Fundamentos da Teoria de Informação

"In fact, the science of thermodynamics began with an analysis, by the great engineer Sadi Carnot, of the problem of how to build the best and most efficient engine, and this constitutes one of the few famous cases in which engineering has contributed to fundamental physical theory. Another example that comes to mind is the more recent analysis of information theory by Claude Shannon. These two analyses, incidentally, turn out to be closely related".

Richard Feynman – Lectures on Physics

1. Introdução

A *teoria de informação*, a cuja elaboração se associa o nome do brilhante engenheiro eletricista e matemático Claude Elwood Shannon, é, sem dúvida, uma das construções intelectuais mais relevantes do século passado. Justificativas para essa afirmação podem advir de duas perspectivas: 1) uma prática, baseada na

"onipresença" de mecanismos de processamento e transmissão de informação no mundo atual e 2) uma teórica, que deve ressaltar o impacto da teoria de informação em campos tão diversos quanto física teórica, biologia, economia, estatística etc.

Em nosso curso, terá particular relevância o uso de conceitos de teoria de informação no âmbito do tratamento de sinais de diversas naturezas. Em outras palavras, buscaremos usar tais conceitos para quantificar e analisar o conteúdo informacional de dados das mais diversas naturezas.

Tendo em vista a amplitude de ideias associadas à teoria, optamos por dedicar um tópico à exposição de conceitos fundamentais, os quais serão usados em tópicos seguintes. Podemos dizer, sem receio, que a familiaridade com esse conteúdo é, atualmente, muito importante para todos aqueles que trabalham com as várias facetas da área de inteligência computacional.

2. Informação

Embora a palavra “informação” seja muito empregada nos dias atuais, seu significado tem algo de etéreo, o que torna desafiadora a tarefa de defini-la.

Neste tópico, atrelaremos essa definição à ideia de probabilidade. Em termos simples, buscaremos estabelecer um valor de informação que se associe a “o que se ganha” conhecendo o resultado de um experimento aleatório. Portanto, informação, de certa forma, se associa a *incerteza*.

Para que entendamos o espírito da definição, imaginemos que nos sejam entregues dois bilhetes: o primeiro diz “Seu amigo, o Sr. Y, acabou de saber que ganhou na loteria.” e o segundo diz “Sua amiga, a Sra. Z, descobriu que o bebê que está esperando é uma menina.”. É razoável esperar que a primeira notícia seja bastante surpreendente, enquanto a segunda, provavelmente, não nos causará tanto espanto.

Isso nos leva a concluir que o primeiro bilhete "trouxe mais informação" que o segundo.

Matematicamente, se supusermos que um evento A tem probabilidade $P(A)$ de ocorrer, a informação associada à observação da ocorrência desse evento poderia ser dada por* (REZA, 1994):

$$I\big(P(A)\big) = \log_2(1/P(A)) = -\log_2(P(A))$$

É importante ressaltar que essa grandeza, devido ao uso do logaritmo na base dois, é medida em *bits*. Outra possibilidade seria usar o logaritmo natural, o que levaria a uma medida em *nats*.

Se considerarmos uma moeda perfeitamente honesta, poderemos dizer que a informação associada à observação do resultado de um lançamento é de 1 bit. A

---

* A grandeza definida a seguir é muitas vezes denominada, em língua inglesa, *self-information*. A mesma expressão em inglês também é usada às vezes para denotar outra grandeza que veremos em breve, a *entropia*.

observação do resultado de lançamento de um dado honesto, por outro lado, leva a uma informação de 2,585 bits. Se estivermos lidando com N eventos equiprováveis, a informação associada à observação de um deles é de $\log_2(N)$.

