Teoria: Sistemas de Processamento de Informação

Nestor Caticha

22 de março de 2013

Sumário

Teoremas de Regraduação de Cox

Probabilidades e Informação

Estas notas para o curso "Mecânica Estatística de Sistemas de Processamento de Informação" estão e parece sempre estarão em estado embrionário e preliminar. Eventuais críticas, correções ou sugestões serão bemvindas. Algumas partes são baseadas nos livros de E. Jaynes [1] e de A. Caticha [2] que estas notas não substituem. Outras não, grande parte são baseadas em conjunto de notas de uma versão anterior deste curso.

A primeira parte discute o uso de probabilidades. O leitor, aluno deste curso não deverá ler isto como uma validação das suas crenças. Afinal, se está estudando Mecânica Estatística já parece natural o uso de probabilidades. Deve procurar falhas no raciocínio. Procurar exceções.

A idéia de apresentar uma forma de pensar que tem aplicações em uma vasta gama de assuntos, pode levar o leitor a pensar que está na presença de alguém que com um martelo, pensa que todos os problemas são pregos. Ou que estamos apresentado dogmas, dos quais não abriremos mão. No fim talvez não saiba como me defender de tais acusações, exceto alegando que o único ponto sobre o qual serie inflexível será que só podemos acreditar naquilo que a informação e evidência permitem, e só enquanto não surgir informação contraditória. Não faz sentido acreditar em algo que não é respaldado por informação [3]

Estudaremos (i) a teoria, (ii) a aplicação da teoria a técnicas de processamento de informação e (iii) a aplicação dessas técnicas a problemas mais ou menos aplicados. Para apreciar a extensão das aplicações, serão estudados problemas teóricos em aprendizagem de máquinas, estatística Bayesiana, modelagem de sistemas sociais e econômicos, modelagem de sistemas cognitivos e neurais. Em machine learning, estudaremos primeiro o átomo de hidrogênio, quer dizer, o percéptron e algumas de suas generalizações. Isto inclui percéptron multicamada, máquinas de vetor de suporte, deep learning, memórias associativas tipo Hopfield. Para isto precisaremos estudar algumas técnicas como método de réplicas e propagação de crenças. Do ponto de vista cognitivo, os tópicos mais recorrentes serão tomada de decisão, categorização e clustering, redução dimensional. Em sistemas sociais mostraremos alguns exemplos atuais de aplicações. Idéias de inferência Bayesiana e inferência Entrópica entrarão em todos estes estudos. Também olharemos aplicações a problemas inversos, caso o tempo permita, como determinação de localização de atividade usando EEG e fMRI.

[1] E. T. Jaynes, Probability Theory: the Logic of Science

[2] Entropic Inference and the Foundations of Physics

[3] Há outras formas de pensar, por exemplo acreditar em algo porque isso me deixa mais feliz. Mas eu não saberia dar um curso sobre isso. "I have a lot of beliefs, and I live by none of them - that's just the way I am... they make me feel good about who I am.--Louis CK

Teoremas de Cox

Há muitas definições matemáticas possíveis que poderiam ser usadas na tentativa de formalizar o conceito coloquial de informação. Uma forma de avançar, que é bastante comum em ciência, começa por definir matematicamente algo e depois tentar interpretar as fórmulas matemáticas para mostrar que esta interpretação esta de acordo com algumas das características que podemos atribuir ao conceito coloquial de informação que temos . Não haverá forma que satisfaça a todos pois cada um terá um exemplo onde este conceito falha.

Em lugar de começar por uma estrutura matemática pré-escolhida para servir de ferramenta de análise, começamos por uma interpretação e depois encontramos a estrutura matemática que se adapte à interpretação. A interpretação passa por estabelecer em alguns casos particulares suficientemente simples, tais que haja algum tipo de consenso, o quê deveria resultar da teoria. É possível que este procedimento pareça novo ao leitor e será surpreendente quantos resultados serão extraidos deste método e do rigor matemático que a teoria se vestirá. Como este procedimento permite saber mais claramente do que estamos falando e do que não estamos, achamos que esta é atualmente a melhor maneira de introduzir a teoria de informação. [4]

Queremos analisar uma asserção, isto é, uma frase A que em princípio é uma proposição que se apresenta como verdadeira. Uma frase pode ser julgada correta ou não de várias maneiras. Podemos pensar se é correta do ponto de vista da sua estrutura gramatical ou sintática. No entanto, nenhuma asserção sozinha pode ser analisada, no que diz respeito a se é verdadeira ou não, de forma independente do resto do universo conceitual. Ela será julgada verdadeira ou não quando analisada dentro de um contexto. A informação trazida por uma asserção C, será usada para atribuir um grau de verdade à asserção A, ou seja dentro do contexto C. Poderiamos chamar esse grau de, por exemplo, probabilidade de que A seja verdade se C for dada. Mas fazendo isto estariamos definindo de antemão que a ferramenta matemática apropriada para descrever informação é a teoria de probabilidades. Isto parece bem razoável mas não escapa às críticas acima e permite que outra ferramenta matemática seja usada por simplesmente expressar o gosto de outras pessoas ou a facilidade de uso em determinados problemas práticos com a mesma justificativa: *parece razoável , eu gosto, funciona, é prático* . Não descartamos o uso de outras ferramentas matemáticas, mas queremos deixar claro que estas poderão ser vistas como aproximações mais ou menos adequadas de uma estrutura que unifica e tem um posição diferente. O **objetivo** deste capítulo é mostrar que a escolha da teoria de probabilidades como a ferramenta matemática adequada para tratar informação é muito mais do que simplesmente

[4] Também ocorrerá que os resultados não serão universalmente satisfatórios, pois há lugar a discussões sobre o tipo de interpretação *a priori* que será imposta. Ver possíveis extensões e críticas leves que talvez não sejam tão relevantes

conveniente. Isto nos levará à teoria de inferência, baseada na teoria de probabilidades, que tem **exatamente** a estrutura da Mecânica Estatística dos pioneiros Boltzmann e Gibbs. Os "exatamentes"não são coincidências. Somos levados a repensar a Mecânica Estatística como uma teoria de inferência, mas muito mais sobre isto será dito adiante. Antes disso há muito o que fazer.

Se a informação em C não permite a certeza sobre a verdade de A então diremos que a crença que temos sobre A esta baseada em informação incompleta. Em casos particulares poderá ocorrer que dado C possa ser concluido, com certeza que a asserção A é verdadeira ou ainda em outros casos que é falsa. Quando não há alternativa para a conclusão, quando ela segue por força da informação disponível, dizemos que a conclusão é racional ou lógica. Dizemos que estamos frente a casos de raciocínio dedutivo. Nestes casos a informação disponível é *completa* pois nada falta para ter certeza. A análise destes casos remonta a Aristóteles.

Exemplos de informação completa são dados pelos silogismos Aristotélicos: suponha que recebemos a informação contida em $C =$ "$A \rightarrow B$", isto é, A implica B. Traduzindo, isto significa "se souber que A é certamente verdade, segue que a proposição B também o é." Dado isso, o que podemos dizer sobre B? Nada com certeza, mas se também recebemos a informação adicional A, isto é, que A é Verdade, então segue B, ou seja "B é Verdade".

Outro caso de informação completa é $\overline{B}$ ou seja "B é Falso", então segue $\overline{A}$, isto é, que "A é Falso".

