag{201}$$

com igualdade se e somente se X possuir distribuição uniforme.

$$\begin{aligned} Pr(X = X') &= Pr(X = x_1 | X' = x_1) Pr(X' = x_1) + \ldots + \\ &\quad Pr(X = x_n | X' = x_n) Pr(X' = x_n) \\ &= Pr(X = x_1) Pr(X' = x_1) + \ldots + \\ &\quad Pr(X = x_n) Pr(X' = x_n) \\ &= p^2(x_1) + \ldots + p^2(x_n) = \sum_x p^2(x) \end{aligned} \tag{202}$$

Desigualdade de Fano II

Demonstração.

Suponha que $X \sim p(x)$. Pela desigualdade de Jensen temos

$$2^{E[\log p(X)]} \leq E[2^{\log p(X)}] \tag{203}$$

pois 2^x é convexa. Logo,

$$\begin{aligned} 2^{-H(X)} &= 2^{\sum_x p(x)\log p(x)} = 2^{E[\log p(X)]} \\ &\leq E[2^{\log p(X)}] \\ &= \sum_x p(x) 2^{\log p(x)} = \sum_x p(x)p(x) \\ &= \sum_x p^2(x) = Pr(X = X') \end{aligned} \tag{204}$$

□

Desigualdade de Fano III

Note que, para maximizar a probabilidade $Pr(X = X')$, devemos minimizar a entropia. No limite, quando $H(X) = 0$, teremos $Pr(X = X') \geq 1$, logo será igual a 1 e assim $X = X'$ sem dúvida.

Desigualdade de Fano I

Corolário

Seja X, X' independentes com $X \sim p(x)$ e $X' \sim q(x)$, $x, x' \in \mathcal{X}$, então

$$\begin{aligned} Pr(X = X') &\geq 2^{-H(p)-D(p||q)} \\ Pr(X = X') &\geq 2^{-H(q)-D(q||p)} \end{aligned} \tag{205}$$

ou seja

$$Pr(X = X') \geq \max\left(2^{-H(p)-D(p||q)}, 2^{-H(q)-D(q||p)}\right) \tag{206}$$

Desigualdade de Fano II

Demonstração.

$$
\begin{aligned}
2^{-H(p)-D(p||q)} &= 2^{\sum_x p(x)\log p(x)+\sum_x p(x)\log\frac{q(x)}{p(x)}} \\
&= 2^{\sum_x p(x)\log q(x)} \\
&= 2^{E_p[\log q(X)]} \\
&\leq \sum_x p(x)2^{\log q(x)} \text{ (Jensen)} \\
&= \sum_x p(x)q(x) \\
&= Pr(X = X') \qquad (207)
\end{aligned}
$$

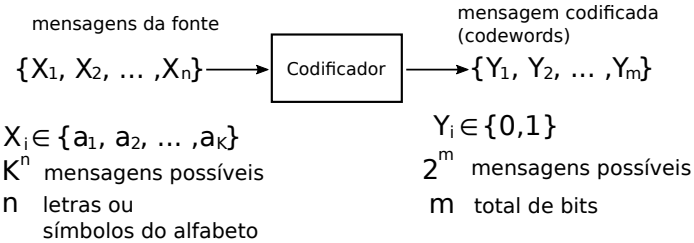
□

Propriedade da Equipartição Assintótica

- Vamos considerar blocos de realizações de uma variável aleatória (i.e., vetores aleatórios de comprimento n). $n =$ tamanho do bloco.
- Sejam $X_1, X_2, \ldots, X_n$ v.a.s i.i.d. com distribuição p (dizemos $X_i \sim p(x)$).
- Existem K símbolos possíveis (alfabeto ou espaço-estado de tamanho K), então $X_i \in \{a_1, a_2, \ldots, a_K\}$.
- Consideram n variáveis aleatórias $(X_1, X_2, \ldots, X_n)$, existem K^n possíveis realizações.

Propriedade da Equipartição Assintótica

Suponha que desejamos codificar as K^n possíveis realizações com uma sequência de dígitos binários de comprimento m. Então, existem 2^m palavras de código (*codewords*).