3. Entropia

Consideremos uma variável aleatória X, de natureza discreta, com função de massa de probabilidade $p(x) = P[X = x]$. A entropia associada a essa variável é dada pela expressão a seguir:

$$H(X) = -\sum_x p(x) log_2[p(x)]$$

Perceba que a entropia nada mais é do que a informação média associada às observações relativas à variável aleatória. Também é possível dizer que a entropia é uma medida da incerteza média associada à variável (COVER & THOMAS, 2006). Note

que a unidade de entropia é a mesma da medida de informação empregada e que, usando os conceitos vistos no tópico anterior, podemos também dizer que:

$$H(X) = -E\{log_2[p(x)]\}$$

sendo $E\{\cdot\}$ o operador de média estatística tomado com respeito à função de massa de probabilidade $p(x)$.

A seguir, daremos alguns exemplos que ajudarão a ilustrar algumas propriedades dos conceitos apresentados.

3.1. Entropia - Exemplos

Consideremos uma variável aleatória que pode assumir dois valores, $X = a$ e $X = b$. Diremos que $P[X = a] = p$ e, naturalmente, que $P[X = b] = 1-p$. Em tal situação, a entropia de X será dada pela expressão:

$$H(X) = -p\log_2(p) - (1 - \mathrm{p})\log_2(1 - p)$$

Consideremos, inicialmente, que os dois valores são equiprováveis, ou seja, que $p = 0{,}5$. Nesse caso, $H(X)$ será igual a 1 bit, ou seja, será de 1 bit a informação média associada à observação.

Consideremos agora que o valor a possua uma maior probabilidade de ocorrência, digamos, $p = 0{,}7$. Nesse caso, $H(X)$ será igual a 0,8813 bit, ou seja, a informação média associada será menor. Intuitivamente, por que isso ocorre? De certa forma, a resposta é que, uma vez que um dos valores é mais provável que o outro, temos uma "expectativa mais precisa" do resultado da observação de X, ou seja, uma menor incerteza. Portanto, é de se esperar que haja uma diminuição de entropia em relação ao caso equiprovável ($p = 0{,}5$). Nessa linha de raciocínio, espera-se também que uma probabilidade ainda maior, digamos, $p = 0{,}9$, produza uma entropia ainda menor. Isso de fato ocorre, uma vez que $H(X)$ será igual a 0,469 bit nesse caso.

Usando um limite, é possível mostrar que, se p tende a um (e $1-p$ tende a zero) ou se p tende a zero (e $1-p$ tende a um), a entropia será nula, já que não haverá incerteza alguma.

Na Figura 1, apresentamos o valor de entropia para todos os possíveis valores de p. O gráfico corresponde ao que foi discutido: a entropia é máxima no caso equiprovável (em que não temos como "tomar partido" de nenhum valor em particular) e tende a zero nos extremos (nos quais não há incerteza).

