Chapter 1

Estudos Iniciais

1.1 Ideias e outras coisas

(Qual a possibilidade de conseguirmos fazer uma regressão, por exemplo, e irmos estimando e atualizando os parametros por meio da informação de fisher?) Estava pesquisando sobre as limitações e achei interessante o artigo 'Notes on the Limitations of the Empirical Fisher Approximation' no qual ele delimita algumas coisas quanto o calculo e aproximações da informação de fisher, em principal quanto seu não uso por falta de um método conveniente para que seja usado, principalmente na area de deep learning. Ademais uma boa aproximação para que seja calculado, com certo grau de precisão e utilidade aparentemente se falta. Eles mostram que enquanto pode ser calculado muitas vezes o valor bruto até rapidamente, quem sabe uma aproximação não seja util, ou talvez tentar traze-lo para o mundo dos computadores com um maior peso.

Kullback–Leibler divergence/Hellinger distance

1.2 Informação de Fisher

Inicialmente, vamos imaginar a seguinte situação: Eu lhe dou um grafico de uma distribuição normal, onde você não sabe onde ela esta centrada, porem sabe qual sua variancia. A partir do grafico dela, voce conseguiria deduzir, ou talvez obter alguma informação sobre o valor real de $\mu^\star$? Observe às seguintes curvas e tente achar suas medias Em uma é muito mais facil identificar onde possivelmente o valor $\mu^\star$ está, mas digamos que precisamos saber o da segunda distribuição e em um caso mais geral, para qualquer distribuição. Que valor poderiamos tentar atribuir a quantidade de informações que podemos tirar de uma base de dados? A definição da informação de fisher, em dados matematicos

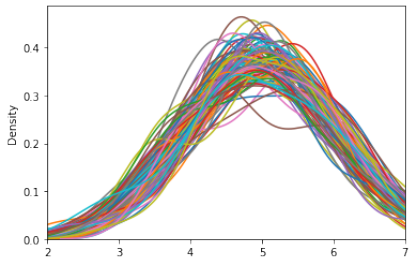

Figure 1.1: Normal com variância 1

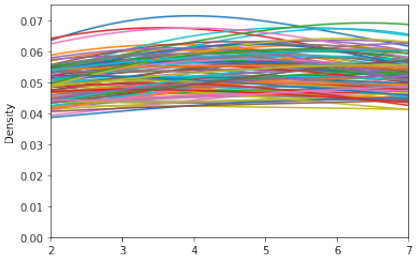

Figure 1.2: Normal com variância 7

é dado por

$$\mathcal{I}_X(\theta) = \begin{cases} \sum_{x \epsilon \chi} (\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x|\theta))^2 p_\theta(x), & \text{se X não é continuo ;} \\ \int_\chi (\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x|\theta))^2 p_\theta(x) \mathrm{d}x, & \text{se X é continuo ;} \end{cases} \tag{1.1}$$

O valor $\frac{\mathrm{d}}{\mathrm{d}\theta} \log f(x|\theta)$ é conhecido como a função score, que descreve quão sensitivo é o modelo a uma mudança no θ em um θ particular. A informação de fisher mede a sensitividade da relação entre f e mudanças no θ pesando na a sensitividade a cada valor de x em respeito a probabilidade definida por $p_\theta f(x|\theta)$ Esse peso em relação a $p(\theta)$ faz com que a informação de fisher seja em relação ao valor esperado de θ. Podemos definir a inforção de fisher em relação a esperança. Vamos assumir algumas coisas para depois conferirmos se estamos corretos. Numa normal, sua variancia não pode ser zero, mas quanto mais perto de zero estivermos, maior a informação que vamos ter. Quanto maior nossa variancia, menor a quantidade de informações que vamos conseguir tirar do nosso parametro. Se nosso modelo seguir isso, estamos no caminho certo

$$\mathcal{I}_X(\theta) = -E(\tfrac{\mathrm{d}^2}{\mathrm{d}x^2} \log f(x|\theta)) = \begin{cases} -\sum_{x \epsilon \chi} (\tfrac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x|\theta))^2 p_\theta(x), & \text{se X não é continuo ;} \\ -\int_\chi (\tfrac{\mathrm{d}^2}{\mathrm{d}\theta^2} \log f(x|\theta))^2 p_\theta(x) \mathrm{d}x, & \text{se X é continuo ;} \end{cases} \tag{1.2}$$

Vamos fazer um exemplo para a Normal.

Sabendo que a equação de uma normal é dada por

$$N(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp^{\frac{(x-\mu)^2}{2\sigma^2}}$$

sua função logaritmo é dado por

$$l(\theta) = -\frac{1}{2}\ln(\sigma^2) - \frac{(x-\mu)^2}{2\sigma^2} + c$$

Como ja temos σ, nosso θ onde queremos obter informaçã é o μ

Logo, colocando $\mu = \theta$ e achando nossas funções

$$l'_\theta(\theta) = \frac{x-\theta}{\sigma^2}$$
$$l''_\theta(\theta) = -\frac{1}{\sigma^2}$$

Fazendo a esperança, temos

$$-E(l''_\theta(\theta)) = \frac{1}{\sigma^2}$$

Então ela depende apenas da variancia da nossa distribuição. Se observamos os graficos 1.1 e 1.2, observamos que na primeira é mais facil tentar assumir o valor de μ enquanto na segunda não tanto. Isso condiz com o comportamento do grafico e seus respectivos valores de variancia. Então essa nossa discussão foi feita corretamente

1.3 Informação de Fisher e Flips ecologicos (Artigo original)

1.3.1 Método e equações

Um dos pontos principais em que precisamos nos apoiar é a ideia de que em sistemas que estão proximos ou em um regime dinamico estatico, possuem valores de variabilidades fixos em suas variaveis de estados e qualquer mudança no sistema é manifestada por meio de mudança nas variabilidades. Ou seja, se tivermos um vetor de variaveis dados por

$$\mathbf{X} = \begin{pmatrix} x_1 & x_2 & \dots & x_n \end{pmatrix}$$
$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_1 & \sigma_2 & \dots & \sigma_n \end{pmatrix}$$

