MATB16

Laboratório de Inteligência Artificial

Aprendizado Probabilístico

Tatiane Nogueira Rios
Ricardo Araújo Rios

LabIA
Instituto de Computação - UFBA

Agenda

Revisão de teoria da probabilidade

Inferência probabilística

Regras de Bayes

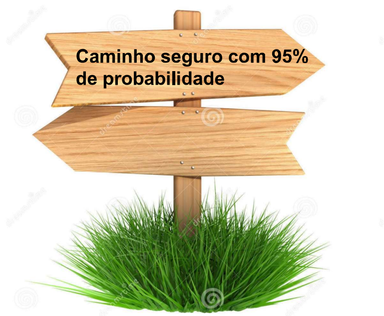
Combinação de evidências

Recapitulando teoria da probabilidade...

- Determina o **grau de crença**
- Baseado em eventos
- Valor muda à medida que a informação aumenta
- Exemplo:
  - Existe 80% de crença/chance de que vai chover.
    - Tendo sido analisados 10 dias, em oito deles choveu e outros 2 não choveu. Se a análise mudar para 11 dias, essa probabilidade muda.

Teoria da probabilidade

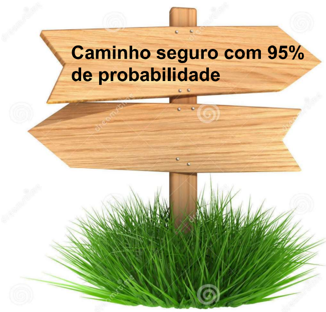
- Qual o significado da sinalização abaixo?

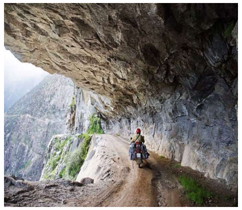


Teoria da probabilidade

- Qual o significado da sinalização abaixo?

O valor 95% é determinado pela **função densidade**: indica a frequência relativa de uma situação x em um conjunto A





Notação básica

- Como as afirmações lógica, as afirmações probabilísticas são acerca de **mundos possíveis** (quão prováveis são os vários mundos)
- O conjunto de todos os mundos possíveis é chamado **espaço amostral**
  - Ex: Se jogamos 2 dados distintos, existem 36 mundo possíveis (combinações).

Notação básica

- Considere: Ω um espaço amostral e ω elementos do espaço
  - Um **modelo de probabilidade** totalmente especificado associa uma probabilidade numérica $P(\omega)$ a cada mundo possível.
  - Todo mundo possível tem uma probabilidade entre 0 e 1
    - $0 \leq P(\omega) \leq 1$
  - A probabilidade do espaço amostral é 1.
    - $\sum_{\omega \in \Omega} P(\omega) = 1$
- Ex: Se assumirmos que os dois dados não são “viciados” e um lançamento não interfere no outro, cada um dos mundos possíveis (combinações) tem probabilidade 1/36.

Notação básica

- Afirmações probabilísticas geralmente não são sobre mundos possíveis particulares, mas sobre os seus conjuntos.
  - Exemplo, poderíamos estar interessados nos casos em que os dois dados chegam ao resultado 11.
- Esses conjuntos são chamados **eventos.**
- Em IA os conjuntos são sempre descritos por **proposições** em uma linguagem formal.
- A probabilidade associada a uma proposição é definida como sendo a soma das probabilidades dos mundos nos quais é válida:
  - Para qualquer proposição $\emptyset, P(\emptyset) = \sum_{\omega \in \emptyset} P(\omega)$
  - Ex: P(Total=11) = P((5,6)) + P((6,5)) = 1/36+1/36 = 1/18

Notação básica

- Probabilidades como P(Total=11) são chamadas **probabilidades incondicionais** ou **anteriores**:
  - Grau de crença em proposições na ausência de qualquer outra informação
  - Ex: **P(cárie) -** probabilidade incondicional (check-up)
- Quando há alguma informação (evidência), tem-se a probabilidade condicional ou posterior
  - Ex: **P (cárie | dor de dente) -** probabilidade condicional
- As probabilidades condicionais são definidas em termos de probabilidades incondicionais:

$$P(a|b) = \frac{P(a \wedge b)}{P(b)}$$

Notação básica

- Um mundo possível é representado por um conjunto de variável/pares de valor
- Na teoria da probabilidade, as variáveis são chamadas **variáveis aleatórias** e seus nomes começam com letra maiúscula.
  - Ex: Variáveis aleatórias: Total e Dado1
- Cada variável aleatória tem um **domínio** (conjunto de valores possíveis que pode assumir)
  - Ex: O domínio de Dado1 é {1, …, 6}, assim como o domínio de Idade pode ser {juvenil, adolescente, adulto}
- Os valores são sempre minúsculos e podem ser finitos ou infinitos