- Para que seja possível termos uma palavra de código para cada mensagem possível, devemos satisfazer a seguinte condição:

$$2^m \geq K^n \tag{208}$$

ou seja

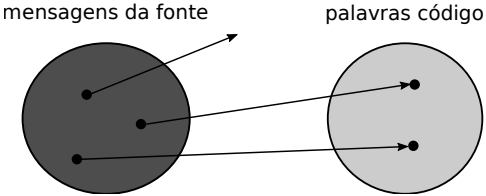
$$m \geq (\log K)n \tag{209}$$

Propriedade da Equipartição Assintótica

- Quantos bits por letra da fonte utilizamos?

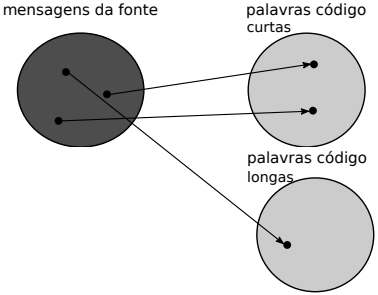
$$\text{taxa} = \frac{m}{n} \geq \log K \text{ bits por letra da fonte} \tag{210}$$

  Exemplo: 26 letras, precisaremos de $\lceil \log K \rceil = 5$bits.
- Podemos utilizar menos bits por símbolo emitido pela fonte (na média) e ainda sim não ter erro? Sim.
- Algumas mensagens da fonte poderiam não ter a elas um código associado.



Propriedade da Equipartição Assintótica

- Ao invés de descartar algumas mensagens, podemos associar a elas palavras longas e às outras palavras associamos palavras curtas.



- Em qualquer um dos casos, quando n é grande suficiente, podemos fazer com que a probabilidade, de se obter uma dessas mensagens da fonte que gerariam erro (ou que teriam palavras longas associadas), muito pequena.

Probabilidade de Palavras da Fonte

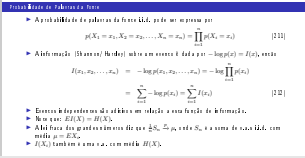
- A probabilidade de palavras da fonte i.i.d. pode ser expressa por

$$p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = \prod_{i=1}^{n} p(X_i = x_i) \tag{211}$$

- A informação (Shannon/Hartley) sobre um evento é dada por $-\log p(x) = I(x)$, então

$$\begin{aligned} I(x_1, x_2, \ldots, x_n) &= -\log p(x_1, x_2, \ldots, x_n) = -\log \prod_{i=1}^{n} p(x_i) \\ &= \sum_{i=1}^{n} -\log p(x_i) = \sum_{i=1}^{n} I(x_i) \end{aligned} \tag{212}$$

- Eventos independentes são aditivos em relação a esta função de informação.
- Note que: $EI(X) = H(X)$.
- A lei fraca dos grandes números diz que $\frac{1}{n} S_n \xrightarrow{p} \mu$, onde S_n é a soma de v.a.s i.i.d. com média $\mu = EX_i$.
- $I(X_i)$ também é uma v.a. com média $H(X)$.



Lei dos Grandes Números

Se um evento de probabilidade p é observado repetidamente em ocasiões independentes, a proporção da frequência observada deste evento em relação ao total número de repetições converge em direção a p à medida que o número de repetições se torna arbitrariamente grande.

Sejam $X_1, X_2, \ldots, X_n$ v.a.s i.i.d. com $\mathrm{E}X_i = \mu$ e $\mathrm{Var}X_i = \sigma^2 < \infty$, para $i = 1, \ldots, n$. Seja a média definida por $\overline{X_n} = \frac{1}{n}\sum_{i=1}^{n} X_i$, então, para $\varepsilon > 0$, a **Lei Fraca dos Grandes Números** diz que $\overline{X_n}$ converge em probabilidade para μ, ou seja,

$$\lim_{n\to\infty} P\left(|\overline{X_n} - \mu| < \varepsilon\right) = 1. \quad (213)$$

Lei fraca dos grandes números e Entropia

- Combinando o que vimos anteriormente, obtemos

$$\frac{1}{n}\sum_{i=1}^{n} I(X_i) \xrightarrow[n\to\infty]{p} H(X) \tag{214}$$

- Se n fica grande suficiente, obteremos

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} I(x_i) &\approx H(X) \text{ onde } \forall i, x_i \sim p(x) \\
-\frac{1}{n}\sum_{i=1}^{n} \log p(x_i) &\approx H(X) \\
-\log \prod_{i=1}^{n} p(x_i) &\approx nH(X) \\
-\log p(x_1, x_2, \ldots, x_n) &\approx nH(X) \\
p(x_1, x_2, \ldots, x_n) &\approx 2^{-nH(X)}
\end{aligned} \tag{215}$$