Nas condições que $C =$ "$A \rightarrow B$" e "A é Falso" , o quê pode ser concluido? Do ponto de vista lógico clássico nada podemos concluir sobre B. Da mesma forma se for dada a informação "B é Verdade", nada podemos concluir sobre A. Estamos frente a casos de informação incompleta e a lógica clássica não serve para chegar a uma conclusão. Não é possível deduzir nada. A indução, [5]. o que quer que isto seja, e que será discutido mais à frente, será necessária para avançar A forma dedutiva da lógica permite somente tres tipos de respostas, *sim*, *não* e *não segue*. A indução nos força ou permite dividir esta última em várias possibilidades e os casos extremos nesse espectro são aqueles onde havendo certeza absoluta, haverá portanto a força da dedução. Podemos falar então sobre quais das alternativas intermediárias é mais razoável acreditar com base no que sabemos. Nota-se então a necessidade de estender a lógica para poder tratar de forma racional casos de informação incompleta. Richard T. Cox, ao se defrontar com este problema por volta da década de 1940, decidiu, como dito acima, estabelecer um conjunto de desejos (*desiderata* [6]) que a teoria deveria satisfazer, e estes serão então os axiomas da extensão da lógica. Aqui podemos discordar, propor outros axiomas, mas uma vez aceitos serão provados

[5] Segundo Harold Jeffreys em seu livro*Theory of Probability*, Bertrand Russell disse que "induction is either disguised deduction or a mere method of making plausible guesses". Jeffreys diz que "é muito melhor trocar a ordem dos dois termos e que muito do que normalmente passa por dedução é indução disfarçada, e que até alguns dos postulados de *Principia Mathematica* foram adotados por motivações indutivas" (e adiciona , são falsos). Com o tempo o próprio Russell mudou de posição, dobrado pela evidência (?) e diz no fim da sua autobiografia: "I was troubled by scepticism and unwillingly forced to the conclusion that most of what passes for knowledge is open to reasonable doubt". Sobre indução disse ainda: "The general principles of science, such as the belief of the reign of law, and the belief that every event must have a cause, are as completeley dependent on the inductive principle as are the beliefs of daily life."(On Induction)

[6]

os teoremas de reparametrização de Cox que mostram que a teoria de probabilidade é a ferramenta para o tratamento de forma racional de situações de informação incompleta. O surpreendente disto é que surge a teoria das probabilidades como a forma para lidar de forma *racional* com a informação e que corremos riscos de ser inconsistentes caso a regras de manipulação de probabilidades não sejam seguidas. Segue que não há probabilidades que não sejam condicionais embora às vezes simplesmente a linguagem esqueca de deixar explícitas as relações de condicionalidade. A amplidão da aplicabilidade da teoria que emerge é impressionante e por exemplo, quando o tipo de asserção for limitado àqueles entendidos em teoria de conjuntos as regras de manipulação serão não mais nem menos que aquelas ditadas pelos axiomas de Kolmogorov. Veremos que emerge uma relação natural entre probabilidade e freqüência e ficará claro de que forma estes conceitos estão ligados e mais importante, de que forma são distintos.

Axiomas de Cox

É interessante notar que os axiomas de Cox descritos por Jaynes não são exatamente iguais aos que Cox apresenta no seu livro *The algebra of probable inference* . A exposição de Jaynes é muito mais simples. Cox, por sua vez, esclarece sua dívida com J. M.Keynes e seu livro *A treatise on Probability*, que deve muito a Laplace e Bernuolli. A exposição de Jaynes teve uma grande influencia, mas ainda recebeu críticas e complementos [7]. Eu seguirei a apresentação de A. Caticha, que é mais completa [8].

A maneira de construir a teoria está baseada na seguinte forma de pensar bastante simples. Queremos construir uma teoria geral para a extensão da lógica nos casos de informação incompleta. Se ela for suficientemente geral, deverá ser válida em casos particulares. Se o caso for suficientemente simples, então podemos saber qual é o resultado esperado que não viole expectativas razoáveis. Poderia ocorrer que ao analisar um número de casos particulares sejam reveladas as inconsistências entre eles, nesse caso não poderemos chegar a uma teoria geral. Mas pode ser que os casos particulares sirvam para restringir e determinar a teoria geral [9].Isto é o que mostraremos a seguir.

Em primeiro lugar queremos falar sobre uma asserção A no caso de informação incompleta. Nos referimos então à crença ou plausibilidade de A ser verdade dado B e a denotamos pelo símbolo $A|B$. Por que não mais provável? Porque já existe uma teoria matemática de probabilidade e não sabemos se esta será a estrutura matemática que emergirá desta análise. Poderiamos usar outras palavras, mas crença ou plausibilidade são conhecidas o suficiente para serem úteis neste contexto.

[7]

[8] Notem que há lugar ainda para avanços nestes primeiros passos. Tentem encontrar defeitos, generalizações, melhorias nos argumentos

[9] Este comentário parece trivial, mas o uso que será dado a seguir é totalmente não trivial. Neste contexto de probabilidades foi colocado primeiro por J.Skilling, mas não de forma explicita. O destaque a este procedimento apareceu por primeira vez no livro de A. Caticha. Usaremos novamente este estilo de fazer teoria ao introduzir o conceito de entropia.

Queremos analisar o primeiro caso simples que lida com o conceito de *mais plausível*. Se A é mais plausível dada informação B do que A dada C, e esta é ainda mais plausível que A dado D então A dado B deveria ser mais plausível que A dado D. Temos assim nosso primeiro desejo, a plausiblidade deverá satisfazer alguma forma de transitividade. Isto é fácil se:

- D_1: A plausibilidade $A|B$ deverá ser representada por um número real.

Dados

$$A|B > A|C$$

e

$$A|C > A|D,$$

segue imediatamente, uma vez que são números reais, que

$$A|B > A|D,$$

de acordo com o axioma 1. Note que dizer que alguma coisa é um número real nos dá imediatamente a transitividade, mas não diz nada sobre que número deve ser atribuido, nem sobre como mudá-lo se a informação passa de B para C.

Através de certas operações e de diferentes asserções podemos criar asserções compostas. Exemplos de operadores são a negação, o produto e a soma lógicos. A negação de A é denotada por $\overline{A}$. O produto ou conjunção de duas asserções é uma terceira asserção: $C = AB$, $C = A \wedge B$ ou ainda $C = A$ **e** B. A soma ou disjunção de duas asserções é uma terceira asserção, que constuma ser denotada por $D = A + B$ ou $D = A \vee B$, ou ainda $D = A$ **ou** B.

A tabela 1.1 mostra a tabela verdade para as operações de soma e produto lógico, onde 1 = Verdade e 0 = Falso. Note que as últimas duas colunas,colocadas aqui para futura referência, mostram que $\overline{A+B}$ e $\overline{A}\,\overline{B}$ são iguais.

A	B	$A+B$	AB	$\overline{A+B}$	$\overline{A}\,\overline{B}$
1	1	1	1	0	0
1	0	1	0	0	0
0	1	1	0	0	0
0	0	0	0	1	1

Tabela 1.1

Suponha que tenhamos um método, usando a teoria geral que procuramos e ainda não temos, de analizar a plausibilidade de uma asserção composta por várias asserções através de conjunções ou disjunções. Esperamos que a plausibilidade possa ser expressa em termos da plausibilidade de asserções mais simples. Talvez haja mais de uma forma de realizar essa análise. Queremos então que:

- D_2: Se a plausibilidade de uma asserção puder ser representada de mais de uma maneira, pela plausibilidade de outras asserções, todas as formas deverão dar o mesmo resultado.