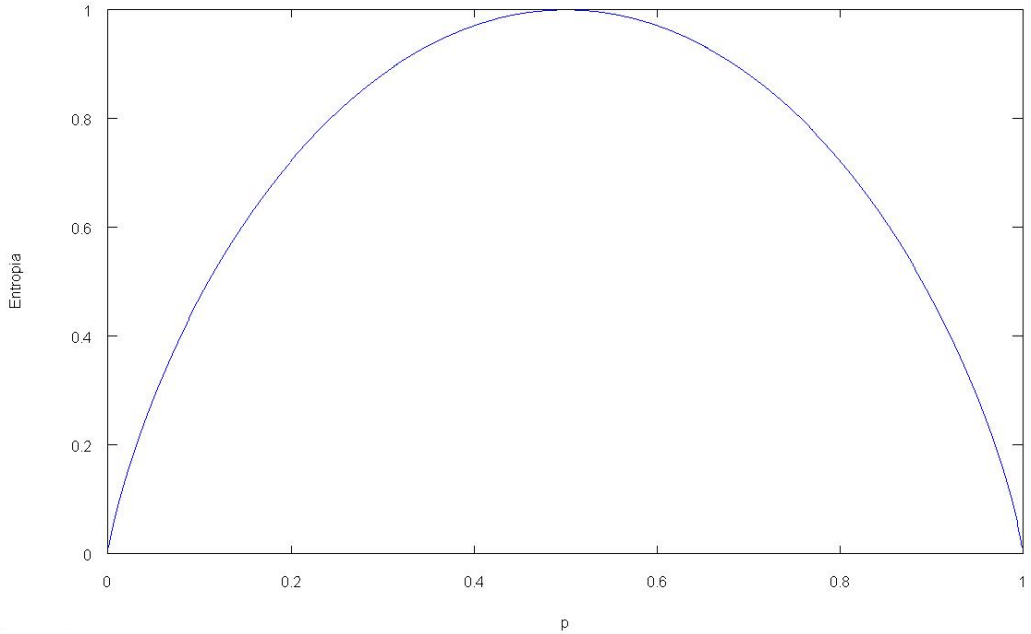

Figura 1 – Entropia de uma variável aleatória com dois possíveis valores.

Com esse exemplo, podemos enunciar duas propriedades importantes relativas à entropia mesmo no caso de uma variável aleatória que assume múltiplos valores: ela é máxima para o caso equiprovável e se anula para o caso em que a probabilidade associada a um determinado valor é igual a um (ou seja, no caso determinístico).

Um outro exemplo, inspirado em (COVER & THOMAS, 2006), evoca a discussão que teremos no curso sobre árvores de decisão. Imagine que uma variável aleatória se vincule às seguintes probabilidades: $P(X = a) = 0{,}5$, $P(X = b) = 0{,}25$, $P(X = c) = 0{,}125$ e $P(X = d) = 0{,}125$. Podemos calcular sem dificuldades a entropia dessa variável, que é $H(X) = 1{,}75$.

Agora imaginemos que estejamos interessados em realizar um procedimento "em árvore" para determinar o valor dessa variável. A primeira pergunta ideal é "X vale a?", já que esse valor é o de maior probabilidade. Em seguida, pelo mesmo motivo, faremos a pergunta "X vale b?". Por fim, podemos fazer a pergunta "X vale c?" ou a pergunta "X vale d?", e isso conclui o processo.

No esquema dado acima, temos 50% de chance de resolver o problema com uma única questão, 25% de resolver o problema com duas questões e 25% de chance de

resolver o problema com três questões. Ou seja, o número médio de questões será $0{,}5 \times 1 + 0{,}25 \times 2 + 0{,}25 \times 3 = 1{,}75$.

A igualdade entre o número médio de questões e o valor da entropia não é incidental. De fato, pode-se mostrar que o mínimo número médio de questões para definir o valor de uma variável aleatória estará entre $H(X)$ e $H(X) + 1$ (COVER & THOMAS, 2006).

3.2. Entropia Conjunta e Entropia Condicional

A entropia conjunta de duas variáveis aleatórias X e Y é dada por (COVER & THOMAS, 2006):

$$H(X,Y) = -\sum_x \sum_y p(x,y) log_2[p(x,y)] = -E\{log_2[p(x,y)]\}$$

A entropia condicional $H(Y|X)$, por sua vez, é dada por:

$$H(Y|X) = -\sum_x \sum_y p(x,y) log_2[p(y|x)]$$

Um resultado muito interessante é que a entropia conjunta pode ser escrita em função da entropia condicional da seguinte forma:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

Uma desigualdade importante é a mostrada a seguir (REZA, 1994):

$$H(X) \geq H(X|Y)$$

sendo válida a igualdade apenas quando X e Y são estatisticamente independentes. A interpretação é direta: a incerteza associada a uma variável aleatória é maior ou igual à incerteza associada a essa variável dado que se conhece uma segunda variável aleatória, já que essa segunda variável pode trazer alguma informação sobre a primeira. No "pior dos casos", a segunda variável aleatória não será relevante, valendo, destarte, a igualdade.

3.3. Entropia e Codificação de Fonte

Para exemplificar o uso do conceito de entropia no âmbito da teoria de codificação, lidaremos com o conceito de codificação de fonte. Trata-se, fundamentalmente, do problema de como "representar a informação" de modo parcimonioso.