Propriedade da Equipartição Assintótica

Quando n é grande suficiente, teremos

$$p(x_1, x_2, \ldots, x_n) \approx 2^{-nH(X)} \tag{216}$$

- Esta probabilidade não depende da sequência em si. Depende apenas do comprimento n e da entropia da v.a.
- Quando n fica grande, podemos dizer que todas as sequências terão a mesma probabilidade: 2^{-nH}.
- Estas sequências que possuem esta probabilidade (praticamente todas as sequências) são chamadas de sequências **típicas**, e são representadas pela conjunto A.

Quase todos eventos são quase equiprováveis

- Se $X_1, X_2, \ldots, X_n$ são i.i.d. e $X_i \sim p(x)$ para todo i, e se n é grande suficiente, então qualquer amostra $x_1, x_2, \ldots, x_n$ terá probabilidade da amostra essencialmente independente da amostra, i.e.,

$$p(x_1, \ldots, x_n) \approx 2^{-nH(X)} \tag{217}$$

  onde $H(X)$ é a entropia de $p(x)$.
- Então, podem existir no máximo 2^{nH} amostras, e pode ser que $2^{nH} \ll K^n$.
- Estas amostras que ocorrem são chamadas de típicas, e são representadas por $A_\epsilon^{(n)}$.
- Uma grande porção de $\mathcal{X}^n$ não irá ocorrer, i.e., pode acontecer que $2^{nH} \ll |\mathcal{X}^n| = K^n$.

Conjunto Típico

- Seja $A_\epsilon^{(n)}$ o conjunto das sequências típicas (i.e., aquelas com probabilidade 2^{-nH}).
- Se "todos" eventos possuem a mesma probabilidade p, então existem $1/p$ deles.
- O número de sequências típicas é

$$|A_\epsilon^{(n)}| \approx 2^{nH(X)}. \tag{218}$$

- Desta forma, para representar (ou codificar) as sequencias típicas, precisaremos de $nH(X)$ bits. Teremos então

$$m = nH(X) \tag{219}$$

no modelo do codificador. Então a taxa será $H(X)$.

Codificando apenas o Conjunto Típico

- Tomando $m = nH$, teremos que o número médio de bits por letra do alfabeto da fonte será dado por

$$\frac{m}{n} = H \text{ que pode ser } \leq \log K \tag{220}$$

- Interpretações para a Entropia na codificação de fonte:
  1) A probabilidade de uma sequência típica é $2^{-nH(X)}$.
  2) O número de sequências típicas é $2^{nH(X)}$.
  3) O número de bits por símbolo da fonte é $H(X)$, quando codificamos apenas o conjunto típico.

Bernoulli

Considere o experimento de Bernoulli com $X_1, \ldots, X_n$ i.i.d. e probabilidade $p(X_i = 1) = p = 1 - p(X_i = 0)$. A probabilidade de uma dada sequência será dada por

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \tag{221}$$

- Existem 2^n sequencias possíveis.
- Todas elas possuem a mesma probabilidade? Não. Considere $p = 0.1$, $(1-p) = 0.9$. A sequência de apenas zeros é a sequência de mais provável.

Bernoulli

Considere a experiência de Bernoulli com $X_1, \ldots, X_n$ i.i.d. e probabilidade $p(X_i = 1) = p = 1 - p(X_i = 0)$. A probabilidade de uma dada sequência será dada por

$$p(x_1, x_2, \ldots, x_n) = \prod_{i=1}^{n} p^{x_i}(1-p)^{1-x_i} = p^{\sum_i x_i}(1-p)^{n-\sum_i x_i} \qquad [221]$$

- Existem 2^n sequências possíveis.
- Todas elas possuem a mesma probabilidade? Não. Considere $p = 0.1$, $(1-p) = 0.9$. A sequência de apenas zeros é a sequência de mais provável.