Há várias formas de usar a a palavra *consistência*. Aqui a usamos da seguinte forma. Impor que duas formas de análise devam dar o mesmo resultado não garante a consistência da teoria geral, no entanto uma teoria onde isso não ocorra será inconsistente. Usamos consistência no sentido de não manifestamente inconsistente, que é o que D_2 acima declara.

Agora olhamos para o caso simples em que a e b são mutuamente exclusivos na condição c e em qualquer outra condição. Então $a|b$ e $b|a$ representam a plausibilidade de algo que sabemos ser falso. Assim como $a|a$ e $b|b$ são a plausibilidade de algo que sabemos ser verdade. Poderia ser que hajam falsidades mais falsas que outras, ou verdades mais verdadeiras que outras, mas achamos razoável impor

- D_3 : Para todo a, $a|a = v_v$ e para a e b mutuamente exclusivos $a|b = v_f$.

Não sabemos que valores dar para v_v ou v_f, mas supomos o mesmo valor em todos os casos que tenhamos certeza de verdade ou falsidade

Todo operador na álgebra Booleana pode ser representado pelas operações conjunção (e) e negação ($\bar{\cdot}$) [10], isto é, o produto e a negação lógicas. A soma lógica pode ser obtida usando $A + B = \overline{\overline{A}\,\overline{B}}$. Precisamos então analisar a plausibilidade de asserções compostas usando esses operadores em termos das plausibilidade de asserções mais simples. Já que este conjunto de operadores é completo, esperamos que só tenhamos que analisar estes dois operadores.

[10] Este conjunto não é mínimo, mas é útil e claro.

Agora olhamos para a soma lógica. Novamente C se refere à informação subjacente e estamos interessados na plausibilidade $y = A_1A_2|C$. Há 4 plausibilidades que serão interessantes para esta análise:

$$x_1 = a|c,\ x_2 = b|c,\ x_3 = a|bc,\ x_4 = b|ac$$

. Notamos que deve haver uma dependência entre $A_1 \vee A_2|C$ e algum subconjunto de $\{x_i\}$, então

- D_4: Deve existir uma função F que relaciona $a \vee b|c$ e algum subconjunto de $\{x_i\}$.

- D_5: Deve existir uma função G que relaciona $ab|c$ e algum subconjunto de $\{x_i\}$.

Não impomos nada além da existência dessas funções, além de que dependam em algumas, se não todas, as variáveis $\{x_i\}$.

Porque um subconjunto? Qual subconjunto? Todos? Como decidir? Há 11 subconjuntos de dois ou mais membros: Seis ($\frac{4!}{2!2!}$) pares (x_i, x_j), quatro ($\frac{4!}{3!1!}$) triplas (x_i, x_j, x_k) e o conjunto inteiro (x_1, x_2, x_3, x_4)

A regra da soma

Começamos com a função F e consideramos a e b mutuamente exclusivos

$$a \vee b|c = F(a|c, b|c, a|bc, b|ac) = F(a|c, b|c, v_f, v_f)$$

mas esta é uma função de apenas duas variáveis, e da constante desconhecida v_f:

$$a \vee b|c = f(a|c, b|c)$$

Agora consideremos tres asserções a, b e c mutuamente excludentes nas condições d. Duas maneiras equivalentes de escrever a disjunção das tres são $(a \vee b) \vee c|d = a \vee (b \vee c)|d$ o que permite usar a função f

$$\begin{aligned} a \vee b \vee c|d &= f(f(a|d, b|d), c|d) \\ &= f(a|d, f(b|d, c|d)) \end{aligned}$$

ou em notação óbvia

$$f(f(x, y), z) = f(x, f(y, z)) \tag{1}$$

chamada equação da associatividade. Pode se provar [11] que existe um bijeção ϕ, dos reais nos reais, monotonicamente crescente, tal que

[11] Aequationes mathematicae 1989, Volume 37, Issue 2-3, pp 306-312 The associativity equation revisited R. Craigen, Z. Páles

$$f(x, y) = \phi^{-1}(\phi(x) + \phi(y)) \tag{2}$$

portanto podemos *regraduar* as atribuições de plausibilidade e nao mais falar dos números do tipo $a|d$ mas de números $\phi(a|d)$. Note que o fato de ser uma bijeição resulta que a ordem de preferencias não se altera, se antes as crenças sobre as asserções tinham uma certa ordem, depois da regraduação, a representação numérica das crenças não se altera. Continuamos sem saber que números são esses, mas avançamos a ponto de poder dizer que para quaisquer eventos mutuamente exclusivos

$$\phi(a \vee b|d) = \phi(a|d) + \phi(b|d). \tag{3}$$

No caso particular que $d = \overline{a}$, isto significa

$$\begin{aligned} \phi(a \vee b|\overline{a}) &= \phi(a|\overline{a}) + \phi(b|\overline{a}) && (4) \\ \phi(b|\overline{a}) &= \phi(a|\overline{a}) + \phi(b|\overline{a}) && (5) \\ & && (6) \end{aligned}$$

pois a crença $\phi(a \vee b|\overline{a})$ é equivalente à crença $\phi(b|\overline{a})$. Segue que

$$\phi(a|\overline{a}) = \phi(v_f) = \phi_f = 0 \tag{7}$$

Embora modesto, heis o primeiro resultado númerico. **O valor regraduado da certeza da falsidade é zero.**

Mas e se não forem mutuamente exclusivos? O interessante é que o resultado anterior serve para o caso geral, mas precisamos usar o truque de escrever

$$a = (a \wedge b) \vee (a \wedge \overline{b}) \;\; e \;\; b = (b \wedge a) \vee (b \wedge \overline{a})$$

Podemos escrever $a \vee b$ como uma disjunção de asserções mutuamente exclusivas:

$$\begin{aligned} a \vee b &= (a \wedge b) \vee (a \wedge \overline{b}) \vee (b \wedge a) \vee (b \wedge \overline{a}) \\ &= (a \wedge b) \vee (a \wedge \overline{b}) \vee (b \wedge \overline{a}) \end{aligned}$$

assim a equação 3 pode ser usada, levando a

$$\begin{aligned} \phi(a \vee b|d) &= \phi(a \wedge b|d) + \phi(a \wedge \overline{b}|d) + \phi(b \wedge \overline{a}|d) \\ &= \phi(a \wedge b|d) + \phi(a \wedge \overline{b}|d) + \phi(b \wedge \overline{a}|d) + \phi(a \wedge b|d) - \phi(a \wedge b|d) \end{aligned}$$

onde, na última linha adicionamos e subtraimos o mesmo número. Usando novamente a equação 3

$$\begin{aligned} \phi(a \vee b|d) &= \phi(a \wedge b \vee a \wedge \overline{b}|d) + \phi(b \wedge \overline{a} \vee a \wedge b|d) - \phi(a \wedge b|d) \\ &= \phi(a|d) + \phi(b|d) - \phi(a \wedge b|d) \end{aligned} \tag{8}$$

Exercício Desenhe o diagrama de Venn adequado a esta situação.

