PROBABILIDADES E INFERÊNCIA BAYESIANA

Probabilidade como Extensão de LP

UM SISTEMA SEMÂNTICO BASEADO EM LP E PROBABILIDADES

Idéia: “herdar” proposições de Lógica Proposicional (mas não o modelo de inferência), e reúsa-las no contexto probabilístico (Teoria dos Conjuntos).

Lógica Proposicional: sintaxe baseada em veracidade de proposições (sentenças) e suas combinações (via operadores lógicos). Não admite quantificação e variáveis (simplificação sobre LPO).

Exemplos de sentenças válidas em LP:

$$BracodoRoboOK \wedge PecaVisivel \Rightarrow RoboPegaPeca$$
$$\neg A \vee B \wedge C \Rightarrow S$$
$$A_{1,1} \wedge East_A \wedge W_{2,1} \Rightarrow \neg Forward$$

Conceitos Básicos de Probabilidade (1)

$\Rightarrow$ **O espaço amostral** Ω - conjunto de <u>todos os possíveis resultados</u> de um experimento (processo).

e.g.: os 6 possíveis resultados para um dado lançado, os 2n possíveis números binários em uma cadeia de n bits.

$\Rightarrow$ $\omega \in \Omega$ é um **evento elementar** (atômico) de Ω.

$\Rightarrow$ **Um campo** σ (σ-field) - coleção $\mathcal{F}$ de subconjuntos de Ω que satisfaz:

a) $\varnothing \in \mathcal{F}$ b) se $A_1, A_2, \ldots \in \mathcal{F}$ então $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ c) se $A \in \mathcal{F}$ então $A^c \in \mathcal{F}$

Exemplos de σ-field:

$\mathcal{F} = \{\varnothing, \Omega\}$.
$\mathcal{F} = \{\varnothing, A, A^c\ \Omega\}$, para todo $A \subset \Omega$.
O *power set* (conjunto dos subconjuntos) de Ω.

Conceitos Básicos de Probabilidade (2)

Posso relacionar "resultados" de experimentos (ou eventos de interesse) a σ-fields:

$(\Omega, \mathcal{F})$

Os possíveis resultados (eventos) elementares

Os resultados (eventos) de interesse

Exemplo: Jogo duas moedas em sequência. Estou interessado no evento "ocorre lado igual nas duas moedas".

Ω = {(cara,cara),(cara,coroa),(coroa,cara),(coroa,coroa)}.
$\mathcal{F}$ = {$\varnothing$, **{(cara,cara),(coroa,coroa)}** , **{(cara,coroa),(coroa,cara)}** , Ω}.

O evento "impossível" (cond. a)

O evento

Complementar do evento (cond. c)

O evento "certo" (cond. b)

Conceitos Básicos de Probabilidade (3)

- Uma **medida de probabilidade P** sobre $\{\Omega, \mathcal{F}\}$ é uma função

  P: $\mathcal{F} \rightarrow [0,1]$

  tal que:

  a) **P**($\varnothing$)=0;

  b) **P**(Ω)=1;

  c) se A_1, A_2, ... é uma coleção de membros disjuntos de $\mathcal{F}$, então

$$\mathrm{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathrm{P}(A_i)$$

- A tripla (Ω, $\mathcal{F}$, **P**) composta por um conjunto amostral Ω, um σ-space $\mathcal{F}$ e uma medida de probabilidade **P** forma um **espaço de probabilidade**.

Exemplos

- Lançamento de uma moeda (possivelmente viciada)

Ω = {H,T}, $\mathcal{F}$ = {$\varnothing$, H, T, Ω}, **P**(H)=p, p $\in$ [0,1].

P($\varnothing$) = 0 (def.), **P**(Ω) = 1 (def.)
H, T disjuntos $\Rightarrow$ **P**(Ω) = **P**(H $\cup$ T) = **P**(H) + **P**(T) = p + **P**(T) $\Rightarrow$ **P**(T) = 1-p.

- Lançamento de um dado.