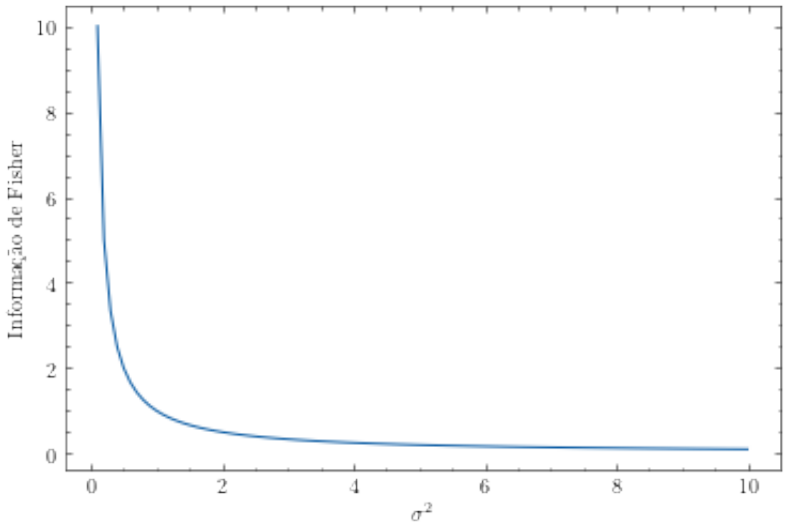

Figure 1.3: Infomação de Fisher por Sigma

Onde $\mathbf{X}$ é o vetor contendo todos as variaveis do regime e $\mathbf{\Sigma}$ é o vetor contendo todas suas variabilidades, com valores iguais a $\sigma_1, \sigma_2, \ldots, \sigma_n$. Se o regime fosse estados, esses valores se materiam constantes, ou em um padrão. Caso ele esteja mudando os valores de σ mudariam tambem. Dado isso, vamos aplicar a informação de fisher. Utilizando a formula:

$$I = \int \frac{1}{p(\varepsilon)} (\frac{\mathrm{d}p(\varepsilon)}{\mathrm{d}\varepsilon})^2 \mathrm{d}\varepsilon \tag{1.3}$$

Aonde $p(\varepsilon)$ é a distribuição de probabilidades do erro,distancia ε do valor real. Assumindo que o sistema pode ser descrito como dinamico continuo e está num ciclo periodico estavel, podemos definir a informação de fisher, depois de uma certa quantia de matematica (fazer isso depois, Marco.), Como:

$$I = \frac{1}{T} \int_0^T \frac{(R''(t))^2}{(R'(t))^4} \, \mathrm{d}t \tag{1.4}$$

Isso para apenas um ciclo do sistema. Na forma vetorial

$$I = \frac{1}{T} \int_0^T \frac{(\mathbf{x}'(t))^T (\mathbf{x}''(t))^2}{\|\mathbf{x}'(t)\|^6} \, \mathrm{d}t \tag{1.5}$$

Fazendo uma interpretação. Sendo $\mathbf{x}'(t)$ a velocidade do nosso sistema e similarmente $\mathbf{x}''(t)$ como a aceleração do sistema e pelo nosso desenvolvimento de como chegamos nela em primeiro lugar, estamos medindo a variabilidade do nosso sistema durante o tempo que ele passa em cada seção da sua trajetoria. Pode ser visto tambem como a mudança de velocidade ao longo da trajetoria.

Em um sistema de velocidade 0, teriamos uma informação infinita, pois qualquer ponto que pegarmos, ele contera todas as informações, pois ele só teria o mesmo estado. Num sistema sem aceleração, de velocidade constante, nossa informação é 0, pois se fossemos olhar um ponto da trajetorio no ciclo 0 e no ciclo 1, estariamos vendo o mesmo estado nos dois ciclos. Para avaliarmos nossa expressão, uma solução analitica é inviavel, numa grande parcela das vezes.

Podemos, então, nos voltar para uma solução numerica que nos dará uma boa solução. A primeira coisa que precisamos fazer é uma estimativa da velocidade e da aceleração do sistema. Como muitas vezes o intervalo de tempo entre a coleta de dados, no mundo real, não está em intervalos regulares, podemos continuar com esses intervalos irregulares, ou podemos interpolar para acharmos dados que são lineares. A estimativa é feita usando 3 pontos, onde escolhemos um que é central, um a uma distancia Δt_p e a distancia Δt_a, pela formula $t - \Delta t_p$ e $t\Delta t_a$ e esses intervalos estão relacionados por $\frac{\Delta t_a}{\Delta t_b} = \alpha$. Pela nossa definição de derivada, temos

$$\mathbf{f}'(t) = \frac{\alpha^2\mathbf{f}(t + \Delta t_a) - (\alpha^2 - 1)\mathbf{f}(t) - \mathbf{f}(t - \alpha\Delta t_a)}{(\alpha^2 + \alpha)\Delta t_a} \tag{1.6}$$

e aceleração

$$\mathbf{f}^{(\prime\prime)} = \frac{\alpha\mathbf{f}(t + \Delta t_a) + \mathbf{f}(t - \alpha\Delta t_a) - (\alpha + 1)\mathbf{f}(t)}{(\alpha^2 + \alpha)\Delta t_a^2/2} \tag{1.7}$$

1.3.2 Discussão e interpretação do modelo

Essas estimativas podem gerar alguma especie de barulho e outros dados. Um jeito de arrumar isso é usando uma média de dois pontos, inves de necessariamente apenas dos valores calculados. Mas o proprio calculo de fisher já da uma ajuda em relação a isso. Ademais, não sabemos a peridiciocidade do nosso sistema, e quando não sabemos isso, nossa informação vai ficar flutuando entre os calculos. Para isso podemos fazer uma estimativa do periodo do sistema, utilizando uma FFT, ou conhecimentos previos. Mas a informação de fisher é muito sensitiva se o sistema não esta em forma periodica, ou está sobre ações de outras forças ela vai flutuar. Podemos claro aplicar a sistemas não periodicos, mas sua interepretação sera dificil (Sera que isso ainda é verdade? Pesquisar depois).