Regra do produto: quais as variáveis relevantes?

Queremos expressar $y = \phi(ab|c)$ em termos da função ainda por determinar G e de algum dos subconjuntos de $\{x_i\}$ Tribus sugeriu a análise das 11 possibilidades para verificar que só há duas que sobrevivem a casos extremos. Os dois conjuntos são (x_1, x_3) e (x_2, x_4). Note que se o primeiro deles fosse um dos sobreviventes, o segundo também deveria ser pela simetria trazida pela comutatividade do produto lógico.

Vejamos como chegar a esta conclusão (novamente seguimos AC)

1. $y = G(\phi(a|c), \phi(b|c))$ (1 possibilidade)
2. $y = G(\phi(a|c), \phi(a|bc))$ (2 possibilidades $a \leftrightarrow b$)
3. $y = G(\phi(a|c), \phi(b|ac))$ (2 possibilidades $a \leftrightarrow b$)
4. $y = G(\phi(a|bc), \phi(b|ac))$ (1 possibilidade)
5. $y = G(\phi(a|c), \phi(b|c), \phi(a|bc))$ (2 possibilidades $a \leftrightarrow b$)
6. $y = G(\phi(a|c), \phi(a|bc), \phi(b|ac))$ (2 possibilidades $a \leftrightarrow b$)
7. $y = G(\phi(a|c), \phi(b|c), \phi(a|bc), \phi(b|ac))$ (1 possibilidade)

Caso 1 Mostraremos que $y = a \wedge b|c = G(\phi(a|c), \phi(b|c)) = G(x_1, x_2)$ não funciona pois nao satisfaz o esperado em um caso simples. Porque não serve o subconjunto mais óbvio (x_1, x_2)? Seja a=' Helena usa um tenis esquerdo vermelho' enquanto que b=' Helena usa um tenis direito preto' . A plausibilidade dessas duas asserções será julgada dada a seguinte informação c='Helena gosta de tenis pretos e de tenis vermelhos', e talvez seja possível concluir que as duas asserções são bastante plausíveis. Mas se tivessemos $y = G(x_1, x_2)$ poderiamos ser levados a pensar que 'Helena usa um tenis esquerdo vermelho e um tenis direito preto' é bastante plausível. Posso acreditar bastante nas duas asseções, mas não que use um tenis de cada cor. Devemos rejeitar esta forma para G.

Para convencer os incrédulos no exposto acima, um argumento mais formal: Suponha que $a|d = a'|d$ e $b|d = b'|d$, mas que embora a e b sejam mutuamente exclusivos, a' e b' não o sejam. Neste caso teriamos que

$$\phi(a'b'|d) = G(\phi(a'|d), \phi(b'|d)) = G(\phi(a|d), \phi(b|d)) = \phi(ab|d) = 0.$$

E isto ocorreria para qualquer par de asserções não mutuamente exclusivas (a', b'), pois sempre poderiamos supor um caso auxiliar (a, b) adequado.

Caso 2 Se $y = G(\phi(a|c), \phi(a|bc))$ em geral, consideramos o caso particular em que $b = ad$ para qualquer d não seja mutuamente exclusivo a a. Logo

$$\begin{aligned} G(\phi(a|c), \phi(a|bc)) &= G(\phi(a|c), \phi(a|adc)) \\ &= G(\phi(a|c), \phi_t) = g(\phi(a|c)) \end{aligned}$$

segue que

$$y = \phi(ab|c) = \phi(b|c)$$
$$\phi(b|c) = g(\phi(a|c))$$

onde o lado esquerdo depende de d mas o lado esquerdo não. Esperamos que isso não ocorra em geral, e portanto eliminamos este candidato.

Caso 3 Para o caso $y = G(a|c, b|ac)$ e a alternativa $G(b|c, a|bc)$ ninguém tem encontrado casos que se oponham ao bom senso. Este será o único candidato a sobreviver e será a pedra de sustentação a toda a teoria que segue.

Caso 4 Se $y = G(\phi(a|bc), \phi(b|ac))$ somos levados a algo inaceitável considerando $a = b$, pois seguiria que

$$\phi(ab|c) = \phi(a|c) = G(\phi(a|ac), \phi(a|ac)) = G(\phi_t, \phi_t)$$

e $\phi(a|c)$ seria constante independente de a.

Caso 5 $y = G(\phi(a|c), \phi(b|c), \phi(a|bc))$. Este caso é mais complicado de analisar. Mostraremos, no entanto que se reduz a algum dos casos anteriores, sob a hipótese razoável de diferenciabilidade de G com respeito a qualquer um dos seus argumentos. Ainda consideraremos a conjunçao de mais de duas asserções , $abc|d$, que pode ser escrito de duas formas diferentes $(ab)c|d = a(bc)|d$, portanto, considerando a primeira forma obtemos

$$\begin{aligned}\phi((ab)c|d) &= G(\phi(ab|d), \phi(c|d), \phi(ab|cd)) \\ &= G(G(\phi(a|d), \phi(b|d), \phi(a|bd)), \phi(c|d), G(\phi(a|cd), \phi(b|cd), \phi(a|bcd)) \\ &= G(G(x,y,z), u, G(v,w,s)) \qquad (9)\end{aligned}$$

$$\begin{aligned}\phi(a(bc)|d) &= G(\phi(a|d), \phi(bc|d), \phi(a|bcd)) \\ &= G(\phi(a|d), G(\phi(b|d), \phi(c|d), \phi(b|cd), \phi(a|bcd)) \\ &= G(x, G(y,u,w), s) \qquad (10)\end{aligned}$$

Notamos duas maneiras de escrever a mesma coisa, por D_2 que declarava que não queremos ser manifestamente inconsistentes, devemos ter

$$G(G(x,y,z), u, G(v,w,s)) = G(x, G(y,u,w), s).$$

Ainda notamos que embora éstas variáveis possam ter quaisquer valores, não ocorre o mesmo conjunto dos dois lados: Lado esquerdo $\{x,y,z,u,v,w,s\}$, lado direito $\{x,y,u,w,s\}$. Portanto o lado esquerdo nao deve depender de $z = \phi(a|bd))$ nem de $v = \phi(a|cd)$ explicitamente. As derivadas parciais com respeito a z ou v devem dar zero:

$$\begin{aligned}0 &= \frac{\partial}{\partial z} G(G(x,y,z), u, G(v,w,s)) \\ &= \frac{\partial}{\partial r} G(r, u, G(v,w,s))_{r=G(x,y,z)} \frac{\partial}{\partial z} G(x,y,z) \qquad (11)\end{aligned}$$

Se um produto é zero, pelo menos um dos fatores é zero, de onde concluimos que ou G não depende do primeiro argumento ou não depende do terceiro. Se não depende do primeiro

$$y = G(\phi(a|c), \phi(b|c), \phi(a|bc)) = G(\phi(b|c), \phi(a|bc)),$$

voltamos ao **Caso 3**. Se não depende do terceiro

$$y = G(\phi(a|c), \phi(b|c), \phi(a|bc)) = G(\phi(a|c), \phi(b|c))$$

e voltamos ao **Caso 1.**

Fica como

Exercício mostrar que o **Caso 6** pode ser reduzido ao **Caso 3** ou ao **Caso 4** e que o **Caso 7** aos **Caso 5** ou **Caso 6**

Concluimos portanto que

$$\begin{aligned} \phi(ab|c) &= G(a|c, b|ac) \\ &= G(b|c, a|bc) \end{aligned} \tag{12}$$

Cox coloca isto como um axioma, mas não precisamos fazer isto, basta dizer que existe uma função G mas que não sabemos *a priori* quais seus argumentos. A eliminação dos casos que contradizem o bom senso em casos suficientemente simples, mostra de forma satisfatória (o leitor pode pular e reclamar, mas terá que encontrar argumentos) que as equações 12 refletem a única opção. Uma das queixas pode ser sobre a diferenciabilidade de G. Mas estamos interessados em situações onde a informação pode mudar e não alterar significativamente as crenças e esperamos ao menos continuidade de G.