Ω = {1,2,3,4,5,6}, $\mathcal{F}$ = $\{0,1\}^{\Omega}$, **P**(A) = $\sum_{i \in A} p_i$ ($p_i \in [0,1]$), para todo A $\subseteq \Omega$.
Evento A_1: aparecer número par.
Evento A_2: aparecer o “5”.

$P(A_1 \cup A_2) = P(A_1) + P(A_2) = [1/6 + 1/6 + 1/6] + 1/6 = 2/3.$

Conceitos Básicos de Probabilidade (4)

Note que o conceito formal de medida de probabilidade não faz referência a freqüência de ocorrência, crenças, chances, etc.

Entretanto, o relacionamento a partir da definição de medida de probabilidade é perfeitamente natural:

a) O impossível nunca acontece...
b) Algum evento tem que resultar de um experimento...
c) A "chance" de ocorrer A ou B ou C disjuntos é a soma das "chances" de ocorrer cada um dos eventos A, B e C.
d) Qualquer "chance" está entre 0 e 1.

Lema 1: $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$
Lema 2: $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - P(A, B)$ (A e B não disjuntos)
Lema 2: $\mathbf{P}(A) = \mathbf{P}(A,B) + \mathbf{P}(A, \neg B)$

Conceitos Básicos de Probabilidade (5)

P(A | K) – probabilidade condicional ou posterior.
Crença em *A*, dado o corpo de informação *K*.

P(A) – probabilidade *a priori*: Crença em A, na falta de informação adicional proveniente de K.

Variável aleatória: função que associa possíveis eventos a uma probabilidade de ocorrência.

P*(Tempo=Sol) = 0.7*
P*(Tempo=Chuva) = 0.2*
P*(Tempo=Nublado) = 0.1*

Em que sentido isto é uma extensão de LP? Vejamos...

proposições e probabilidades.pdf

prob prior e dist.pdf

Probabilidade condicional

Probabilidade condicional ou posterior, e.g., **P**(cárie|dordedente) = 0.8

i.e., dado que dordedente é tudo que conheço, a chance de cárie (vista por mim) é de 80%.

NÃO "se dordedente *então* 80% de chances de cárie"

(Notação: **P**(cárie|dordedente) = vetor de 2-elementos vetores de 2-elementos)

Se sabemos mais, e.g., *cárie é também observada*, então

P(cárie|dordedente, *cárie*) = 1

OBS:

1) A crença menos específica permanece válida, mas pode ficar inútil.

2) A nova evidência pode ser inútil:

P(cárie|dordedente, Corinthians derrotado) = **P**(cárie|dordedente) = 0.8

NOTE A IMPORTÂNCIA DO CONHECIMENTO DO DOMÍNIO PARA QUALQUER PROCESSO DE INFERÊNCIA.

O Axioma Básico

$$\mathrm{P}(A|B) = \frac{\mathrm{P}(A,B)}{\mathrm{P}(B)}$$

Isto parece intuitivo?
A falta de naturalidade deste axioma é uma das grandes críticas a teorias de IA baseadas em Probabilidade.

Mais natural é: $\mathrm{P}(A,B) = \mathrm{P}(A|B)\,\mathrm{P}(B)$

onde *B* forma um "contexto" para o evento *A*.

Há uma versão geral para distribuições completas, e.g.,
P(Tempo,cárie) = **P**(Tempo|cárie)**P**(cárie)
(Um conjunto de 4 × 2 equações, e não multiplicação de matrizes)

Corolário:

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A|B_i)\,\mathrm{P}(B_i)$$

Regra da Cadeia

Generalizando:

$$\mathrm{P}(A \mid k) = \sum_i \mathrm{P}(A \mid B_i, k)\,\mathrm{P}(B_i \mid K)$$

A Regra da Cadeia:

$$\mathrm{P}(E_1, E_2, \ldots E_n) = \mathrm{P}(E_n \mid BE_{n-1}, \ldots E_2, E_1) \ldots \mathrm{P}(E_2 \mid E_1) P(E_1)$$