Todas as sequências que possuem 'alguma' probabilidade, terão a mesma probabilidade? Depende do que queremos dizer com 'alguma'. Para valores pequenos de n, não, mas à medida que n cresce, algo acontece e a resposta 'sim' começa a ser a mais apropriada.

Propriedade de Equipartição Assintótica

- É possível prever a probabilidade de que uma determinada sequências terá uma probabilidade particular?
$$\Pr(p(X_1, X_2, \ldots, X_n) = \alpha) = ? \tag{222}$$
- Note que $p(X_1, X_2, \ldots, X_n)$ é uma variável aleatória. É uma probabilidade que é uma função do conjunto de variáveis aleatórias.
- Teremos
$$\Pr(p(X_1, X_2, \ldots, X_n) \approx 2^{-nH}) \approx 1 \tag{223}$$
quando n é grande suficiente.
- Quase todos os eventos (que ocorrem com alguma probabilidade) são todos equiprováveis.

Ensaio de Bernoulli

Exemplo

Seja $S_n \sim$ Binomial(n,p) com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim$ Bernoulli(p). Teremos então $ES_n = np$ e var$(S) = npq$, onde $q = 1-p$, e

$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{224}$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n(-p\log p-(1-p)\log(1-p))} \\ &= 2^{\log p^{np}+\log(1-p)^{n(1-p)}} \\ &= p^{np}q^{nq} \end{aligned} \tag{225}$$

$H = H(p)$ é a entropia binária com probabilidade p. np é o número esperado de 1s e nq é o número esperado de 0s.

Ensaio de Bernoulli

Exemplo

Seja $S_n \sim$ **Binomial**(n, p) com $S_n = X_1 + X_2 + \ldots + X_n$, $X_i \sim$ **Bernoulli**(p). Teremos então $ES_n = np$ e $var(S) = np\,q$, onde $q = 1 - p$, e

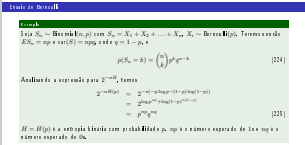
$$p(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{224}$$

Analisando a expressão para 2^{-nH}, temos

$$\begin{aligned} 2^{-nH(p)} &= 2^{-n(-p\log p-(1-p)\log(1-p))} \\ &= 2^{\log p^{np}+\log(1-p)^{n(1-p)}} \\ &= p^{np}q^{nq} \end{aligned} \tag{225}$$

$H = H(p)$ é a entropia binária com probabilidade p, np é o número esperado de 1s e nq é o número esperado de 0s.

- Todas as sequências que ocorrem são aquelas cujo número de 1s e 0s são aproximadamente iguais aos seus valores esperados.
- Nenhum outra sequência possui probabilidade significativa.
- A sequência $X_1, X_2, \ldots, X_n$ foi assumida como sendo i.i.d., entretanto podemos estender para cadeias de Markov e processos aleatórios estacionários ergódicos.

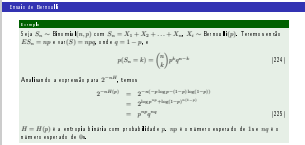


O cálculo computacional do coeficiente binomial pode apresentar perda de precisão numérica, por envolver fatoriais e frações de números muito grandes. Exemplo de execussão no GNU-Octave para calcular $\binom{100}{20}$:

```
» nchoosek(100,20)
warning: nchoosek: possible loss of precision
warning: called from
```

Uma possível solução é utilizar o pacote simbólico para efetuar os cálculos. Podemos verificar que realmente ocorreu erro de precisão numérica ao realizar o cálculo computacional.

```
» double(nchoosek(sym(100),sym(20))) - nchoosek(100,20)
ans = -65536
```