Suponhamos que desejemos produzir um código que seja capaz de representar, efetivamente, a informação produzida por uma fonte que emite quatro símbolos, A, B, C e D, com probabilidades P_A, P_B, P_C e P_D, respectivamente. Nosso código será binário, ou seja, formado por palavras construídas como grupos de zeros e uns. Imaginemos agora que $P_A = 0{,}5$, $P_B = 0{,}25$ e $P_C = P_D = 0{,}125$.

Uma possibilidade natural seria usar palavras de dois bits, ou seja, por exemplo, "00" para representar o símbolo A, "01" para representar o símbolo B, "10" para representar o símbolo C e "11" para representar o símbolo D. Isso levaria a um

código com comprimento médio $L_{\text{médio}} = 2\ bits$. No entanto, será essa a forma mais eficiente?

Considere agora um código que explore o fato de haver diferenças entre probabilidades de geração dos símbolos. Para tanto, adotaremos a palavra "0" para representar o símbolo A, a palavra "10" para representar o símbolo B, a palavra "110" para representar o símbolo C e a palavra "111" para representar o símbolo D. Nesse caso, perceba que $L_{\text{médio}} = 1 \times 0{,}5 + 2 \times 0{,}25 + 3 \times 0{,}125 + 3 \times 0{,}125 = 1{,}75\ bits$. Obtivemos, portanto, um código mais parcimonioso para representar a mesma informação! Há alguns aspectos relacionados à questão de decodificabilidade que não explicitaremos aqui, mas os códigos apresentados no curso serão sempre consistentes nesse sentido.

O clássico trabalho de Shannon (1948) contém um teorema que fundamenta a questão dos limites para a codificação de fonte. Basicamente, o teorema diz que o

comprimento médio de um código que represente a informação trazida por uma fonte discreta será sempre maior ou igual à entropia dessa fonte. Em outras palavras (COVER & THOMAS, 2006):

$$L_{\text{médio}} \geq H_D(X)$$

sendo D a base dos logaritmos usados para calcular a entropia e, também, a cardinalidade do alfabeto usado para construir os códigos. Nos exemplos que demos acima, usamos códigos com palavras formadas por dígitos "0" e "1", ou seja, usamos $D = 2$. Portanto, a entropia "na base 2" será o limitante adequado. Perceba que, em nosso exemplo, $H_2(X) = 1{,}75$, ou seja, o código que apresentamos era um código ótimo! De fato, só há igualdade na condição (9) quando $P_i = D^{-L_i}$, sendo P_i a probabilidade associada à emissão do *i-ésimo* símbolo da fonte e L_i o comprimento da palavra usada para representar esse símbolo no contexto de um alfabeto *D-ário*.

A entropia da fonte, portanto, corresponde a um limite inviolável quando se trata de compactar dados (ou seja, "enxugar" a informação sem perdas). Para obter

códigos com comprimento médio menor que o dado pela entropia, será necessário admitir algum grau de distorção.

4. Informação Mútua e Divergência de Kullback-Leibler

Uma grandeza fundamental para nós é a informação mútua entre duas variáveis aleatórias X e Y. Essa grandeza, de certa forma, indica a redução no nível de incerteza associado a uma variável graças à informação trazida pela outra variável. Em outras palavras, a informação mútua associada a um par de variáveis X e Y é dada por:

$$I(X,Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

Em consonância com o que foi visto na seção 3.2, $I(X,Y) \geq 0$, com igualdade apenas se X e Y forem estatisticamente independentes. Isso é intuitivo: se X for independente de Y, não haverá redução de incerteza, pois não é trazida informação relevante! Note que a informação mútua se torna, portanto, *uma ferramenta poderosa*

para quantificar independência estatística, o que será deveras útil para nós na sequência do curso.