Para observamos a transição, podemos simplesmete avaliar a integral em um periodo, colocando o valor como o valor do periodo todo. Isso da uma estimativa grosseira, digamos assim, onde ela vem, geralmente, em formato de blocos. Ou podemos fazer uma quase convolução, onde fazer a média movel, colocando o valor da integral como um unico ponto no espaço, onde a informação de fisher usa uma janela de tamanho equivalente ao periodo, centrada no ponto, posteriormente mudando nossa janela para o proximo ponto e assim sucessivamente. Os resultados realmente mostram uma mudança na informação de fisher entre um periodo e outro, então é sim um método válido.

(Não sei dizer se necessariamente é quando muda a informação e como eles definem ser transição ou quando ele muda pra outro estado ciclo, talvez seja algo em relação a ecologia, ou já seja algo definido e na aplicação real a gente so conseguiria definir que esta mudando, sem necessariamente saber se esta indo para transição ou para outro estavel.)

1.3.3 Conclusões finais

Alguns pontos importantes a serem reforçados é que por mais que sejam uteis, a pesquisa ainda estão incompletas (isso é so uma preliminar do primeiro artigo, posteriormente adicionarei sobre novas biografias). Um dos pontos que precisam ser estudados é a regularização do intervalo de integração da informação, pois diferenças nesses tempos podem levar a graficos que são de dificil interpretação. Ademais, para sistemas aciclicos, não é de muita utlidade esse informação, pois ela variaria muito. Para sistemas muitos complexos, com muitas variaveis, ainda é meio dificil e não tão bom. Geralmente ela se da melhor com escalas temporais e sistemas um pouco menos complexos. O valor em si da informação ainda não é muito bem definido, valores altos indicam um sistema mais estavel, porem valores proximos necessariamente indicariam regimes parecidos, ou apenas com variabilidades parecidas? Quantas variaveis deveriamos considerar? Para escalas de tempo pequenas, ainda vamos conseguir usar isso para detectar? A informação ainda é extremamente sensitiva a forças internas, qualquer pertubação e casos em que não necessariamente é ciclico, já existe algum modelo que seja mais robusto a isso? Claramente a informação é muito util, porem é preciso ver seu desenvolvimento.

Chapter 2

Importancia

Mais do que nunca no mundo vem se falando sobre a importancia dos nossos sistemas ecologicos. É inegavel o profundo impacto que fizemos no planeta, principalmente nos ultimos anos e agora estamos começando a lidar com as consequencias. Não é segredo que alteramos muito a ordem natural dos ecossistemas, seja de forma negativa ou de forma positiva. imaginando a situação em que queremos recuperar um certa área que foi afetada, seria muito util termos uma métrica ou uma ideia de se está funcionando, se o que está sendo feito está realmente sendo util. Claro que, teoricamente, apenas visualmente poderiamos tentar imaginar, mas algo que seja mais certeiro é desejado. Se selecionarmos uma certa quantia de variaveis, como por exemplo, tamanho populacional de uma certa especie, quantidade de um certo nutriente no solo, massa de liquens nas arvores, dentre outras, uma métrica que pudesse avaliar a evolução das variaveis com o tempo e que nos dissese com um numero, ou algo similar, que a evolução está acontecendo seria de grande ajuda.

Identificação de mudanças em sistemas ecologicos tem algo que tem crescido muito nos ultimos e o uso de estatistica, principalmente na sua forma mais moderna, pode se tornar algo fundamental para essa área.(pensar em algo pra escrever melhor. Não muito bom)

Chapter 3

Uma analise profunda da informação de Fisher

3.1 Um passo para trás

Eu acho que é importante para realmente entender e fazer as corretas derivações necessarias, ter o conhecimento de infomações que precedem o nosso objetivo final, tanto para realmente entender, não so copiar uma formula, tanto para entender de onde possivelmente são os erros e possiveis locais de melhoria. Então por enquanto, farei uma pesquisa sobre as materias que levam à informação de fisher.

3.1.1 Estatistica suficiente

Colocando que temos um conjunto de dados $X_1, X_2, \ldots, X_n$ gerados por uma função de probabilidade qualquer com parametro θ, se existir uma estatistica $Y = u(X_1, X_2, \ldots, X_n)$, tal que a condicional $P(X_1, X_2, \ldots, X_n | Y = y)$ não depender de θ. Exemplo:

Dado uma distribuição de bernoulli, com parametro θ desconhecido, onde se fazem n testes tal que existam n dados. Uma estatistica suficiente para esse parametro é a soma de todos os X_i, pois:

$$Y = \sum_i X_i$$

$$\begin{aligned} P(X_1, X_2, \ldots, X_n | Y = y) &= \frac{p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, Y = y)}{P(Y = y)} \\ &= \frac{P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n, \sum_i X_i = x_i)}{P(\sum_i X_i = x_i)} \end{aligned}$$

Não dependendo do parametro θ. Ou seja, conseguimos toda informação possivel sobre o parametro, apenas com essa estatistica, pois independente dele,

adicionando, retirando dados, nenhuma informação a mais sobre o parametro será ganhada.