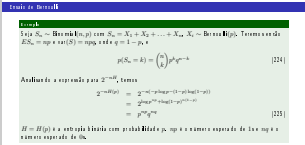


Outra alternativa é realizar uma aproximação para calcular $\binom{N}{k}$.
Iremos utilizar a aproximação de Stirling para a função fatorial:

$$\ln n! = n \ln n - n + \mathcal{O}(\ln n) \tag{226}$$

Utilizando esta aproximação em $\ln \binom{N}{k}$, teremos

$$\begin{aligned} \ln \binom{N}{k} &\equiv \ln \frac{N!}{(N-k)!k!} = \ln N! - \ln(N-k)! - \ln k! \\ &\simeq N \ln N - N - (N-k)\ln(N-k) + (N-k) - k \ln k + k \\ &= \underbrace{(N-k)\ln N - (N-k)\ln N}_{=0} + N \ln N - N - (N-k)\ln(N-k) + (N-k) - k\ln k + k \\ &= (N-k)\ln\frac{N}{N-k} + k \ln \frac{N}{k} \\ &= N\left(-\frac{N-k}{N}\ln\frac{N-k}{k} - \frac{k}{N}\ln\frac{k}{N}\right) = NH_e\left(\frac{k}{N}\right). \end{aligned} \tag{227}$$



Concluímos então que

$$\ln \binom{N}{k} \simeq N H_e \left(\frac{k}{N}\right) \tag{228}$$

e, como os temos em ambos os lados envolvem logaritmos, podemos realizar a mudança de base em ambos os lados (basta multiplicar por $\ln 2$),

$$\log \binom{N}{k} \simeq N H \left(\frac{k}{N}\right), \tag{229}$$

onde agora utilizamos a entropia binária em bits. Assim, teremos

$$\binom{N}{k} \simeq 2^{N H\left(\frac{k}{N}\right)}. \tag{230}$$

Propriedade da Equipartição Assintótica

Teorema (Propriedade da Equipartição Assintótica)

Se $X_1, X_2, \ldots, X_n$ são i.i.d. e $X_i \sim p(x)$ para todo i, então

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \xrightarrow{p} H(X) \tag{231}$$

Demonstração.

$$\begin{aligned} -\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) &= -\frac{1}{n}\log \prod_{i=1}^{n} p(X_i) \\ &= -\frac{1}{n}\sum_{i} \log p(X_i) \xrightarrow{p} E\log p(X) \\ &\quad \text{onde utilizamos a lei fraca dos números grandes} \\ &= H(X) \end{aligned} \tag{232}$$

Conjunto Típico

Definição (Conjunto Típico)

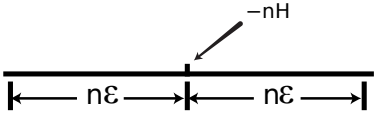
Um conjunto típico $A_\epsilon^{(n)}$ em relação a $p(x)$ é o conjunto de sequências $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ com propriedade

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)} \tag{233}$$

De forma equivalente, podemos escrever

$$A_\epsilon^{(n)} = \left\{ (x_1, x_2, \ldots, x_n) : | -\frac{1}{n} \log p(x_1, \ldots, x_n) - H| < \epsilon \right\} \tag{234}$$

O conjunto típico é formado pelas sequências com $\log$ da probabilidade dentro de seguinte extensão



Conjunto Típico

- O tamanho do conjunto típico de sequências produzidas pela fonte é tipicamente muito menor que o tamanho do conjunto de todas as sequências produzidas pela fonte.
- Uma sequencia típica não precisa ter probabilidade próxima daquela que é a sequência mais provável.
- Geralmente a sequência mais provável não está no conjunto típico.

Propriedades do Conjunto Típico

Teorema (Propriedades do Conjunto Típico $A_\epsilon^{(n)}$)

1) *Se* $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, *então*

$$H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon \tag{235}$$

2) $p(A_\epsilon^{(n)}) = p\left(\left\{x : x \in A_\epsilon^{(n)}\right\}\right) > 1 - \epsilon$ *para* n *grande suficiente, para todo* $\epsilon > 0$.
3) *Limite superior:* $|A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$, *onde* $|A_\epsilon^{(n)}|$ *é o número de elementos no conjunto* $A_\epsilon^{(n)}$.
4) *Limite inferior:* $|A_\epsilon^{(n)}| \geq (1-\epsilon)2^{n(H(X)-\epsilon)}$ *para* n *grande suficiente.*

- O conjunto típico possui, essencialmente, probabilidade 1 (algo típico irá tipicamente ocorrer).
- Todos os itens neste conjunto terão a mesma probabilidade $\approx 2^{-nH}$.