Notação básica

- Pode-se combinar as proposições elementares usando conectivos da lógica proposicional:

$$P(cárie|\neg dor\ de\ dente\ \wedge adolescente) = 1$$

- As probabilidades de todos os valores possíveis de uma variável aleatória pode ser escrita como **distribuição de probabilidade**: **P**(Tempo) = <0,6;0,1;0,29;0,01> que equivale à:

  P(Tempo=ensolarado)=0,6
  P(Tempo=chuvoso)=0,1
  P(Tempo=nublado)=0,29
  P(Tempo=neve)=0,01

Notação básica

- Quando a variável aleatória pode assumir um número infinito de valores, tem-se uma **função densidade de probabilidade (fdp).**
  - Ex: P(TempMeioDia=x) = $\text{Uniforme}_{[18C,26C]}(x)$
  - A temperatura ao meio dia é distribuída uniformemente entre 18 e 26 graus Celsius.
- A distribuição sobre variáveis múltiplas é chamada **distribuição de probabilidade conjunta**
  - Ex: P(Tempo, Cárie) = tabela de probabilidades 4 x 2.

Notação básica

- Um mundo possível é definido para ser uma atribuição de valores a todas as variáveis aleatórias consideradas.
  - Ex: se as variáveis aleatórias são Cárie, Dor de dente e Tempo, então existem 2 x 2 x 4 = 16 mundos possíveis.

Inferência probabilística

- Computação de probabilidades posteriores de proposições de consulta dada uma evidência observada.
- Base de conhecimento: **distribuição conjunta total**
- Considere a distribuição conjunta total para o mundo DorDeDente, Cárie, Boticão:

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Inferência probabilística

- P(cárie) = ?

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Inferência probabilística

- P(cárie) = 0,108+0,012+0,072+0,008=0,2

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Inferência probabilística

- P(cárie) = 0,108+0,012+0,072+0,008=0,2

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Probabilidade marginal

Inferência probabilística

- P(cárie|dordedente) = ?

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Inferência probabilística

- P(cárie|dordedente) = P(cárie ∧ dordedente) / P(dordedente) = 0,108+0,012 / 0,108+0,012+0,016+0,064 = 0,6

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Probabilidade condicional

Independência

- Para expandir a distribuição conjunta total, podemos adicionar a variável Tempo.
- A distribuição conjunta total agora será: P(dordedente, boticão, cárie, tempo)=2x2x2x4 = 32
- Considerando que o tempo é **independente** dos problemas dentários de alguém, então a tabela de 32 elementos pode ser reduzida para uma tabela de oito elementos e uma tabela de 4 elementos.
- A independência entre as proposições a e b pode ser escrita como:
  - $P(a|b)=P(a)$ *ou* $P(b|a) = P(b)$ *ou* $P(a \wedge b)=P(a)P(b)$

Independência condicional

- Se eu tenho cárie, a probabilidade do boticão acertar esse dente não depende de minha dor de dente:
  - *P(boticão | dordedente, cárie) = P(boticão | cárie)*
- A mesma independência ocorre se eu não tiver uma cárie:
  - *P(boticão | dordedente, ~cárie) = P(boticão | ~cárie)*
- Logo, Boticão é **condicionalmente independente** da dordedente dado cárie:
  - *P(boticão | dordedente, cárie) = P(boticão | cárie)*

Regra de Bayes

Sabe-se que as probabilidades condicionais são definidas em termos de probabilidades incondicionais:

$$P(a|b) = \frac{P(a \land b)}{P(b)}$$

Esta equação pode ser escrita de maneira diferente, estabelecendo a **regra do produto**:

$$P(a \land b) = P(a|b)\,P(b)$$

Regra de Bayes

A regra do produto pode ser escrita de duas formas:

$$P(a \land b) = P(a|b)\,P(b) \quad \text{e} \quad P(a \land b) = P(b|a)\,P(a)$$

Igualando os dois membros da direita e dividindo por P(a) tem-se o **teorema de bayes**:

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Essa equação simples é a base de todos os sistemas modernos de IA para inferência probabilística

Exemplo

- Problema de diagnóstico médico
- Mundos possíveis:
  - O paciente tem câncer ou não tem câncer
  - O resultado é Positivo (+) ou Negativo (-)
- Conhecimento a priori:
  - De toda a população, somente uma fração de 0.008 tem a doença
  - Resultado verdadeiramente positivo em 98% dos casos em que a doença realmente ocorre
  - Resultado verdadeiramente negativo em 97% dos casos em que a doença não ocorre

Exemplo

- Uma paciente teve resultado de laboratório positivo. Diagnostica-se a paciente como tendo câncer ou não?

Exemplo

- Uma paciente teve resultado de laboratório positivo. Diagnostica-se a paciente como tendo câncer ou não?
- Tem-se que:

	Cancer (0,008)	*~Cancer (0,992)*
Positivo	0,98	0,03
Negativo	0,02	0,97

- De toda a população, somente uma fração de 0.008 tem a doença
- Resultado verdadeiramente positivo em 98% dos casos em que a doença realmente ocorre
- Resultado verdadeiramente negativo em 97% dos casos em que a doença não ocorre

Exemplo

- Uma paciente teve resultado de laboratório positivo. Diagnostica-se a paciente como tendo câncer ou não?
- Tem-se que:

	Cancer (0,008)	*~Cancer (0,992)*
Positivo	0,98	0,03
Negativo	0,02	0,97

$$P(Cancer|+) = P(+ \mid cancer) * P(cancer) = 0{,}98 * 0{,}008 = 0{,}007$$
$$P(\sim Cancer|+) = P(+ \mid \sim cancer) * P(\sim cancer) = 0{,}03 * 0{,}992 = 0{,}0298$$

Lei da probabilidade total

- Frequentemente não sabemos calcular P(X)
  - Se o acontecimento “Y” pode ocorrer em “m” condições diferentes Y_1, Y_2, Y_3, … Y_m; e esses m sub-acontecimentos forem mutuamente exclusivos, então, pode-se escrever a probabilidade de X como

$$P(X) = \sum_{i=1\ldots m} P(X|Y_i) * P(Y_i))$$

Recapitulando...

- Teorema de Bayes
  - Expressa a probabilidade a posteriori de uma hipótese baseada
    - na probabilidade a priori dessa hipótese
    - na probabilidade a priori da evidência
    - verossimilhança entre causa e efeito

$$P(b|a) = \frac{P(a|b)P(b)}{P(a)}$$

Combinação de evidências

- A regra de Bayes pode ser útil para responder a consultas probabilísticas condicionadas sobre uma única peça de evidência.
- As informações probabilísticas estão frequentemente disponíveis sob a forma: $P(efeito|causa)$.

Combinação de evidências

- Mas o que acontece se tivermos duas ou mais evidências?
- Ex: $P(cárie|dordedente \wedge boticão)$ = ?

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Combinação de evidências

- Mas o que acontece se tivermos duas ou mais evidências?
- Ex: $P(cárie|dordedente \wedge boticão)$ = ?

	dordedente		*~dordedente*	
	boticão	*~boticão*	*boticão*	*~boticão*
cárie	0,108	0,012	0,072	0,008
~cárie	0,016	0,064	0,144	0,576

Distribuição conjunta total

Combinação de evidências

- A reformulação anterior só funciona **se forem conhecidas todas as probabilidades condicionais** da conjunção $dordedente \wedge botic\tilde{a}o$ para cada valor de cárie.
    - Só é viável para apenas duas variáveis de evidência.
    - Se há n variáveis de evidência possíveis (raio X, dieta, higiene oral etc..), então haveria 2^n combinações possíveis de valores observados para os quais precisaríamos conhecer probabilidades condicionais.

Combinação de evidências

- É preciso encontrar asserções adicionais sobre o domínio para simplificar as expressões.
- Essas variáveis são independentes, dada a presença ou a ausência de uma cárie.