$\mathbf{P}(X_1, \ldots, X_n) = \mathbf{P}(X_1, \ldots, X_{n-1})\ \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$

$= \mathbf{P}(X_1, \ldots, X_{n-2})\ \mathbf{P}(X_{n-1} \mid X_1, \ldots, X_{n-2})\ \mathbf{P}(X_n \mid X_1, \ldots, X_{n-1})$

$= \ldots$

$= \Pi\ \mathbf{P}(X_i \mid X_1, \ldots, X_{i-1})$

Interpretação e Modelos Probabilísticos

Desnecessário num curso de probabilidade, mas útil do ponto de vista de IA. Desnecessário do ponto de vista matemático, mas importante para adeptos de Probabilidade Bayesiana como ferramenta para descrever a realidade.

$$\mathrm{P}(B \mid A) = \frac{\mathrm{P}(A,B)}{\mathrm{P}(A)}$$

Crença em *B* depois de se descobrir *A* nunca é menor do que a crença em *A*,*B* antes de se descobrir *A*.

Definindo Grau de Surpresa = $[\mathbf{P}(A)]^{-1}$, observo que a razão entre $\mathbf{P}(B|A)$ e $\mathbf{P}(A,B)$ aumenta com este.

Modelo Probabilístico: codificação de informação probabilística que permite calcular a probabilidade associada a qualquer sentença formada a partir de proposições atômicas. Normalmente, o modelo é especificado por uma *distribuição conjunta* associada à conjunção das variáveis.

Exemplo: Sentenças *A*,*B*,*C*. O modelo especifica probabilidades para as sentenças $(A \wedge B \wedge C)$, $(A \wedge B \wedge \neg C)$, etc., de modo que sua soma seja 1.

Suficiência de Modelos Probabilísticos

- Cada conjunção de variáveis é um evento elementar(ponto).
- Uma fórmula qualquer é um conjunto de pontos.
- Qualquer fórmula booleana pode ser expressa como uma disjunção de eventos elementares.
- Os eventos elementares são mutuamente exclusivos.

Do axioma: $P(S) = P(E_1) + P(E_2) + \ldots$

E uso o axioma básico para calcular qualquer $P(A|B)$.

Uma medida de probabilidade (no sentido estrito)!
Um modelo completo (no sentido de modelos lógicos)!

Inferência por enumeração:

inference by enumeration.pdf

Inversão Bayesiana (Regra de Bayes)

$$P(H|e) = \frac{P(e\,|\,H)P(H)}{P(e)}$$

P(*H*| *e*): Probabilidade posterior
P(*H*): Probabilidade *a priori*

Prova . . .

Por quê esta fórmula é importante?
P(*e*|*H*) é fácil de calcular, ao contrário de **P**(*H*|*e*).

Exemplo. No cassino, um croupier fala 12! Ele jogou os dados ou estava comandando um jogo de roleta?

P(*12*|*dados*), **P**(*12*|*roleta*): fácil de modelar. **P**(*dados*), **P**(*roleta*): fácil, basta ver número de mesas de dado ou roleta no cassino. **P**(*dados*|*12*), **P**(*roleta*|*12*): não é tão fácil estimar . . .

Predição e Diagnóstico

Definindo:

Chances *a priori*: $O(H)=\frac{P(H)}{P(\neg H)}=\frac{P(H)}{1-P(H)}$

Taxa de verossimilhança: $L(e\,|\,H)=\frac{P(e\,|\,H)}{P(e\,|\,\neg H)}$

Chances *a posteriori*: $O(H\,|\,e)=\frac{P(H\,|\,e)}{P(\neg H\,|\,e)}$

Temos:

O(H|e) = L(e|H)O(H)

O(H): predição baseada apenas na informação de *background*.

L(e|H): diagnóstico baseado na evidência..

Acumulando Evidências

Considere *N* alarmes diferentes, cada um sensível a um mecanismo distinto e produzindo um som diferente.