Infelizmente, essa condição não nos da o necessario para calcularmos que estatistica será essa, então ainda é um grande de um trabalho imaginarmos isso. Por sorte, temos um teorema que nos ajudará nesse quesito, o *Teorema da Fatorização* em que ele fala que se uma estatistica $Y = u(X_1, X_2, \ldots, X_n)$, so é suficiente se e apenas se, a distribuição $f(x_1, x_2, \ldots, x_n; \theta)$, que gerou dos dados, pode ser fatorizada em dois fatores, ϕ, h, da forma $f(x_1, x_2, \ldots, x_n; \theta) = \phi(u(x_1, x_2, \ldots, x_n); \theta)h(x_1, x_2, \ldots, x_n)$, onde ϕ que é dependente dos dados apenas por meio de $u(x_1, x_2, \ldots, x_n)$ e h é independente do parametro. Vamos à um exemplo: Suponha uma bernoulli de parametro θ desconhecido e constante. Coletando n dados de forma independente, podemos achar nossa estatistica de modo

$$\begin{aligned} f(x_1, x_2, \ldots, x_n; \theta) &= f(x_1, \theta) * f(x_2, \theta) * \ldots * f(x_n, \theta) \\ &= \theta^{x_1}(1-\theta)^{1-x_1} * \theta^{x_2}(1-\theta)^{1-x_2} * \ldots * \theta^{x_n}(1-\theta)^{1-x_n} \\ &= \theta^{\sum_i x_i}(1-\theta)^{n-\sum_i x_i} \end{aligned}$$

O parametro θ apenas interage com a estatistica e a função h é uma constante de valor 1.

3.1.2 Maxima Verossemelhança

Outro tópica que sera discutido e posteriormente usado será o de maxima verossemelhança. Esse conceito é muito usado em questão de estimação de parametros e estatistica Bayesiana e nos servira bem para o proposito final.

Supondo que temos uma certa quantia de dados, vindos de uma distribuição qualquer, com um parametro θ, fixo, com valores podendo estar num intervalo fechado ou não. Idealmente, gostariamos de tentar achar a curva de melhor encaixe nesses dados. Por facilidade, vamos colocar que sabemos que eles se originam de uma curva normal de parametros desconhecidos. Sabemos a priori que a curva normal está centrada na média e sua dispersão e peso da cauda relaciado à sua variacia. Seus parametros pode assumir qualquer valor na reta dos reais, então, em tese, teriamos infinitas curvas que se encaixariam nos nossos dados. Porem, podemos supor, até mesmo empiricamente, que alguma delas tem que se encaixar a melhor, até por que nossos dados foram tirados dessa curva com parametros fixos. Como poderiamos tentar determinar, ou testar, se algum parametro é melhor do que outro?

Para isso, temos a ideia das chances e da maxima verossemelhança. Vamos começar com as chances. Elas, diferente do entendimento popular, são diferentes das probabilidades. Elas não seguem as mesmas regras, de que, por exemplo, tem de ser somada 1 no seu inteiro. Alem disso, elas se diferem de uma probabilidade condicional por uma constante K, da forma L(θ)=K*P($X = x_i|\theta$). Onde θ seja nosso parametro, ou parametros. Sozinha, não temos muita interpretação passa essa chance, dado a constante K que não tem valor de interpretação para nós. Então, ela vem sempre a titulo de comparação para nos. Nossa chance,

é em relação ao parametro, então é justo assumir que uma comparação entre duas chances, na forma de razão, nos diria o quanto uma é superior a outra. Supondo por exemplo, que a comparação $\mathrm{L}(\theta_1)/\mathrm{L}(\theta_2\ )$ nos de um valor de 1.5. Significaria que os dados tem 1.5 vezes maior probabilidade de estarem sobre o parametro θ_1 do que do parametro θ_2.

Isso nos leva a pensar, existiria algum parametro tal que sua razão de chances sempre dê um valor maior que 1? Sim, isso é dito pela Lei da verossimilhança. *Dentro de um modelo estatistico, um conjunto de dados sempre vai se encaixar melhor em um parametro do que em outro, tal que as chances do primeiro parametro sempre vai ser maior que a do segundo.*

3.1.3 Chances

Enquanto ja é um começo, isso não nos da uma função clara de como achar esse parametro. Para isso, temos alguns meios de seguirmos. O primeiro, usando um pouco de calculo, é achar o maximo e minimo de funções. Como temos uma função das chances pelo valor de θ podemos tentar achar o nosso pico da função. Lembrando de calculo, seria derivar e igualar a zero. Porem, é muitas vezes mais facil achar o maximo do log da função. Maximizar o log da função é igual a maximizar a função. Para uma binomial, por exemplo, se quisermos estimar o parametro θ fazemos

$$P(X = x_i) = \binom{x}{n}\theta^x * (1-\theta)^{n-x}$$

$$\log(P(X = x_i)) = \log(\binom{x_i}{n}) + x\log(\hat{\theta}) + (n-x)\log(1-\hat{\theta})$$

$$\frac{\mathrm{d}P}{\mathrm{d}\hat{\theta}} = \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}\log(\binom{x_i}{n}) + \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}x\log(\hat{\theta}) + \frac{\mathrm{d}}{\mathrm{d}\hat{\theta}}(n-x_i)\log(1-\hat{\theta})$$

$$0 = \frac{x_i}{\hat{\theta}} - \frac{n-x_i}{1-\hat{\theta}}$$

$$n\hat{\theta} - x\hat{\theta} = x - x\hat{\theta}$$

$$\hat{\theta} = \frac{x}{n}$$

Se nossa distribuição fosse multimodal, uma solução explicitamente analitica talvez não fosse possivel, então metodos numericos teriam de ser utilizados.