Note que agora será possível concluir que 'Helena usa um tenis esquerdo vermelho e um tenis direito preto' pode ser pouco plausível por que precisamos saber a plausibilidade de 'Helena usa um tenis esquerdo vermelho dado que Helena usa um tenis direito preto' e isto pode ser pouco plausível.

Mas ainda não acabamos. Precisamos determinar a função especifica G.

Regra do produto: qual é a função G?

Novamente olhamos para um caso simples, onde podemos escrever o resultado de duas maneiras. Considere a, b, c e d com $b|d$ e $c|d$ mutuamente exclusivos, e a asserção $a(b \vee c)$ uma conjunção que pode ser escrita como uma disjunção:

$$a(b \vee c) = (ab) \vee (ac). \tag{13}$$

Podemos usar o resultado para a soma para estudar o produto $\phi(a(b \vee c)|d)$:

$$\begin{aligned} \phi(a(b \vee c)|d) &= G(\phi(a|d), \phi(b \vee c)|ad)) \\ &= G(\phi(a|d), \phi(b|ad) + \phi(c|ad)) \end{aligned} \tag{14}$$

$$\begin{aligned} \phi((ab) \vee (ac))|d) &= \phi(ab|d) + \phi(ac)|d)) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) \end{aligned} \tag{15}$$

onde a equação 14 usa primeiro quev $a(b \vee c)$ é um produto e em segundo lugar a regra da soma para asserções mutuamente exclusivas $b|d$ e $c|d$. A equação 15 mostra o resultado de considerar a soma $(ab) \vee (ac)$. Mas devido à equação 13 e D_2, estas duas formas devem dar o mesmo resultado:

$$G(x, y + z) = G(x, y) + G(x, z). \tag{16}$$

Novamente requerindo a diferenciabilidade, desta vez duas vezes, e definindo $w = y + z$ obtemos a equação diferencial

$$\frac{\partial^2 G(x,w)}{\partial w^2} = 0 \tag{17}$$

que tem solução geral $G(x,w) = A(x)w + B(x)$ em termos de duas funções desconhecidas, mas fáceis de determinar. Substituindo esta forma em 16 obtemos

$$A(x)(y+z) + B(x) = A(x)(y+z) + 2B(x), \tag{18}$$

portanto $B(x) = 0$, ou seja $G(x,w) = A(x)w$. Agora olhamos para $a|d$ e usamos $a|d = ad|d$ para a e d quaisquers.

$$\begin{aligned} \phi(a|d) &= \phi(ad|d) = G(\phi(a|d), \phi(d|ad)) \\ &= G(\phi(a|d), \phi_t) = A(\phi(a|d))\phi_t \end{aligned} \tag{19}$$

onde $\phi(d|ad) = \phi_t$ pois, obviamente d é informação completa para d. Ou seja $x = A(x)\phi_t$, logo

$$G(x,w) = \frac{xw}{\phi_t} \tag{20}$$

isto significa que, para $e = b \vee c$, b e c mutuamente exclusivos

$$\phi(ae|d) = \frac{\phi(a|d)\phi(e|ad)}{\phi_t} \tag{21}$$

o que permite regraduar mais uma vez os números associados as crenças sem mudar a ordem.

Mas resta um problema: e se retirarmos a restrição de b e c mutuamente exclusivos? Precisamos usar a equação 8 para obter:

$$\phi(a \vee b|d) = \phi(a|d) + \phi(b|d) - \phi(ab|d) \tag{22}$$

$$\begin{aligned} \phi(a(b \vee c)|d) &= G(\phi(a|d), \phi(b \vee c)|ad)) \\ &= G(\phi(a|d), \phi(b|ad) + \phi(c|ad) - \phi(bc|ad)) \end{aligned} \tag{23}$$

$$\begin{aligned} \phi((ab) \vee (ac)|d) &= \phi(ab|d) + \phi(ac|d) - \phi(abc|d) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) - G(\phi(a|d), \phi(bc|ad)) \\ &= G(\phi(a|d), \phi(b|ad))) + G(\phi(a|d), \phi(c|ad)) - G(\phi(a|d), G(\phi(b|ad), \phi(c|abd)) \end{aligned} \tag{24}$$

igualando os lados direitos das equações 23 e 24, obtemos uma nova equação funcional. Substituindo a forma para G da equação 20, vemos que o produto também funciona neste caso geral.

Exercício Mostre que a forma produto (eq. 20) é solução da equação funcional. Mostre que esta é a única forma se G for diferenciável duas vezes em cada argumento.

Da equação 21 obtemos

$$\frac{\phi(ae|d)}{\phi_t} = \frac{\phi(a|d)}{\phi_t}\frac{\phi(e|ad)}{\phi_t} \tag{25}$$

o que permite regraduar mais uma vez os números associados as crenças sem mudar sua ordem. Crenças regraduadas, de forma bijetora representam o mesmo ordenamento e portanto podem ser ainda chamados de crenças. Definimos os novos números

$$p(a|b) = \frac{\phi(a|b)}{\phi_t} \tag{26}$$

e reescrevemos os resultados

$$\begin{aligned} p(a|a) &= p_t = 1 \\ p(a|\overline{a}) &= p_f = 0 \\ p(a \vee b|c) &= p(a|c) + p(b|c) - p(ab|c) \\ p(ab|c) &= p(a|c)p(b|ac) \\ &= p(b|c)p(a|bc) \end{aligned} \tag{27}$$

Começamos a reconhecer as fórmulas que descrevem as probabilidades da soma e do produto. Mas ainda não acabamos. Precisamos determinar o que acontece com a negação. Começamos com $a \vee \overline{a}|d$ que deve ser sempre verdade e com $a\overline{a}|d$ que deve ser sempre falso:

$$\begin{aligned} 1 &= p(a \vee \overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d) - p(a\overline{a}|d) \\ &= p(a|d) + p(\overline{a}|d), \end{aligned} \tag{28}$$

ou a soma das crenças regraduadas de uma asserção e da sua negação é um.

Isso completa a identificação das crenças ou plausibilidade regraduadas em números que satisfazem as regras da probabilidade. Concluimos que a estrutura matemática adequada, e que usaremos nestas notas, para descrever situações de informação incompleta é a teoria de probabilidades.