É possível mostrar que a informação mútua também pode ser dada pela expressão a seguir:

$$I(X,Y) = \sum_x \sum_y p(x,y) log_2 \left[\frac{p(x,y)}{p(x)p(y)} \right]$$

Essa definição pode ser interpretada de um modo muito interessante, mas, para isso, é preciso definir uma grandeza denominada *divergência de Kullback-Leibler*. A divergência de Kullback-Leibler entre duas funções de massa de probabilidade de uma variável, $p_1(x)$ e $p_2(x)$, é dada por:

$$D(p_1||p_2) = \sum_x p_1(x) log_2 \left[\frac{p_1(x)}{p_2(x)} \right]$$

Duas propriedades muito relevantes são: $D(p_1||p_2) \geq 0$ e $D(p_1||p_2) = 0$ se e somente se $p_1(x) = p_2(x)$. No entanto, não se pode dizer que se trata de uma *distância*, pois

não são satisfeitas a desigualdade do triângulo e a propriedade de simetria – note que $D(p_1||p_2) \neq D(p_2||p_1)$ (COVER & THOMAS, 2006).

Avaliando a expressão apresentada para a informação mútua, percebemos que ela, na verdade, pode ser vista como a divergência de Kullback-Leibler (num contexto de duas variáveis aleatórias – X e Y) entre $p(x,y)$ e o produto de funções de massa de probabilidade marginais $p(x)p(y)$. Quando as variáveis são independentes, temos exatamente que $p(x,y) = p(x)p(y)$, ou seja, a informação mútua quantifica, basicamente, o grau de dependência estatística entre as variáveis ao medir a divergência entre a função de massa de probabilidade conjunta e o produto das funções marginais.

4.1. Informação Mútua e Capacidade de Canal

A informação mútua, que será muito importante na sequência do curso, tem um papel fundamental na teoria de codificação. Uma razão para isso é que, avaliando a

informação mútua entre a entrada de um canal de comunicação e a sua saída, temos condição de avaliar determinados limites para o envio eficiente de mensagens. Esse ponto é simplesmente crucial para a moderna teoria de transmissão de dados.

No espírito do que foi dito acima, se associarmos a variável aleatória X à mensagem transmitida e a variável aleatória Y à mensagem recebida após a passagem por um canal de comunicação, teremos condições de definir a grandeza $I(X,Y)$, a qual, pelo que foi discutido acima, revela aspectos essenciais da atuação do canal. Por exemplo, imagine um canal tão destrutivo que, simplesmente, seja capaz de fazer com que as mensagens recebidas sejam independentes do que foi enviado: nesse caso, a informação mútua será nula, e podemos intuir que não haverá fluxo efetivo de informação do transmissor para o receptor. Por outro lado, caso Y seja bastante dependente de X, podemos julgar que é possível estabelecer algum grau de comunicação apropriada.

No clássico trabalho de Shannon (1948), define-se uma grandeza fundamental para a comunicação, a *capacidade de canal*, em termos da informação mútua. No cenário de transmissão que delineamos, pode-se dizer que a capacidade de canal C é dada por:

$$C = \max_{p(x)} I(X,Y)$$

Em outras palavras, a capacidade de canal revela, digamos, "no melhor caso", qual é a informação mútua entre mensagem transmitida e mensagem recebida. Ou seja, a capacidade de canal indica a condição de maior dependência (ou seja, de "melhor" fluxo de informação) entre transmissor e receptor, condição esta que deve ser atingida por meio da manipulação das probabilidades de envio *p(x)*. Isso é fundamental: manipula-se a informação mútua por meio da modificação das probabilidades de envio de símbolos, o que, intuitivamente, já revela que uma codificação eficiente pode "facilitar" a transmissão de dados. A capacidade de canal,

caso usemos a base dois para os logaritmos, será dada em bits/transmissão ou bits/uso do canal.

Vejamos a seguir alguns exemplos clássicos de cálculo de capacidade de canal para que possamos nos familiarizar com a ideia. O primeiro exemplo que daremos é o de um canal ideal, ou seja, um canal que não produz erros, como mostrado na Figura 2.