Exemplo $K = |\{0,1\}|$: Conjunto Típico

- Suponha o Ensaio de Bernoulli com distribuição uniforme, $K = 2$, $\mathcal{X} = \{0,1\}$, $p = 0.5$, entropia $H = 1$ e $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, então todas as sequências ocorreram com igual probabilidade.
- Considere agora uma distribuição não-uniforme, $p = 0.1$ e $q = 1 - p = 0.9$, a entropia será $H \approx 0.469$. Considere $n = 100$, então $K^{100} = 2^{100} = 10^{\log_{10} 2^{100}} \approx 10^{100 \times 0.30103} \approx 10^{30}$, a capacidade representacional das sequencias da fonte. Mas $|A_\epsilon^{(n)}| = 2^{nH} = 10^{nH \times \log_{10} 2} \approx 10^{14} \ll 10^{30} \approx K^{100}$. O número de sequências típicas é muito menor que o número de possíveis sequências.
- Ineficiência: capacidade representacional é muito maior do que as coisas que ocorrem. O alfabeto da fonte é pobre para realizar compressão.
- Assuma ϵ muito pequeno, então onde foi parar a massa das $\approx 10^{30} - 10^{14}$ sequencias? (veremos adiante)

Conjunto Típicos são Típicos

- Pela definição
$$p(A_\epsilon^{(n)}) > 1 - \epsilon \text{ para qualquer } \epsilon > 0 \tag{236}$$
- Então $A_\epsilon^{(n)}$ possui praticamente toda probabilidade, e cada elemento em $A_\epsilon^{(n)}$ possui a mesma probabilidade, então
$$p(x) \approx 2^{-nH} \forall x \in A_\epsilon^{(n)} \tag{237}$$
- Exemplo: Ensaio de Bernoulli $X_i \sim$ Bernoulli(p) com $p(X_i = 1) = p = 1 - p(X_i = 0)$ e $p > 0.5$.
- Probabilidade de n 1s sucessivos é p^n e esta é a sequência mais provável.
- Probabilidade de uma sequência típica é 2^{-nH}.
- Para $n = 100$, $p - 0.9 = 1 - q$, a sequência mais provável possui probabilidade $p^n \approx 2.66 \times 10^{-5}$, mas uma sequencia típica possui probabilidade $2^{-nH} \approx 7.62 \times 10^{-15}$.

Exemplo $K = |\{0,1\}|$: Conjunto Típico

- Suponha o Ensaio de Bernoulli com distribuição uniforme, $K = 2$, $\mathcal{X} = \{0,1\}$, $p = 0.5$, entropia $H = 1$ e $|A_\epsilon^{(n)}| = 2^{nH} = 2^n = K^n$, então todas as sequências ocorreram com igual probabilidade.
- Considere agora uma distribuição não-uniforme, $p = 0.1$ e $q = 1 - p = 0.9$, a entropia será $H \approx 0.469$. Considere $n = 100$, então $K^{100} = 2^{100} = 10^{\log_{10} 2^{100}} \approx 10^{100 \times 0.30103} \approx 10^{30}$, a capacidade representacional das sequencias da fonte. Mas $|A_\epsilon^{(n)}| = 2^{nH} = 10^{nH \times \log_{10} 2} \approx 10^{14} \ll 10^{30} \approx K^{100}$. O número de sequências típicas é muito menor que o número de possíveis sequências.
- Ineficiência: capacidade representacional é muito maior do que as coisas que ocorrem. O alfabeto da fonte é pobre para realizar compressão.
- Assuma ϵ muito pequeno, então onde foi parar a massa das $\approx 10^{30} - 10^{14}$ sequencias? (veremos adiante)

Conjunto Típicos são Típicos

- Pela definição
$$p(A_\epsilon^{(n)}) > 1 - \epsilon \text{ para qualquer } \epsilon > 0 \tag{238}$$
- Então $A_\epsilon^{(n)}$ possui praticamente toda probabilidade, e cada elemento em $A_\epsilon^{(n)}$ possui a mesma probabilidade, então
$$p(x) \approx 2^{-nH} \forall x \in A_\epsilon^{(n)} \tag{239}$$
- Exemplo: Ensaio de Bernoulli $X_i \sim$ Bernoulli(p) com $p(X_i = 1) = p = 1 - p(X_i = 0)$ e $p > 0.5$.
- Probabilidade de n 1s sucessivos é p^n e esta é a sequência mais provável.
- Probabilidade de uma sequência típica é 2^{-nH}.
- Para $n = 100$, $p - 0.9 = 1 - q$, a sequência mais provável possui probabilidade $p^n \approx 2.66 \times 10^{-5}$, mas uma sequencia típica possui probabilidade $2^{-nH} \approx 7.62 \times 10^{-15}$.