3.1.4 Desigualdade de Cramer-Rao

Se tivermos um conjunto de dados todos retirados de uma mesma função de probabilidade na forma $f(\vec{x}, \theta)$, onde se tem n dados e 1 parametro. Se cada dado for independente, pela regra de probabilidade, sua distribuição conjunta é $f_{\mathbf{X}}(\vec{x}, \theta)$. Similarmente, seu produto de chances é dado por $L_{\mathbf{X}}(\theta) = \prod_{i=1}^{n} = f_{\mathbf{X}}(x_i, \theta)$. Supondo agora, que $L_{\mathbf{X}}(\theta)$ seja diferenciável em todo espaço e não dependa de θ, mais especificadamente, não tenha suporte em θ. Se integramos

a função de chances, $L_{\mathbf{X}}$ em todo espaço, teremos o valor de 1. Isso é dado que ela é so o produtorio de todas as funções de densidade. O produto de varias funções de densidade nos da uma função de densidade e por definição, em todo espaço, sua probabilidade é 1. Ou seja

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = 1 \tag{3.1}$$

Derivando-a, parcialmente em θ temos:

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = 0 \tag{3.2}$$

Como definimos que não há suporte em θ nas nossas distribuições, podemos colocar a diferencial dentro da integral.

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \tag{3.3}$$

fazendo uma certa manipulação albegrica, temos

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \frac{1}{L_{\mathbf{X}}(\theta)}\frac{\partial}{\partial\theta}L_{\mathbf{X}}(\theta)L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\ln(L_{\mathbf{X}}(\theta))\cdot L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = (\theta) \tag{3.4}$$

$$\theta\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\ln(L_{\mathbf{X}}(\theta))\cdot L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \tag{3.5}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \frac{\partial}{\partial\theta}\ln(L_{\mathbf{X}}(\theta))\cdot L_{\mathbf{X}}(\theta)\cdot\theta\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \tag{3.6}$$

Guardemos essa nossa equação. Vamos trabalhar agora estimando nosso parametro θ. Seja $\hat{\theta}$ um estimar do meu parametro orginal, tal que $E(\hat{\theta}) = \theta$. Por definição de esperança, podemos reescrever a equação como:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \hat{\theta}L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \theta \tag{3.7}$$

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \hat{\theta}L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \frac{\partial}{\partial\theta}\theta \tag{3.8}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \hat{\theta}\frac{\partial}{\partial\theta}L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = 1 \tag{3.9}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty} \hat{\theta}\frac{\partial}{\partial\theta}\log(L_{\mathbf{X}}(\theta))\,\mathrm{d}x_1\,\mathrm{d}x_2\,...\mathrm{d}x_n = \tag{3.10}$$

Subtraindo as equações (3.10) e (3.6)

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}(\hat{\theta}-\theta)\frac{\partial}{\partial\theta}\log(L_{\mathbf{X}}(\theta))\,\mathrm{d}x_1\,\mathrm{d}x_2\,\dots,\mathrm{d}x_n = 1 \tag{3.11}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}(\hat{\theta}-\theta)\frac{\partial}{\partial\theta}\log(L_{\mathbf{X}}(\theta))\,\mathrm{d}x_1\,\mathrm{d}x_2\,\dots,\mathrm{d}x_n = \tag{3.12}$$

$$(\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}\sqrt{L_{\mathbf{X}}(\theta)}(\hat{\theta}-\theta)\frac{\partial}{\partial\theta}\log(L_{\mathbf{X}}(\theta))\sqrt{L_{\mathbf{X}}(\theta)}\,\mathrm{d}x_1\,\mathrm{d}x_2\,\dots,\mathrm{d}x_n)^2 = \tag{3.13}$$

Lembrando de algebra linear, tal forma é reconhecida como um produto escalar entre dois vetores onde podemos, então, separa-los e aplicar a regra de cauchy-schwartz

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}L_{\mathbf{X}}(\theta)(\hat{\theta}-\theta)^2\,\mathrm{d}x_1\,\mathrm{d}x_2\,\dots,\mathrm{d}x_n\cdot \tag{3.14}$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\dots\int_{-\infty}^{\infty}(\frac{\partial}{\partial\theta}\log(L_{\mathbf{X}}(\theta)))^2L_{\mathbf{X}}(\theta)\,\mathrm{d}x_1\,\mathrm{d}x_2\,\dots,\mathrm{d}x_n \geq 1$$

A primeira parte reconhecemos como a variancia dos nossos dados. A segunda parte, nada mais é que a definição da informação de fisher, onde é o valor esperado, esperança, da função score

$$Var(\theta)\cdot E[\frac{\partial}{\partial\theta}\ln(L(\theta)^2)L(\theta)] \geq 1 \tag{3.15}$$

Alem de mostrar de onde surge a equação de fisher, foi mostrado tambem outro ponto muito importante, que foi a desigualdade de Cramer-Rao. Essa desigualdade mostra que a informação de fisher é inversamente proporcional ao erro, ou variancia do nosso modelo. Ou seja, com uma informação de fisher extremamente alta, nosso erro sera extremamente baixo e vice-versa

Falando um pouco mais sobre a função score. Ela é definida da seguinte forma:

$$S(\theta) = \frac{\partial}{\partial\theta}\ln L(\theta) \tag{3.16}$$

Seu valor tende a 0 quando nosso parametro estimado $\hat{\theta}$ tende a θ. Vamos achar

sua esperança.

$$E_\theta S(\theta) = \int_{-\infty}^{\infty} p_\theta(x) S(\theta)\,\mathrm{d}x \tag{3.17}$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} \ln L(\theta) p_\theta(x)\,\mathrm{d}x \tag{3.18}$$

$$= \int_{-\infty}^{\infty} \frac{\frac{\partial}{\partial\theta} L(\theta)}{L(\theta)} p_\theta(x)\,\mathrm{d}x \tag{3.19}$$

$$= \int_{-\infty}^{\infty} \frac{\partial}{\partial\theta} L(\theta)\,\mathrm{d}x \tag{3.20}$$

$$= \frac{\partial}{\partial\theta} \int_{-\infty}^{\infty} p_\theta(x)\,\mathrm{d}x \tag{3.21}$$

$$= 0 \tag{3.22}$$

Sua variancia, por sua vez é definida pela esperança da informação de fisher, $\mathcal{I}$, dado o valor real de θ.