Figura 2 – Estrutura de um canal ideal.

Uma vez que $P(Y=1|X=1) = P(Y=0|X=0) = 1$ e $P(Y=1|X=0) = P(Y=0|X=1) = 0$, temos que, como o canal não gera erros, se $P(X=1) = P(X=0) = 0{,}5$, será possível enviar 1 bit de informação, sendo exatamente essa a capacidade.

O outro exemplo que daremos é o de um canal que produz erros, o canal binário simétrico (BSC, do inglês *binary symmetric channel*). Nesse caso, existe uma chance p de que um símbolo "0" seja recebido como "1" e de que um símbolo "1" seja recebido como "0", *i.e.*, $P(Y=1|X=1) = P(Y=0|X=0) = 1-p$ e $P(Y=1|X=0) = P(Y=0|X=1) = p$. A Figura 3 ilustra o comportamento do canal.

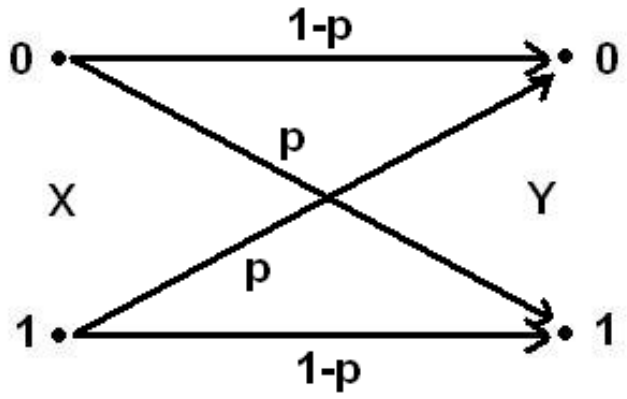

Figura 3 – Canal binário simétrico (BSC).

É possível mostrar (Cover & Thomas, 2006) que a capacidade desse canal é $C = 1 - H(p)$, sendo $H(p)$ a entropia de uma variável aleatória binária. A capacidade é atingida com uma distribuição equiprovável dos símbolos transmitidos. É bastante interessante perceber que, se $p = 0$, o BSC se torna o canal ideal visto anteriormente, mas, se $p = 0{,}5$, a capacidade será nula, já que, uma vez que a probabilidade de acerto se igual à probabilidade de erro, o receptor estará totalmente "às cegas". Em tal situação, existe uma completa destruição da informação transmitida.

Antes de encerramos a seção, é preciso mencionar um estupendo resultado exposto em (SHANNON, 1948). Basicamente, Shannon mostrou que todas as taxas menores que C são atingíveis no sentido de que, usando um processo adequado de codificação, é possível operar nessas taxas com probabilidade de erro arbitrariamente pequena. O resultado de Shannon é um resultado de existência no

sentido de que ele garante que haverá um processo de codificação capaz de garantir transmissão eficiente, mas não indica que processo será esse. De fato, em teoria de codificação, quando um código é proposto e aplicado a certa classe de tarefas, tipicamente se busca verificar quão próximo ele está do limite de Shannon.

Perceba que o teorema indica que a capacidade de canal, de certo modo, é uma espécie de "limite de velocidade" para comunicação segura. Caso violemos esse limite, estaremos transcendendo a capacidade de fluxo de informação do meio, o que acarretará erros inevitáveis. Metáforas possíveis são o ritmo de movimento numa avenida congestionada ou mesmo o escoamento de um fluido no gargalo. Por outro lado, o teorema diz mais: ele diz que, se não tentarmos violar o limite, *será possível, em tese, realizar comunicação confiável*, desde que a informação seja adequadamente codificada. Muito do que se faz em pesquisa de ponta nas diversas

áreas de transmissão de dados gravita em torno de formas de abordar essa tarefa de modo eficiente.

5. Variáveis Aleatórias Contínuas

Os conceitos fundamentais abordados continuam essencialmente válidos quando se lida com variáveis aleatórias que assumem valores reais. No entanto, é importante dizer algumas palavras, já que o caso de variáveis contínuas é aquele com que mais trabalharemos no restante do curso.