Sequencias Não-Típicas não são típicas

- $p^n \gg 2^{-nH}$: a sequência mais provável é muito mais provável do que uma sequencia típica.
- O que ocorre com a probabilidade da sequencia mais provável, na média (por símbolo), quando n cresce

$$-\frac{1}{n}\log p^n = -\log p = \log 1/p \xrightarrow[n\to\infty]{p} H? \tag{240}$$

- O conjunto típico possui, essencialmente, toda a probabilidade $p(A_\epsilon^{(n)}) > 1-\epsilon$.
- A sequencia mais provável está no conjunto típico? Não, já que a probabilidade da sequencia mais provável não é 2^{-nH}.
- $n = 100$, $p = 0.9 = 1-q$. Considere uma sequência com noventa 1s e dez 0s. A probabilidade é $p^{90}(1-p)^{10} \approx 7.62 \times 10^{-15} \approx 2^{-nH}$.
- Esta sequencia muito improvável é típica.

Probabilidade Média de Sequencias

- Pela PEA (Propriedade da Equipartição Assintótica), temos que se $(x_1, \ldots, x_n)$ é típico (i.e., $(x_1, \ldots, x_n) \in A_\epsilon^{(n)}$, para qualquer $\epsilon > 0$), então

$$-\frac{1}{n} \log p(x_1, \ldots, x_n) \xrightarrow[n \to \infty]{p} H \tag{241}$$

- Para as sequencias mais prováveis, quando n é grande suficiente, teremos (para $p > 0.5$)

$$-\frac{1}{n} \log p^n = -\log p = \log 1/p \xrightarrow[n \to \infty]{p} -\log p \tag{242}$$

- A probabilidade média da sequencia mais provável é bem diferente da sequencia típica.

Sequencias Típicas

- o conjunto típico possui essencialmente toda da probabilidade.
- Como pode uma sequencia com a maior probabilidade não ser típica, mas uma sequência com probabilidade muito menor ser típica?
- Existe um número exponencialmente crescente de sequencias típicas, cada uma com probabilidade menor do que a sequencia de maior probabilidade.
- A probabilidade individual de cada sequencia vai a zero quando n cresce.
- O tamanho do conjunto típico cresce rapidamente, quando $n \to \infty$, de forma que a probabilidade de $A_\epsilon^{(n)}$ vai a 1.
- O tamanho do conjunto das sequencias muito prováveis cresce lentamente, de forma que a probabilidade do conjunto vai a zero quando $n \to \infty$.

Sequencias Típicas I

A probabilidade de uma sequência $x_{1:n}$, onde $x_i \in \mathcal{X} = \{a_1, \ldots, a_K\}$, é dada por

$$P(x_{1:n}) = P(x_1)P(x_2)\ldots P(x_n) \approx p_1^{p_1 n} p_2^{p_2 n} \ldots p_K^{p_K n}, \tag{243}$$

onde consideramos que, em uma sequência muito longa, esperamos observar $p_1 n$ ocorrências do símbolo a_1, $p_2 n$ ocorrências do símbolo a_2, etc.
A informação associada a esta sequência típica é

$$\log \frac{1}{P(x_{1:n})} \approx \sum_{i=1}^{K} \log p_i^{-p_i n} \tag{244}$$

$$= n \sum_{i=1}^{K} p_i \log \frac{1}{p_i} = nH(X). \tag{245}$$

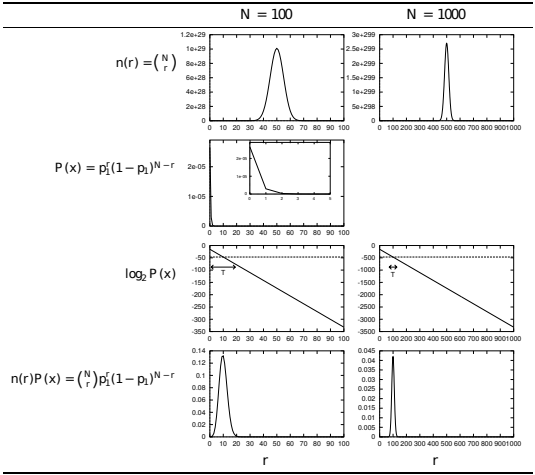
Uma sequência típica terá probabilidade próxima de $2^{-nH(X)}$.