$$\mathcal{I} = E[\frac{\partial}{\partial\theta} \ln(L(\theta)^2)] = var_\theta S(\theta) \tag{3.23}$$

$$\mathcal{I} = \int_{-\infty}^{\infty} (\frac{\partial}{\partial\theta} \ln L(\theta))^2 L(\theta)\,\mathrm{d}x \tag{3.24}$$

Ela é util para todos os valores possiveis de θ, enquanto a informação de fisher observada, dada pelas equações nos capitulos inicias so é util nas vizinhanças de $\hat{\theta}$. De forma mais simples, $\mathcal{I}$ é uma função de θ para todos os valores admissiveis de θ, dando uma média, por se dizer, atraves de todos os set de dados. Esse valor nos da um valor geral para a curvatura da função score. O valor observado de fisher, nosso I comum, nos da apenas para um unico dataset. O valor esperado, nosso $\mathcal{I}$ nos diz quão dificil vai ser estimar nosso parametro, valores maiores implicam maior facilidade de estimação. Para um modelo de Cauchy, dado a densidade de probabilidade $p(x) = [\pi(1 + x - \theta)^2]^{-1}$

$$I(\theta) = -\sum_i 2\frac{[(x_i - \theta)^2 - 1]}{[(x_i - \theta)^2 + 1]^2}$$

$$\mathcal{I} = -2nE_\theta \frac{(X_1 - \theta)^2 - 1}{[(X_1 - \theta)^2 + 1]^2} = \frac{n}{2}$$

$I(\theta) \neq \mathcal{I}$, qual dos dois é melhor escolher? $I(\theta)$ para $n = N$ nos diz uma probabilidade de $I(\hat{\theta})$ cair dentro de um intervalo. $\mathcal{I}$ nos diz a curvatura média. Qual é mais relevante? Posteriormente retornarei a esse problema.

Voltando a ideia da desigualdade de Cramer-Rao. Demonstrado que ela é valida no caso de $E(\hat{\theta}) = \theta$, mas se formos para um caso mais geral, onde $E(\hat{\theta}) = g(\theta)$, como nossa desigualdade fica? Bom, como tivemos que derivar e elevar ao quadrado para chegarmos na desigualdade, sendo $g(\theta)$ diferenciavel, teremos

$$Var(\theta) \cdot E(\frac{\partial}{\partial\theta} \ln(L(\theta))^2) \geq (g'(\theta))^2 \tag{3.25}$$

3.1.5 Uma analise sobre as diferenças entre I e $\mathcal{I}$

Como foi colocado no capitulo anterior, existe uma diferença entre esses dois valores e irei me debruçar um pouco mais sobre suas diferenças. I é definido como $-\frac{\partial^2}{\partial\theta^2}\ln L(\theta)$ Em termos matematicos, analisando seu significado geometrico e sua derivada. Como isso, com valores para todos as variaveis, nos da um escalar, essa informação nos da o valor da curvatura da função score em um determinado θ. Valores de pico, muito altos, nos dizem que se tem uma menor incerteza quanto ao parametro, valores proximos a zero nos dizem que há uma grande incerteza quanto ao valor, indicando maior incerteza. Ele é estimado para o valor de maxima verossimilhança (MLE). O MLE, na teoria de estimativa, nos conta uma estimativa no qual se tem maior chance do nosso parametro está. Para perguntas do tipo, com um conjunto de dados, qual a melhor estimativa, geralmente o MLE nós da uma resposta bem decente. I geralmente, vem junto com o estimador $\hat{\theta}$ então se é util acha-lo. Para isso, podemos resolver a equação $S(\theta) = 0$, lembrando que $S(\theta) = \frac{\partial}{\partial\theta}\ln(L(\theta))$. Onde achar essa solução nos da o MLE, ou seja, o $\hat{\theta}$. De forma mais importante, essa informação, dependente do estimador, varia de dataset a dataset. Pensando numa normal, por exemplo, se fossemos tirar 100 dados, 5 vezes da mesma distribuição, a probabilidade da média ser igual é praticamente zero. Então, o valor de I seria diferente para cada uma delas. So seria uma util uma analise nas redondezas de $\hat{\theta}$. Essa informação de fisher é local e sensivel a mundaças de coordenadas.

Por sua vez $\mathcal{I}$ é definida como a varianca da função score, ou em outros termos, $E_\theta(\frac{\partial}{\partial\theta}L(\theta))^2$, ou, $-E_\theta(\frac{\partial^2}{\partial\theta^2}L(\theta))$. Repare nas diferenças, $\mathcal{I}$ é o valor esperado da informação de fisher. $\mathcal{I}$ é a esperança de I. Então, se I era a curvatura da função score em um determinado $\hat{\theta}$, $\mathcal{I}$ vai nos dizer a curvatura média da função. Ademais, ele apresenta algumas propriedades mais interessantes. Ele é invariante quanto a mudanças de coordenadas, ou seja, se nossa função $p(y|\theta)$ puder ser descrita como $p_X(y-\theta)$, ambas nos darão o mesmo valor de $\mathcal{I}$. Para uma transformação de parametros, tal que $\phi = g(\theta)$ para qualquer distribuição $g()$. A função score de ϕ é dada por.

$$S(\phi) = \frac{\partial}{\partial\phi}\ln L(\theta) \tag{3.26}$$

$$= \frac{\partial\theta}{\partial\phi}\ln L(\theta) \tag{3.27}$$

$$= \frac{\partial\theta}{\partial\phi}S(\theta) \tag{3.28}$$

E sua informação de fisher $\mathcal{I}$ é

$$\mathcal{I}(\phi) = varS(\phi) \tag{3.29}$$

$$= (\frac{\partial\theta}{\partial\phi})^2\mathcal{I}(\theta) \tag{3.30}$$

$$= \frac{\mathcal{I}(\theta)}{(\partial\phi/\partial\theta)^2} \tag{3.31}$$

Para uma poisson:

$$\mathcal{I}(\theta) = \frac{n}{\theta}$$
$$\mathcal{I}(\ln\theta) = \frac{n/\theta}{1/\theta^2} = \theta$$