Combinação de evidências

- Cada uma é diretamente causada pela cárie, mas nenhuma delas tem efeito direto sobre a outra: a dor de dente depende do estado dos nervos do dente, enquanto a precisão da ferramenta depende da habilidade do dentista, para o qual a dor de dente é irrelevante.

Logo, tem-se que:

$$P\,(dordedente \wedge botic\tilde{a}o|c\acute{a}rie) = P(dordedente|carie)P(botic\tilde{a}o|c\acute{a}rie)$$

Combinação de evidências

- Cada uma é diretamente causada pela cárie, mas nenhuma delas tem efeito direto sobre a outra: a dor de dente depende do estado dos nervos do dente, enquanto a precisão da ferramenta depende da habilidade do dentista, para o qual a dor de dente é irrelevante.

Logo, tem-se que:

$$P\,(dordedente \wedge botic\tilde{a}o|c\acute{a}rie) = P(dordedente|carie)P(botic\tilde{a}o|c\acute{a}rie)$$

Independência condicional

Combinação de evidências

Quando há mais de uma evidência, os requisitos de informações são idênticos aos da inferência, utilizando cada item de evidência separadamente: a probabilidade a priori para a variável de consulta e a probabilidade condicional de cada efeito, dada sua causa (Regra de Bayes + Independência condicional):

$$P(cárie|dordedente \wedge boticão) = P(dordedente|cárie)\ P(boticão|cárie)\ P(cárie)$$

Combinação de evidências

- A decomposição de grandes domínios probabilísticos em subconjuntos conectados livremente por meio de independência condicional é um dos desenvolvimentos mais importantes na história recente da IA.
- Quando uma única causa influencia de maneira direta vários efeitos, todos condicionalmente independentes, dada a causa, tem-se a distribuição conjunta total:

$$P(Causa, Efeito_1, \dots, Efeito_n) = P(Causa) \prod_i P(Efeito_n|Causa)$$

Combinação de evidências

- A **decomposição de grandes domínios probabilísticos** em subconjuntos conectados livremente por meio de independência condicional é **um dos desenvolvimentos mais importantes na história recente da IA**.
- Quando uma única causa influencia de maneira direta vários efeitos, todos condicionalmente independentes, dada a causa, tem-se a distribuição conjunta total:

$$P(Causa, Efeito_1, \dots, Efeito_n) = P(Causa) \prod_i P(Efeito_n|Causa)$$

Classificador Naive-Bayes

- Junto com árvores de decisão, redes neurais, método dos vizinhos mais próximos, Naive-Bayes é um dos métodos de aprendizagem mais práticos.
- Quando usar:
  - Quando o conjunto de dados disponível é de tamanho moderado à grande
  - Quando os atributos que descrevem as instâncias são condicionalmente independentes das classes
- Aplicações bem sucedidas:
  - Diagnóstico médico
  - Classificação de documentos textuais

Classificador Naive-Bayes

- Tarefa de atribuir a uma nova instância a classe mais provável dados os valores de atributos que descrevem a instância.

$$v_{MAP} = \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 \ldots a_n)$$

Teorema de Bayes

$$v_{MAP} = \arg\max_{v_j \in V} \frac{P(v_j \mid a_1, a_2 \ldots a_n) P(v_j)}{P(a_1, a_2 \ldots a_n)}$$

$$= \arg\max_{v_j \in V} P(v_j \mid a_1, a_2 \ldots a_n) P(v_j)$$

Por que P(a1,...,an) sumiu?

Classificador Naive-Bayes

- Para determinar a classe, basta estimar os termos da equação a partir do conjunto de treinamento

$$v_{MAP} = \arg\max_{v_j \in V} P(a_1, a_2, \ldots, a_n \mid v_j) P(v_j)$$

- $P(v_j)$ é simplesmente a frequência da classe v_j no conjunto de treinamento.