Sejam *H* = evento "Roubo Ocorreu", e^k = evidência fornecida pelo alarme *k* (e^k_0: inativo, e^k_1 : ativo)

A sensibilidade do alarme é caracterizada por: $L(e_1^k\,|\,H)=\frac{P(e_1^k\,|\,H)}{P(e_1^k\,|\,\neg H)}$

Alguns alarmes ativos e outros não: evidência conflitante.

$$O(H|e^1,e^2,\ldots e^n)=L(e^1,e^2,\ldots e^n\,|\,H)O(H)$$

Difícil de calcular, mas se assumirmos independência entre sensores:

$$P(e^1,e^2,\ldots e^n\,|\,H)=\prod_{k=1}^{n}P(e^k\,|\,H),\; P(e^1,e^2,\ldots e^n\,|\,\neg H)=\prod_{k=1}^{n}P(e^k\,|\,\neg H)$$

e portanto, $O(H|e^1,e^2,\ldots e^n)=O(H)\prod_{k=1}^{n}L(e^k\,|\,H)$

Ou seja: características individuais suficientes para determinar impacto combinado dos alarmes.

A **<u>independência de evidências</u>** usualmente simplifica muito o problema.

Recursão Bayesiana

Sejam:

- ***H***: hipótese
- $\mathbf{e}_n = e^1, e^2, \ldots, e^n$: dados observados no passado (evidências)
- *e*: um novo fato

Como calcular $P(H|\, \mathbf{e}_n, e)$?

Método Animal:
- adiciono *e* à coleção $\mathbf{e}_n$
- calculo o impacto de ***H*** no novo conjunto $\mathbf{e}_{n+1} = \{\mathbf{e}_n, e\}$

Trabalho insano: preciso de toda a seqüência histórica de dados a cada passo.

Recursão Bayesiana

Sob certas condições, posso fazer:

$$P(H|e_n, e) = P(H|e_n)\frac{P(e\,|\,e_n, H)}{P(e\,|\,e_n)}$$

Observe que $P(H|\mathbf{e}_n)$ faz o papel da probabilidade *a priori* no cálculo do impacto da nova informação *e*. Isto ainda pode ser trabalhoso. Freqüentemente, porém, há independência entre a nova evidência e a coleção de evidências passadas:

$P(e|\mathbf{e}_n\, H) = P(e|H)$ e $P(\, e|\mathbf{e}_n, \neg H) = P(e|\, \neg H\,)$

e portanto

$$O(H|e_{n+1}) = O(H\,|\,e_n)L(e\,|\,H)$$

que é um procedimento recursivo para calcular chances à medida que nova informação é adquirida.

Calculando logaritmos (*log-likelihood*)

$$\log O(H|e_{n+1}) = \log O(H\,|\,e_n) + \log L(e\,|\,H)$$

que permite uma interpretação intuitiva clara para o efeito da nova informação.

Hipóteses a Várias Variáveis

Independência Condicional: OK se variáveis que influenciam hipótese são dependentes de mecanismos intrínsecos a cada uma.

Em geral, circunstâncias externas podem afetar grupos de variáveis, introduzindo uma dependência “escondida”. O que fazer?

Solução: aumentar o refinamento do espaço de hipóteses.

Exemplo: ao invés de $H = roubo$ e $H = \neg\, roubo$, defino:

- $H_1 =$ roubo, porta arrombada
- $H_2 =$ roubo, janela arrombada
- $H_3 =$ sem roubo, silêncio total
- $H_4 =$ sem roubo, terremoto.

de modo que cada hipótese corresponda a um único estado dos sensores.

Modelagem Hierárquica: Evidência Incerta

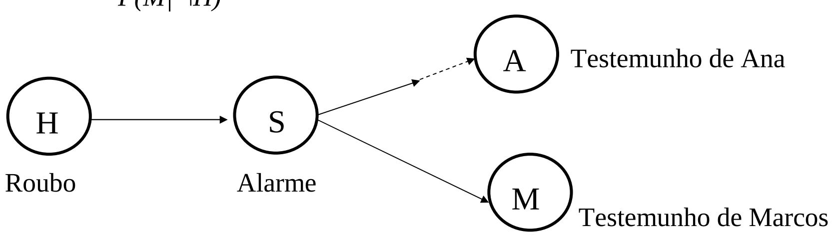
Exemplo. João recebe ligação do vizinho Marcos, que afirma ter ouvido o som de um alarme anti-roubo vindo da direção da casa de João. Enquanto se prepara para ir para casa e verificar o que houve, João lembra que Marcos é um brincalhão de péssimo gosto, e decide ligar para sua outra vizinha Ana, mais confiável.