No caso contínuo, define-se a *entropia diferencial*† de uma variável aleatória como:

† Embora haja diferenças conceituais entre a definição de entropia apresentada anteriormente a ideia de entropia diferencial, usaremos a mesma notação para ambas as grandezas por uma questão de simplicidade. Sempre que lidarmos com variáveis discretas, estaremos falando da entropia e, sempre que falarmos de variáveis contínuas, estaremos falando da entropia diferencial.

$$H(X) = -\int_x p(x) ln[p(x)] dx$$

Ao contrário do que ocorria no caso discreto, a entropia diferencial pode assumir valores negativos. Não obstante, continua válida a propriedade:

$$H(X,Y) = H(X) + H(Y|X) = H(Y) + H(X|Y)$$

e a definição de informação mútua entre duas variáveis aleatórias é análoga:

$$I(X,Y) = \iint_{x,y} p(x,y) ln\left[\frac{p(x,y)}{p(x)p(y)}\right] dxdy$$

A informação mútua continua a ser uma divergência de Kullback-Leibler, já que esta também pode ser definida, com as mesmas propriedades fundamentais, para o caso contínuo. Portanto, $I(X,Y) = D\big(p(x,y)||p(x)p(y)\big)$.

5.1. Densidades de Máxima Entropia

Tendo em vista a definição de entropia diferencial, poderíamos perguntar: qual será a densidade de probabilidade de máxima entropia? Essa pergunta, por si só, não faz muito sentido, pois a entropia diferencial é ilimitada, mas, sob certas restrições, a questão se torna assaz interessante.

É possível obter as densidades de máxima entropia, por exemplo, sob um conjunto de restrições de momentos. Por exemplo, podemos desejar a densidade de máxima entropia que possua primeiro momento igual a m_1, segundo momento igual a m_2 e terceiro momento igual a m_3. Nesse caso, as restrições tornam o problema significativo e evitam qualquer divergência trivial.

No caso de restrição no primeiro e no segundo momentos, pode-se mostrar (COVER & THOMAS, 2006) que a densidade de máxima entropia é a gaussiana.

Por outro lado, se adotarmos uma restrição de outra natureza – supor que a densidade só possui valor não-nulo numa faixa de valores que vai de x_1 a x_2 – a densidade de máxima entropia é a uniforme.

Não nos estenderemos nessa análise aqui, mas é importante que tenhamos em mente o problema geral e as propriedades de densidades gaussianas e uniformes expostas, que serão úteis quando falarmos das relações entre teoria de informação e processamento de sinais.

6. Entropia Cruzada

Consideremos que $P(\cdot)$ e $Q(\cdot)$ sejam duas funções de massa de probabilidade ou que $p(\cdot)$ e $q(\cdot)$ sejam duas densidades de probabilidade. A entropia cruzada é:

$$H_C(P,Q) = E_P\{-\log Q\}$$

ou

$$H_C(p,q) = E_p\{-\log q\}$$

Vale a seguinte relação:

$$H_C(P, Q) = H(P) + D(P||Q)$$

Ou

$$H_C(p, q) = H(p) + D(p||q)$$

Uma propriedade muito importante é que minimizar a entropia cruzada com respeito a Q equivale a minimizar a divergência $D(\cdot)$, pois o termo $H(P)$ não é afetado (GOODFELLOW ET AL., 2016).

7. Referências bibliográficas

T. M. COVER & J. A. THOMAS, **Elements of Information Theory**, Wiley, 2ª ed., 2006.

GOODFELLOW, I., BENGIO, Y., COURVILLE, A. **Deep Learning**, MIT Press, 2016.

F. M. REZA, **An Introduction to Information Theory**, Dover, 1994.

C. E. SHANNON, "A Mathematical Theory of Communication", *Bell System Technical Journal*, No. 27, pp. 379-423, 623-656, 1948.