O que foi obtido pode ser comparado com os axiomas de Kolmogorov [12]. Vemos uma diferença importante. Na formulação da teoria de probabilidade como um capítulo da teoria da medida, as probabilidades são medidas e não há menção a condicionais. Rao adicionou mais tarde a complementação introduzindo, como uma idéia tardia, a probabilidade condicional definida a partir do teorema de Bayes, que

[12] Kolmogorov

Cox obteve como uma consequência direta da consistência. A partir de

$$\begin{aligned} p(ab|c) &= p(a|c)p(b|ac) \\ &= p(b|c)p(a|bc) \end{aligned} \tag{29}$$

obtemos o teorema de Bayes [13]

$$p(a|bc) = \frac{p(a|c)p(b|ac)}{p(b|c)} \tag{30}$$

[13] T. Bayes formulou a parte da inversão: $p(a|bc) \propto p(b|ac)$, Laplace o escreveu pela primeira vez e deu-lhe a devida importância

que esconde, atrás de sua grande simplicidade, uma importancia enorme, que deriva em parte do número ilimitado que tem encontrada em várias áreas da ciência.

Este é o conteúdo dos teoremas de Cox: uma atribuição de números para descrever as crenças em asserções, dada a informação, que satisfaça os casos particulares, pode ser mudada de forma a não alterar o ordenamento das crenças e preferências e a satisfazer as regras da probabilidade. Tem cheiro e cor de probabilidade e tem todas as propriedades das probabilidades. Não falaremos mais sobre plausibilidade. Não sabiamos o que era, e a abandonamos como a um andaime, após ter construido o edifício da teoria de probabilidades. Obviamente este exercício não forneceu os valores das probabilidades. Que bom, senão fechariam os institutos dedicados ao estudo e às aplicações das probabilidades. Mais sérios, podemos dizer que a nossa grande preocupação agora será dirigida à busca de técnicas que baseadas na informação disponível permitam atribuições ou talvez o problema associado mas diferente, de atualização dos números associados a probabilidades dos eventos ou asserções de interesse quando recebemos nova informação. Esta é a preocupação central da inferência e da teoria de aprendizado e nos levará à introdução da idéia de entropia. A entropia no sentido de teoria de informação está intimamente ligada à idéia de entropia termodinâmica e mais ainda à de Mecânica Estatística como veremos mais tarde. Poderemos afirmar que a Mecânica Estatística foi a primeira teoria de informação, embora não seja costumeiro colocá-la nessa luz.

O teorema de Bayes e Informação Incompleta

Vejamos agora alguns exemplos da utilização destes resultados em casos simples onde há informação incompleta.

Voltemos agora aos silogismos iniciais. Suponha que

- A="Está chovendo"
- B="Há nuvens"
- $C =$ "$A \rightarrow B$"

Note que a implicação lógica não segue da causalidade física. Chove porque há nuvens do ponto de vista de causalidade, mas do ponto de vista lógico saber que chove obriga à conclusão que deve haver nuvens. Suponha que seja dada a informação B, ou seja é dado que há nuvens. Dentro da lógica aristotélica nada podemos dizer. Devemos com base nisso desprezar por ilógicos quem nos aconselha a levar um guarda-chuva porque há nuvens? Vejamos o que nos diz a teoria das probabilidades. Neste caso o teorema de Bayes começa a mostrar a sua força. A probabilidade $P(A|CI)$ representa a crença que esteja chovendo, sob a informação C, mas não levando em conta se há ou não nuvens. Também leva em conta I, tudo o que é sabido sobre o clima nesta estação do ano, podendo ser muita informação ou nenhuma. Não importa efetivamente que número $P(A|CI)$ seja, estará entre zero e um. Esta probabilidade é dita *a priori* em relação a B. Uma vez que se recebe e incorpora a informação que efetivamente há nuvens, ou seja B, então passaremos a $P(A|BCI)$, outro número, que é chamada a probabilidade *a posteriori* ou simplesmente posterior. Aplicando Bayes

$$P(A|BCI) = \frac{P(A|CI)P(B|ACI)}{P(B|CI)}, \tag{31}$$

que relaciona a probabilidade *a priori* e a posterior. Cortando e deixando para depois uma discussão longa sobre inferência, podemos dizer que é razoável que usemos a posterior para decidir se levaremos ou não o guarda-chuvas. A probabilidade $P(B|ACI)$ recebe o nome de verossimilhança (*likelihoood* e poderia ser calculada se tivessemos um modelo sobre a influência de A em B, mas é isso o que temos, este é um caso de informação completa! Temos certeza da veracidade de B se AC for dado. Assim

$$P(B|ACI) = 1. \tag{32}$$

O quê pode ser dito sobre o denominador $P(B|CI)$? O mínimo que pode ser dito é que

$$P(B|CI) \leq 1. \tag{33}$$

Substituindo estes resultados obtemos

$$P(A|BCI) \geq P(A|CI), \tag{34}$$

a probabilidade que atribuiremos a que A seja verdade é maior ou igual se levarmos em conta o fato que há nuvens, que aquela que atribuimos sem saber se há nuvens ou não. Finalmente nos diz que a pessoa que percebe que há nuvens e leva o quarda-chuvas está agindo de forma lógica, não dentro da lógica aristotélica, mas segunda a extensão da lógica para casos de informação incompleta, representada pela teoria das probabilidades. Vemos que o bom senso diário desta situação pode ser deduzido dos desejos impostos por Cox.

Suponha outro caso de informação imcompleta. Agora A é dado como falso. Continuaremos a insistir que não podemos dizer nada sobre B do ponto de vista da lógica? O teorema de Bayes, nos diz

$$P(B|\overline{A}CI) = \frac{P(B|CI)P(\overline{A}|BCI)}{P(\overline{A}|CI)}, \tag{35}$$

e também sabemos que $P(A|BCI) \geq P(A|CI)$ da análise anterior. Ainda mais, temos que $P(A|BCI) = 1 - P(\overline{A}|BCI)$ e $P(A|CI) = 1 - P(\overline{A}|CI)$, portanto

$$P(B|\overline{A}CI) \leq P(B|CI) \tag{36}$$

levando à conclusão que se não está chovendo, devemos atribuir uma probabilidade menor a que haja nuvens. Quem está mais disposto a carregar um chapeú de sol porque recebeu informação que não está chovendo, age de forma lógica.

Exemplo

Consideremos um exemplo clássico de testes médicos. Um teste médico serve para ajudar a determinar se um paciente está doente, mas ele não é perfeito e há evidência, baseado na história que há falsos positivos e falsos negativos. O que significa um resultado positivo? Para proceder, o mais importante é esclarecer quais são as asserções relevantes.

Consideremos as asserções

- D="paciente está doente"
- A="resultado do teste é positivo"

junto com os dados sobre

- especificidade: $P(A|D) = .90$, a probabilidade de dar positivo no teste na condição de estar doente
- sensibilidade: $1 - P(A|\overline{D}) = 1. - .2 = .8$, a probabilidade de teste dar positivo no caso em que o paciente não está doente,

Vemos que o teste é bastante específico (90%) e bastante sensível ($(80 = 100 - 20)$%).