Evidência $S = Som$ é incerta: não posso simplesmente escrever

$$O(H|S) = L(S|H)O(H).$$

A única evidência real é o testemunho de Marcos: só tenho $O(H|M) = L(M|H)\,O(H)$.

$L(M \mid H) = \dfrac{P(M|H)}{P(M|\neg H)}$ não trivial: depende de inferência em 2 passos (ver figura).



Modelagem Hierárquica: Evidência Incerta

Mesmo que tenha *L(M|H)*, não posso combiná-lo de modo simples com outras evidências (como o testemunho de Ana) como no exemplo anterior, pois os depoimentos não são condicionalmente independentes com respeito a *H*. Em outras palavras, não posso afirmar *P(A|H,M) = P(A|H)* porque o testemunho de Marcos provê evidência mais forte de roubo (ou seja, de disparo de alarme ouvido por Ana).

Posso porém assumir independência entre *A* e *H* com respeito a *S* , uma vez que saibamos se o alarme disparou ou não.

Resolvemos o problema incorporando a variável intermediária *S*:

$$P(H_i \mid A,M) = \alpha P(A,M \mid H_i)P(H_i) = \alpha P(H_i)\sum_j P(A,M \mid H_i,S_j)P(S_j \mid H_i)$$

onde j corresponde a cada um dos possíveis estados do alarme. A independência condicional de *A,M,Hi* com respeito a *S* permite:

$$P(A,M \mid H_i,S_j) = P(A \mid S_j)P(M \mid S_j)$$

e portanto

$$P(H_i \mid A,M) = \alpha P(H_i)\sum_j P(A,S_j)P(M \mid S_j)P(S_j,H_i)$$

CT215 21

Modelagem Hierárquica: Independência Condicional como Simplificador

Computacionalmente, o processo de usar a independência condicional relativa à *S* permite divisão do problema em **estágios independentes** para formar uma inferência global (um processo de encadeamento).

O processo parece ser comum no processo de raciocínio humano:

Exemplo - Medicina e definição de quadros clínicos para variáveis que produzem independência condicional.

CT215 22

Modelagem Hierárquica: Evidência Virtual

Considere a seguinte modificação da estória. Quando João liga para Ana, esta se mostra prolixa e dispersiva. Ao invés de responder se de fato ouviu o alarme, Ana discorre sobre sua última operação, fala de futebol e comenta o barulho da vizinhança nos últimos meses. Da conversa, João conclui que provavelmente existe algo como *80%* de chance de que Ana tenha de fato ouvido o alarme de sua casa. Isto não é fácil de modelar. .

P(e|Som) para este tipo de evidência certamente não pode ser modelado: cada evidência corresponde a uma possível “conversa”. E *P(Som|e)* requer especificação precisa sobre como a evidência *e* foi obtida.

Problema da interpretação autônoma: intérprete não consegue explicar processo interpretativo, mas sua informação é valiosa.

No exemplo: João estabeleceu uma medida de confiança de *0.8* para a hipótese *Som*, mas o processo de **obtenção da evidência** está escondido. Como combinar este julgamento com crenças prévias e ter certeza de que uma informação não é usada mais de uma vez?

Simplificação: assume-se que sumários de evidência virtual são produzidos independente de informação prévia. A evidência acima portanto não pode ser interpretada como *P(Som|A)=0.8*, pois este é sensível, por exemplo, à características do alarme (*P(Som|H)*).

Normalmente, interpreto evidência virtual como medida de verossimilhança:

$$P(A \mid Som) : P(A \mid \neg Som) = 4:1$$