Sequencias Típicas II

x	$\log_2(P(x))$
...1...................1.....1....1.1.......1........1...........1.....................1.......11...	−50.1
......................1.....1.....1.......1....1.........1.....................................1....	−37.3
........1....1..1...1....11..1.1.........11.........................1...1.1..1...1................1.	−65.9
1.1...1................1.......................11.1..1............................1.....1..1.11.....	−56.4
...11...........1...1.....1.1......1..........1....1...1.....1............1.........................	−53.2
..............1......1.........1.1.......1..........1............1...1......................1.......	−43.7
.....1........1.......1...1............1............1...........1......1..11........................	−46.8
.....1..1..1...............111...................1...............1.........1.1...1...1.............1	−56.4
.........1..........1.....1......1..........1....1..1...	−37.3
......1........................1..............1.....1..1.1.1..1...................................1.	−43.7
1.......................1..........1...1...................1....1....1........1..11..1.1...1........	−56.4
...........11.1.........1................1......1.....................1.............................	−37.3
.1..........1...1.1.............1.......11...........1.1...1..............1.............11..........	−56.4
......1...1..1.....1..11.1.1.1...1.....................1............1.............1..1..............	−59.5
............11.1......1....1..1............................1.......1..............1.......1.........	−46.8
....	−15.2
1111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111111	−332.1

Figura 5: Sequências regadas por um ensaio de Bernoulli com $n = 100$ e $P(X = 1) = p = 0.1$. As 15 sequências superiores representam amostras típicas. As duas últimas sequências representam a sequência mais provável e a menos provável MacKay (2003).

Sequencias Típicas III



Sequencias Típicas IV

Figura 6: Para $p = 0.1$, $n = 100$ e $n = 1000$ os gráficos ilustram $n(r)$, o número de strings contendo r 1s; a probabilidade $P(x_{1:n})$ para uma string contendo r 1s; a mesma probabilidade em escala logarítmica; e a probabilidade total $n(r)P(x_{1:n})$ de todas as strings contendo r 1s MacKay (2003).

Sequencias Típicas V

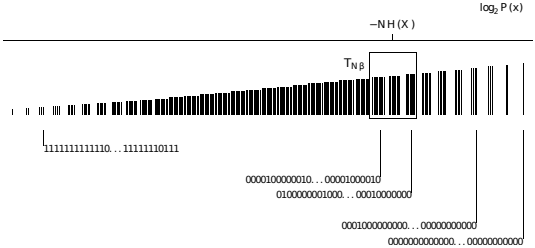


Figura 7: Diagrama esquemático ilustrando todas as sequencias no conjunto $\mathcal{X}^n$ ordenadas pela probabilidade MacKay (2003).

O termo **equipartição** é utilizado para descrever a ideia de que os membros do conjunto típico possuem aproximadamente a mesma probabilidade.

Hartley, R. V. L. (1928). Transmission of information1. *Bell System Technical Journal*, 7(3):535–563.

MacKay, D. J. C. (2003). *Information theory, inference, and learning algorithms*. Cambridge University Press, Cambridge, UK ; New York.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.*, 27(3):379–423.

Wikipedia (2020a). Etaoin shrdlu. `https://en.wikipedia.org/wiki/Etaoin_shrdlu`. [Online; accessed 10-August-2020].

Wikipedia (2020b). Jean-dominique bauby. `https://en.wikipedia.org/wiki/Jean-Dominique_Bauby`. [Online; accessed 10-August-2020].

Wikipedia (2020c). Letter frequency. `https://en.wikipedia.org/wiki/Letter_frequency`. [Online; accessed 10-August-2020].

Wikipedia (2020d). Morse code. `https://en.wikipedia.org/wiki/Morse_code`. [Online; accessed 10-August-2020].