Suponha que seu resultado no teste deu positivo, A é verdade. Isto significa que está doente? Há possibilidade de erros portanto não temos informação completa. Qual é a pergunta que devemos fazer? Pode não ser o mais óbvio a se fazer quando se recebe uma notícia ruim, mas em geral devemos aplicar o teorema de Bayes. Assim poderemos calcular $P(D|AI)$ que é o que realmente interessa, a probabilidade de ter a doença,

$$P(D|AI) = \frac{P(D|I)P(A|DI)}{P(A|I)}, \tag{37}$$

e também

$$P(\overline{D}|AI) = \frac{P(\overline{D}|I)P(A|\overline{D}I)}{P(A|I)}, \tag{38}$$

os denominadores são inconvenientes e os eliminamos olhando para a razão

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)}. \tag{39}$$

Após considerar a equação acima percebemos que não temos dados suficientes para entrar em pânico. A razão entre as probabilidades que nos interessa é $P(D|AI)/P(\overline{D}|AI)$ depende de dados que temos, sobre a especificidade e sensibilidade do teste e de dados que não temos sobre a distribuição da doença na população. A teoria que não pode nesta altura nos dar a resposta que buscamos, faz a segunda melhor coisa, indicando que informaçao adicional devemos procurar. Após esta análise voltamos ao médico e perguntamos se ele tem informação sobre a distribuição *a priori* da doença na população caracterizada por I. Suponha que recebamos informação que $\frac{P(D|I)}{P(\overline{D}|I)} = .99/.01$, só 1% da população tem a doença. Segue que

$$\frac{P(D|AI)}{P(\overline{D}|AI)} = \frac{P(D|I)P(A|DI)}{P(\overline{D}|I)P(A|\overline{D}I)} = \frac{.01 \times .90}{.99 \times .20} = 0.045. \tag{40}$$

ou seja a probabilidade de não ter a doença é aproximadamente .95. Não que isto seja uma boa notícia, afinal a probabilidade que era de 1% de ter a doença passou para 4.5% : aumentou quase cinco vezes. Mas não devemos ainda entrar em pânico nem jogar fora a informação que ganhamos com o teste.

Jaynes e o bom senso

O próximo caso simples lida com informação neutra. Suponha que

$$A|C \geq A|C',$$

ou seja a plausibilidade de A diminui quando a informação disponível passa de C para C'. Suponha que para B isso não aconteça. Pensemos no caso que B é indiferente ante a mudança de C para C'. Isto é

$$B|C = B|C'.$$

Parece razoável que se a asserção conjunta AB for considerada, esta seria mais plausível nas condições C que C'; isto é seria desejável que a teoria satisfizesse

- $A|C \geq A|C'$ e $B|DC = B|DC'$, para qualquer D, implicam que $AB|C \geq AB|C'$

Jaynes defende que este desejo está de acordo com o *bom senso*. Talvez seja difícil definir o que é bom senso, mas sería mais difícil negar que isto seja razoável. Jaynes coloca isto como um dos axiomas para chegar à teoria de probabilidades.

O leitor talvez possa se convencer através de um simples exemplo. Seja A='Há vida em Marte', C= 'Há água em Marte', $C' = \overline{C}$, a negação de C. Suponhamos óbvio que $A|C \geq A|C'$. Suponha que B='Hoje é segunda feira'. Certamente $B|C = B|C'$. e também é razoável que a plausibilidade de que haja vida em Marte e hoje seja segunda feira' dado que 'há agua em Marte' é maior ou igual a plausibilidade de que 'haja vida em Marte e hoje seja segunda' dado que 'não há agua em Marte'.

Pelo regra do produto, podemos provar isto

$$P(AB|C) = P(A|C)P(B|AC) = P(A|C)P(B|AC')$$

$$P(AB|C) \geq P(A|C')P(B|AC') = P(AB|C').$$

Exemplo do Teorema de Bayes e Ajuste de funções

Uma das primeiras lições que os estudantes de física tem ao entrarem num laboratório é sobre ajuste de curvas usando conjuntos de medidas empíricas.

Um objeto cai e medimos as posições ou velocidades como função do tempo. Estão de acordo com o que se espera de um objeto que cai na presença de um campo gravitacional? Qual é o valor de g, a aceleração da gravidade? Só para deixar isto claro, não faltarão exemplos compicados mais adiante nestas notas, olharemos para o caso em que obtemos um conjunto de dados

$$D = \{v_1, v_2, ...v_N\} \tag{41}$$

para as velocidades medidas em

$$T = \{t_1, t_2, ...t_N\}. \tag{42}$$

O modelo que queremos avaliar, refutar ou aceitar (pelo menos até ter mais dados) é

$$\mathcal{M} : v = v_0 + gt \tag{43}$$

A pergunta que quer ser respondida diz respeito a asserções do tipo $H(g)$:"O valor da aceleração da gravidade é g". Para cada valor de g que for inserido nessa frase, teremos uma asserção diferente. O que queremos é comparar o mérito de cada asserção, qual é a probabilidade de cada uma delas, para todos os valores que possam ser inseridos.

O teorema de Bayes nos permite escrever

$$P(H|DI) = \frac{P(H|I)P(D|HI)}{P(D|I)}. \tag{44}$$

O que será discutido a seguir é fundamental para este curso. Será discutido em contextos mais compĺicados e portanto vale a pena o esforço de entender cada passo. É tão importante que cada termo recebe um nome.

Em primeiro lugar temos que definir as asserções relevantes ao problema. A parte que parece menos importante, mas que na realidade é fundamental é I, que define várias coisas que de tão importantes são consideradas desnecessárias pois, para que falar o óbvio? I denota toda a informação sobre a experiência. Qual é a teoria que queremos confrontar com os dados? Quais são as características do aparelho de medida? Em que instantes de tempo t_i fazemos as medidas, quais as incertezas que estas medidas têm? Em que planeta estamos? e muito mais que ficará tácitamente escondida, mas ainda relevante.

D é o conjunto de dados, H é a hipótese que quer ser testada.

Agora o significado das probabilidades que aparecem na equação 44. Começamos pelo conhecimento que temos sobre o contexto experimental mas sem levar em consideração os dados. A distribuição de probabilidades *a priori* $P(H|I)$ codifica tudo o que sabemos sobre a gravitação antes de entrar no laboratório. Se não soubermos o planeta onde a experência é realizada, fica difícil esperar um valor e não outro. Todas as gerações de estudantes que fizeram esta experiência, dos quais temos noticia, o fizeram na terra. O resultado deu algo que se parece com 9.8 ms^{-2}. Se o resultado final fosse 9.8 kms^{-2} o aluno ficaria tentado a mudar seu resultado, mudaria de forma ad hoc seu valor no relatório, o que seria desonesto, ou faria novamente as contas. Se ainda persistir o problema, jogaria fora os dados. Isto é desonesto? Não se estiver de acordo com a sua probabilidade a priori. Qual é a probabilidade que a aceleração da gravidade seja 9.8 kms^{-2} em São Paulo? qual é a probabilidade que você atribuiria antes de entrar no laboratório? Quanto voce estaria disposto a apostar contra a veracidade dessa asserção? A priori, o estudante sabe que o valor estará por volta de dez, e pode ser constante entre 7 e 15. Muito mais que isso ou muito menos, deve ser erro, e é melhor jogar o que o estudante chama de ponto fora da curva. Isso é perfeitamente lógico e deve ser feito a não ser que em I haja a possibilidade de que algo possa mudar o valor esperado. Por exemplo a experiência esta sendo feita em cima de uma cratéra aberta por um meteorito composto do elemento X. Então podemos permitir a suposição que novos valores sejam encontrados. Seriamos cegos se considerassemos a probabilidade a priori de encontrar valores muito diferentes, nula e se assim for feito, certamente não

os encontraremos.