Classificador Naive-Bayes

- Calcular $P(a_1, a_2, \ldots, a_n \mid v_j)$ é inviável, pois o número de termos como este é o número possível de instâncias (com todas as possibilidades de valores de atributos) multiplicado pelo número possível de classes v_j.
  - Processo caro
  - Necessitaria de um conjunto de treinamento muito grande para evitar o problema dos dados serem esparsos

Classificador Naive-Bayes

- O classificador naive-bayes resolve o problema com a suposição ingênua de que os atributos são condicionalmente independentes dada a classe

- A conjunção de atributos <a1, a2, ..., an> corresponde à multiplicação das probabilidades de cada atributo

$$P(a_1, \ldots, a_n \mid v_j) = \prod_i P(a_i \mid v_j)$$

Classificador Naive-Bayes

- Reescrevendo a equação anterior:

$$v_{map} = \arg\max_{v_j \in V} P(a1, a2, ..., an \mid v_j) P(v_j)$$

$$\Downarrow$$

$$v = \arg\max_{v_j \in V} P(v_j) \prod_i P(ai \mid v_j)$$

- Agora, o número de P(ai|vj) é o número de valores do atributo ai multiplicado pelo número de classes, o que é viável, pois é muito menos do que o cálculo anterior de P(a1, a2, ..., an|vj)

Classificador Naive-Bayes

- Portanto,
  - O processo de aprendizado consiste em estimar $P(v_j)$ e $P(a_i|v_j)$ a partir do conjunto de treinamento.
  - O conjunto de estimativas (probabilidades) consiste na hipótese aprendida, usada para classificar novos exemplos.

Exemplo 1

- Considere o conjunto de treinamento:

dia	aparência	temperatura	umidade	vento	jogar_tênis
D1	ensolarado	quente	alta	fraco	**não**
D2	ensolarado	quente	alta	forte	**não**
D3	nublado	quente	alta	fraco	**sim**
D4	chuva	moderada	alta	fraco	**sim**
D5	chuva	fria	normal	fraco	**sim**
D6	chuva	fria	normal	forte	**não**
D7	nublado	fria	normal	forte	**sim**
D8	ensolarado	moderada	alta	fraco	**não**
D9	ensolarado	fria	normal	fraco	**sim**
D10	chuva	moderada	normal	fraco	**sim**
D11	ensolarado	moderada	normal	forte	**sim**
D12	nublado	moderada	alta	forte	**sim**
D13	nublado	quente	normal	fraco	**sim**
D14	chuva	moderada	alta	forte	**não**

Exemplo 1

- Classifique a instância abaixo como sim (jogar tênis) ou não
  - <aparência=ensolarado, temperatura=moderada, umidade=alta, vento=forte>

Exemplo 1

<aparência=ensolarado, temperatura=moderada, umidade=alta, vento=forte>

$$P(\texttt{Sim}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = P(\texttt{Sim})P(\texttt{Ensolarado}|\texttt{Sim})P(\texttt{Moderada}|\texttt{Sim})P(\texttt{Alta}|\texttt{Sim})P(\texttt{Forte}|\texttt{Sim})$$

$$P(\texttt{Sim}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = \frac{9}{14}*\frac{2}{9}*\frac{4}{9}*\frac{3}{9}*\frac{3}{9}$$

$$P(\texttt{Sim}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = 0.0071$$

$$P(\texttt{Nao}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = P(\texttt{Nao})P(\texttt{Ensolarado}|\texttt{Nao})P(\texttt{Moderada}|\texttt{Nao})P(\texttt{Alta}|\texttt{Nao})P(\texttt{Forte}|\texttt{Nao})$$

$$P(\texttt{Nao}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = \frac{5}{14}*\frac{3}{5}*\frac{2}{5}*\frac{4}{5}*\frac{3}{5}$$

$$P(\texttt{Nao}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = 0.0411$$

Exemplo 1

- <aparência=ensolarado, temperatura=moderada, umidade=alta, vento=forte>

$$P(\texttt{Sim}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = 0.0071$$

$$P(\texttt{Nao}|\texttt{Ensolarado},\texttt{Moderada},\texttt{Alta},\texttt{Forte}) = 0.0411$$

$$P(\texttt{jogar\_tenis} = \texttt{Sim}) = \frac{0.0071}{0.0071 + 0.0411} \approx 15\%$$

$$P(\texttt{jogar\_tenis} = \texttt{Nao}) = \frac{0.0411}{0.0071 + 0.0411} \approx 85\%$$

Referências

- RUSSEL, S.; NORVIK, P. Inteligência Artificial. Editora Campus, 2013, Capítulos 12, 13 e 14.
- Faceli et al., Inteligência Artificial – Uma Abordagem de Aprendizado de Máquina, LTC, 2015.