Matrizes de informação de Fisher

Até dito momento so analisamos a informação de fisher para um parametro θ, ou seja, em uma normal, assumimos que conheciamos a variancia σ^2, mas, se por exemplo, não conhecermos nem a media μ nem a variancia σ^2 como podemos descrever a informação observada e esperada de fisher? Bom, para isso temos nossa matriz, a qual podemos colocar nossas informações nela. A informação de fisher observada, na sua forma matricial pode ser dada por:

Seja $\vec{\theta}$ o vetor coluna que contem todos os parametros variando de 1 a n. Para a i observação, a informação de fisher pode ser descrita como

$$F_i(\vec{\theta}) = (\frac{\mathrm{d}}{\mathrm{d}\vec{\theta}} \ln L(\vec{\theta}))(\frac{\mathrm{d}}{\mathrm{d}\vec{\theta}} \ln L(\vec{\theta})^\top) \tag{3.32}$$

Onde $\frac{\mathrm{d}}{\mathrm{d}\vec{\theta}} \ln L(\vec{\theta})$ é um vetor coluna n x 1, onde nossa matriz de informação sera n x n. Para a informação esperada, basta apenas fazer a esperança da matriz.

$$F_i(\vec{\theta}) = E[(\frac{\mathrm{d}}{\mathrm{d}\vec{\theta}} \ln L(\vec{\theta}))(\frac{\mathrm{d}}{\mathrm{d}\vec{\theta}} \ln L(\vec{\theta})^\top)] \tag{3.33}$$

Para o caso de derivadas de segunda ordem. Temos

Para a observação m, n temos:

$$F_i(\vec{\theta}) = -\frac{\partial^2}{\partial\theta_m \partial\theta_j} \ln L(\theta) \tag{3.34}$$

Similarmente para a esperada

$$F_i(\vec{\theta}) = -E(\frac{\partial^2}{\partial\theta_m \partial\theta_j} \ln L(\theta)) \tag{3.35}$$

Pensando agora em uma limitação dela, se fossemos pegar a definição integral da informação de fisher, $\int_{-\infty}^{\infty} (\frac{\partial}{\partial\theta} \ln L(\theta))^2 * L(\theta)\,\mathrm{d}x$, $\frac{\partial}{\partial\theta} L(\theta) = \frac{f'(x|\theta)}{f(x|\theta)}$, essa função ela é indefinida se $f(x|\theta) \to 0$, o que necessariamente ocorre dado um certo x. Para contornarmos essa situação podemos fazer uma substituição

$p(x) = g^2(x)$, de forma que nossa nova equação fica:

$$\mathcal{I} = \int_{-\infty}^{\infty} (\frac{\partial}{\partial\theta} \ln L(\theta))^2 \, \mathrm{d}x \tag{3.36}$$

$$= \int_{-\infty}^{\infty} (\frac{f'(x)}{f(x)})^2 f(x) \, \mathrm{d}x \tag{3.37}$$

$$= \int_{-\infty}^{\infty} \frac{(2g(x))^2 g'(x)^2}{g(x)^2} \, \mathrm{d}x \tag{3.38}$$

$$= 4 \int_{-\infty}^{\infty} g'(x)^2 \, \mathrm{d}x \tag{3.39}$$

Essa nova forma é a forma de amplitude da nossa informação. Ela então evitaria esse problema da não existencia se $p(x) \to 0$. Ademais, talvez não se aplicaria muito ao nosso caso, mas a integral e o gradiente tem de existir, sendo uma condição necessaria essa.

Chapter 4

Entropia de Shannon

4.1 Uma Introdução ao conceito de entropia

4.1.1 A diferença entre Termodinamcia e Estatistica

Bom, inicialmente muito se fala quanto a entropia, porém muito no sentido termodinamico. Em suma é uma medida da quantidade de estados que as moleculas de uma determinada podem estar. No exemplo classico, a agua no estado solido é aquela que apresenta menor entropia, a liquida com uma entropia média e o vapor com uma entropia alta. Tal valor, geralmente denominado de S é calculado pela formula de Boltzmann para gas ideal, mas muitas vezes pode ser calculada e observada empiricamente.

Enquanto a entropia de Shannon, num sentido mais estatistico e de teoria da informação, tem uma certa relação com isso, não será o foco da escrita. Mas claro, não se poderia deixar passar batido essa comparação. Em seu trabalho seminal, lançando em 1948, onde ele coloca as fundações da teoria da informação, tal trabalho de Shannon foi muito produto de sua epoca, pois a grande maioria do que foi colocado foi desenvolvido para ser usado na segunda guerra mundial, tanto para classificar a informação de transmissão de radios, quanto para transmitir mensagens codificadas, quanto para decifra-las. Enquanto tambem não o foco do que sera escrito, é interessante saber o contexto historico

.

4.1.2 Informações e Bits

Em computação vemos os bits em tudo quanto é lugar. Nossa velocidade da internet é medida em bits, em geral uma ordem de grande superior, megas ou até gigas, e nosso armazenamento tambem, geralmente em gigas ou teras. Mas o que realmente são bits? Imaginemos que, por exemplo, uma variavel possa assumir apenas dois valores 0 ou 1. Ela teria apenas um bit de informação, pois assumiria apenas 2 valores. Se tivermos apenas uma letra, por exemplo, quantos bits ela teria? Se tivermos n tipos de informações, vamos precisar de

$\log_2(n)$ bits. Para um alfabeto de 26 letras, vamos precisar de $\log_2(26) \approx 4.7$ para essa uma letra. Isso supondo que cada letra tenha uma probabilidade igual de ser escolhida.