A probabilidade $P(D|HI)$ descreve quão verossímel seria encontrar esse conjunto de dados se além de I, o valor particular de g representado por H fosse o correto. Esta é a famosa contribuição do reverendo Thomas Bayes[14]: a inversão. Queriamos saber a probabilidade de g ter um certo valor nas condições que os dados foram observados, mas estamos olhando para a probabilidade dos dados no caso que a teoria (contida em I) e um valor particular do parâmetro g sejam verdade. Este termo recebe o nome verossimilhança (likelihood em inglés).

[14] referencia de bayes

O denominador $P(D|I)$ será interessante em outros contextos. Em geral é chamado de evidência. Pode ser obtido usando o fato que g não pode ter dois valores diferentes. As asserções para valores de de g diferentes são mutuamente exclusivas. Portando a soma sobre todas as possibilidades é um. Neste caso em que g toma valores reais, é interessante considerar que as asserções tem o significado que o valor da aceleração da gravidade está entre g e $g + dg$, somas são substituidas por integrais.

O resultado de toda a análise será a obtenção de $P(H|DI)$ que se chama a distribuição de (densidade de) probabilidade posterior, ou simplesmente a posterior.

A crítica mais comum é que a *realidade objetiva* é única e portanto não é possível que haja uma probabilidade para o valor de g. Mas não é isso o que esta probabilidade significa. g pode ter um valor único objetivo [15]. O que a posterior, ou a a priori significam é que não temos informação completa e que só podemos atribuir probabilidades às diferentes asserções sobre o valor de g. Mais dados, ou seja mais informação, permitirão novas estimativas. O que estas probabilidades codificam não é o valor de g, mas a crença que esse seja o valor correto.

[15] Sabemos que g uniforme, constante é só uma aproximação válida para quedas em distâncias pequenas em comparação ao raio da terra dentro da teoria de Newton. Mas também sabemos que essa teoria não é final, tendo sido substituida pela de Einstein, e certamente não sabemos por qual teoria vai ser substituida em anos futuros. Não sobra muito do conceito de um g que descreve uma realidade objetiva

Obtendo a posterior

Há vários exemplos que mostram a importância de determinar a distribuição a priori com muito cuidado. Podemos dizer que a probabilidade que $g < 0$ deve ser zero. Os objetos mais densos que o ar não caem para cima. Também podemos limitar os valores superiores. Poderiamos dizer que $P(H|I) = c$ se $g_{min} < g < g_{max}$ e zero fora desse intervalo. A constante c é tal que $\int_{g_{min}}^{g_{max}} P(H|I)dg = 1$ ou $c^{-1} = g_{max} - g_{min}$.

A verossimilhança $P(D|HI)$ leva em conta que as medidas são sujeitas a erros. Poderiamos dizer, por exemplo, que o modelo teórico e o modelo sobre o aparelho de medidas, juntos nos levam a esperar, que para os valor de tempo t_i, onde é feita a medida,

$$v_i = v_0 + gt_i + \eta_i. \tag{45}$$

O resultado esperado puramente pelo modelo teórico (eq. 43) é corrompido por algo que chamamos ruído. Isto esconde uma grande quantidade de ignorância sobre o processo de medida. Se pudessemos aumentar o contrôle sobre o aparelho de medida (e.g. temperatura, vento, correntes elétricas, valores das resistências, ...etc.) a amplitude de η_i poderia ser menor. Mas sempre há uma incerteza sobre o valor medido. Temos que fazer algumas hipotéses sobre η_i. Estas, supostas verdadeiras, serão incluidas na asserção I. Como não temos informação completa, devemos descrever o conjunto de ηs por uma distribuição de probabilidade $P(\eta_1....\eta_N|I_{exp})$. É razoável supor que as diferentes medidas são independentes, e usando a regra do produto lógico

$$\begin{aligned} P(\eta_1\eta_2....\eta_N|I_{exp}) &= P(\eta_1|I_{exp})P(\eta_2....\eta_N|\eta_1 I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2\eta_3....\eta_N|I_{exp}) \\ &= P(\eta_1|I_{exp})P(\eta_2|I_{exp})P(\eta_3....\eta_N|\eta_2 I_{exp}) \\ &... \\ &= \prod_i^N P(\eta_i|I_{exp}), \end{aligned} \tag{46}$$

onde usamos na primeira e terceira linha a regra do produto e na segunda a independência dos valores de $\eta_2, \eta_3....\eta_N$ e o de η_1. Temos que a distribuição conjunta é o produto das distribuições inddividuais.

Qual é a distribuição $P(\eta|I_{exp})$ a ser usada. Ainda devemos supor algo mais, por exemplo média nula e variância finita σ^2. No capítulo sobre entropia justificaremos porque isto nos leva a uma distribuição gaussiana

$$P(\eta_1, \eta_2....\eta_N|I_{\rm exp}) = \frac{e^{-\sum_{i=1}^N \frac{\eta_i^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}$$

mas $\eta_i = v_i - v_0 - gt_i$, portanto

$$P(\eta_1, \eta_2....\eta_N|I_{\rm exp}) = \frac{e^{-\sum_{i=1}^N \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}$$

e a posterior

$$P(H|DI) = \frac{P(H|I)}{PD|I)} \frac{e^{-\sum_{i=1}^N \frac{(v_i - v_0 - gt_i)^2}{2\sigma^2}}}{(2\pi\sigma^2)^{N/2}}. \tag{47}$$

O problema de inferência está pronto. Mas qual é a resposta a ser dada? Há várias quantidades que podem ser extraidas da posterior. Por simplicidade nos contentamos com o valor de g que é mais provável. Se a a priori é constante na região que a gaussiana é relevante, podemos esquecer o prefator. Teremos a estimativa conhecida como

máxima verossimilhança. A resposta é simplesmente o valor que torna o argumento da exponencial máximo,

$$g_{MV} = \arg\min_{g} \sum_{i=1}^{N} \frac{(v_i - v_0 - g t_i)^2}{2\sigma^2} \tag{48}$$

que é o velho método de mínimos quadrados. Se $P(H|I)$ fosse relevante, teriamos o máximo aposteriori g_{MAP}.

Mas escolher um valor sobre os outros esconde que não temos certeza absoluta. A largura da gaussiana $\sigma/\sqrt{N}$ nos dá uma medida da incerteza.

Ainda podemos levar em conta que valores vizinhos de g_{MAP} tem probabilidade não desprezível e apresentar o valor esperado

$$g^* = \int g P(H|DI) dg. \tag{49}$$

O que ganhamos em apresentar assim o método dos mínimos quadrados que os estudantes devem ter visto há muito tempo? Suponha por exemplo, que voce colha mais informação sobre o aparelho de medida e chegue à conclusão que a distribuição dos η não é gaussiana. Ainda assim usaria o método dos mínimos quadrados? Podemos ver quais as suposições necessárias e tentar verificar se cada uma delas é razoável ou não. Isto não é pouco, a apresentação cuidadosa pode evitar suposições que não gostariamos de fazer ao analisar os dados de uma experiência. Tão importante quanto usar a informação disponível é não usar a que não o é. O próximo capítulo levará esta idéia adiante.