Vamos pensar agora como podemos construir para nossa entropia de Shannon. Com mais um exemplo de bits. Com a mesma ideia do alfabeto, se quisermos, por exemplo, escrever uma frase com 5 palavras, quantos bits precisariamos? Para cada letra, teriamos 4.7 bits, se cada uma delas for independente e de igual probabilidades, podemos so pegar e somar 4.7 5 vezes. De fato, podemos fazer isso para qualquer palavra, desde que não tenha espaço, ou caracteres especiais, claro. Então para uma palavra de m letras, precisariamos de $m * 4.7$ bits. Vamos pensar então, é muito mais facil, por exemplo, tentar acertar uma palavra de 3 letras, do que uma palavra de 10. A quantidade de bits está relacionado com essa 'facilidade'? Mantenhamos esse questionamento em mente

4.1.3 Conhecimento

Continuando nossa analogia no alfabeto. Vamos supor que temos uma palavra de 5 letras e você precise decifra-la. Inicialmente você não tem conhecimento nenhum, todas as letras são igualmente provaveis. Mas agora, suponha que eu te diga, existe a letra a nessa palavra. Você ja ganhou algum conhecimento. Mas você conseguiria quantifica-lo? Em uma ideia mais trivial e popular, você pode até pensar, tal letra é bastante comum na lingua portuguesa, foi-se revelado, obviamente, que apenas palavras contendo apenas as outras vogais são validas. Não reduziu-se tanto, então, nosso espaço de possibilidades. Por outro lado, se eu te falasse que contem a letra z. Já é uma boa ajuda, pois a letra z é bastante incomum, então você acaba ganhando mais informação no geral quando te falo tal dado. Para quantificar esse conhecimento ganho, podemos colocar como

$$G(B|A) = \log_2 \frac{1}{p(B)} = -\log_2(p(B)) \tag{4.1}$$

Onde temos A e B como um espaço de probabilidade, tal que $B \subset A$. O ganho de informação é sempre positivo. Ademais, ele obdece a propriedade que se $B \subset C \subset A$, temos que

$$G(C|A) = -\log_2 p(C) \tag{4.2}$$

$$G(B|C) = -\log_2 p(B|C) = -\log_2 \frac{p(B)}{p(C)} \tag{4.3}$$

Ou seja, o ganho que se tem descobrir que ele pertence a B depois de saber que pertence a C é a razão das probabilidades de B e C. Por consequencia, $G(B|A) = G(C|A) + G(B|C)$ Portanto, dado o espaço de probabilidades $A = (a_1, a_2, ..., a_n)$, se tivermos um ganho dado por $G = -\log p(a_n)$ o ganho médio esperado é dado por:

$$H(A) = \sum_i p(a_i) - \log p(a_i) \tag{4.4}$$

Esse ganho médio tambem pode ser entendido como a quantidade de informação da função densidade de probabilidade. Ela não esta relacionada, exclusivamente, com a variavel aleatoria, mas sim com sua distribuição. Não é incomum ve-la escrita como $H(p)$ inves da forma que foi escrita, para não ser criada uma confusão.

4.1.4 Diferenças entre Informação de fisher e Entropia de Shannon

Lembramos que a informação de fisher se relaciona à curvatura média da nossa função score Como isso se relaciona com a entropia. Se fossemos pensar em um sentido mais empirico quanto maior a nossa informção, menor seria o nosso caos, ou seja menor seria nossa entropia. Então necessariamente a informaçoinformação de fisher seria o inverso da entropia de shannon Não necessarimente. Existe uma diferença fundamental entre os dois. A entropia de shannon fundamentalmente se relaciona com o conteudo da mennsagem, quanto de informação ela carrega. Porem, por exemplo, digamos que você receba uma mensagem Completamente aleatoria, por assim dizer, euma mensagem com um texto claro. Qual delas teria a menor entropia: Claramente é aquela que vem com a menagem claramente mais identificada, já de antemão para nos. Porem, se fossemos decifrar as duas mensagens, qual você aprenderia mais: Seria aquela de maior entropia pois se conseguissemos identificar o que faz com que ela fique aleatoria e conseguissemos reproduzi-la de forma similar, aprenderiamos muito mais. Porém geralmente se assume que transmissões com maior entropia carregam menos infromações pois demandam menos bits para serem transportados. Dentro da teoria original de fisher ele não expressou uma relação direta com a entropia, porem, posteriormente foi-se expressado tal relação. Segue que:

4.1.5 Pontos importante na Entropia de Shannon

Até o momento so fizemos a analise de um fator, ou seja $H(A)$, onde apenas consideramos a distribuição de probabilidades A, mas se por exemplo, fomos considerar um segundo fator, ou seja, uma distribuição, B, como ela interagiria com A. Chamando $p(a_i)$ como a probabilidade de a_i acontecer e $p(b_j)$ como a probabilidade de b_j acontecer, temos alguns tipos de probabilidades conjuntas. Por exemplo, $p(a_i, b_j)$ sendo a probabilidade de a_i, b_j acontecerem simultaneamente, $p(a_i|b_j)$ sendo a probabilidade condicional de a_i acontecer dado que b_j aconteceu. Chamando a informação media conjunta de $M(A, B)$, pode ser mostrado que:

$$M(A, B) = \sum_{i,j} p(a_i, b_j) \log \frac{p(a_i|b_j)}{p(b_j)} \tag{4.5}$$

De forma que pode ser mostrado que $H(A) \geq M(A, B) \geq 0$. A interpretação de $M(A, B)$ é que ele nos diz quanto da complexidade de $H(A)$ é explicada por B , ou seja, o quanto de A é restringido por B. Ademais, podemos decompor a

entropia de Shannon como sendo

$$H(A) = M(A, B) + H(A|B) \tag{4.6}$$

Onde $M(A, B)$ mostra o quanto A é restringido por B e $H(A|B)$ fala a 'liberdade' de A dado a presença de B . Em Um sistema ecologico, então, isso nos mostraria, por exemplo, num sistema presa predator, onde ele é muito restringido pela eficiencia de caça e comida, a entropia de Shannon exprime tanto a eficiencia quanto a liberdade. Liberdade essa que se relacionaria à liberdade do predador dado a presa.