Aprendizado de Máquina: Aprendizado Bayesiano

Prof. Arnaldo Candido Junior
UTFPR – Medianeira

Revisão de probabilidades

- Considere a tabela ao lado
- Uma pessoa nasce com olho azul quando possui dois genes recessivos
- Para simplificar:
  - Só serão analisados olhos azuis e castanhos
  - Será considerado que os genes estão igualmente distribuídos na população

Mãe	**Pai**	**Filho**
a	a	aa
a	C	aC
C	a	Ca
C	C	CC

Revisão de probabilidades (2)

- Revisão:
  - Probabilidade a priori
  - Probabilidade conjunta
  - Probabilidade a posteriori
- Teorema de Bayes
- Hipótese ingênua

Revisão de probabilidades (3)

- Probabilidade **a priori**: calculada a partir daquilo que se sabe no início do problema
- Calcule as probabilidades dos eventos
  - A: filho tem olhos azuis
  - B: filho tem olhos castanhos
  - C: mãe forneceu gene recessivo
  - D: mãe forneceu gene dominante

Revisão de probabilidades (4)

- Para isso, é só contar as linhas da tabela que casam com os eventos (abordagem frequentista)
- p(A) = 25%
- p(B) = 75%
- p(C) = 50%
- p(D) = 50%

Revisão de probabilidades (5)

- Probabilidade **conjunta**: a chance de dois (ou mais) eventos ocorrem juntos
- Calcule
  - $p(A \wedge C)$
  - $p(B \wedge C)$
  - $p(B \wedge D)$

Revisão de probabilidades (6)

- Basta contar as linhas da tabela que casam com ambos eventos
  - $p(A \wedge C) = 25\%$
  - $p(B \wedge C) = 25\%$
  - $p(B \wedge D) = 50\%$

Revisão de probabilidades (7)

- Probabilidade **a posteori**: calculada quando se recebe uma informação nova do problema
- Calcule
  - p(A|C), lê-se p(A) dado C
  - p(B|D)

Revisão de probabilidades (8)

- Por contagem: filtre pelo evento dado e conte o evento de interesse
  - p(A|C) = 50%
  - p(B|D) = 100%
- Por fórmula:

$$p(A|B) = \frac{p(A \wedge B)}{p(B)}$$

Teorema de Bayes

- Assuma p(A) ≠ 0 e p(B) ≠ 0

$$p(A|B) = \frac{p(A \wedge B)}{p(B)} \quad \rightarrow \quad p(A \wedge B) = p(A|B)\, p(B)$$

$$p(B|A) = \frac{p(A \wedge B)}{p(A)} \quad \rightarrow \quad p(A \wedge B) = p(B|A)\, p(A)$$

- Combinando as duas equações:

$$p(A|B)\, p(B) = p(B|A)\, p(A) \quad \rightarrow \quad p(A|B) = \frac{p(B|A)\, p(A)}{p(B)}$$

Teorema de Bayes (3)

- Exercício: para a tabela da cor dos olhes:
  - Calcule pelo método de contagem:
    p(B)
    p(D)
    p(B|D)
    p(D|B)
  - Calcule pelo teorema: p(D|B)

Teorema de Bayes (4)

- Abordagem frequentista: consiste em contar linhas na tabela
- Abordagem bayesiana: consiste em usar aproximações para montar a tabela
  - Não dá para ter “clones” de um paciente para saber quantos dos clones tem determinada doença a partir dos sintomas
  - Usa-se outros pacientes com sintomas parecidos como uma aproximação do paciente para saber a doença

Hipótese Ingênua

- Hipótese Ingênua: $p(A \wedge B) = p(A)P(B)$
- Válida apenas quando A e B são dois eventos independentes
- Nem sempre é verdade: peso é completamente independente da altura?
- Exercício: use a hipótese ingênua para calcular $p(A \wedge C)$ na tabela de genética

Classificador Bayesiano Ingênuo

- Seja f: X $\rightarrow$ Y um problema de classificação
- Seja $\hat{x}$ uma instância deste problema:
  $\hat{x} = \{x_1, ..., x_n\}$, $\hat{x} \in X$
- Sejam v_i os valores dos atributos
- Desejamos calcular:

$$f = \underset{y_j \in Y}{argmax}\, p(y = y_j | x_1 = v_1 \wedge x_2 = v_2 \wedge ... \wedge x_n = v_n)$$

Classificador Bayesiano Ingênuo (2)

- Ou seja, desejamos encontrar a classe mais provável da instância
- Primeiro, vamos simplificar a notação:

$$f = \underset{y_j}{argmax}\, p(y_j \mid v_1 \wedge v_2 \wedge \ldots \wedge v_n)$$

- Pelo **teorema de bayes**

$$f = \underset{y_i}{argmax} \frac{p(v_1 \wedge v_2 \wedge \ldots \wedge v_n | y_j)\, p(y_j)}{p(v_1 \wedge v_2 \wedge \ldots \wedge v_n)}$$

Classificador Bayesiano Ingênuo (3)

- Para analisar uma dada instância, o denominador é constante, podemos simplificar o cálculo para:

$$f = \underset{y_j}{argmax}\, p(v_1 \wedge v_2 \wedge \ldots \wedge v_n | y_j)\, p(y_j)$$

- Pela **hipótese ingênua**:

$$f = \underset{y_j}{argmax}\, p(v_1 | y_j)\, p(v_2 | y_j) \ldots p(v_n | y_j)\, p(y_j)$$

$$f = \underset{y_j}{argmax}\, p(y_j) \prod_{v_i} p(v_i | y_j)$$

Classificador Bayesiano Ingênuo (4)

- Estimativa das Probabilidades $p(v_i \mid y_j)$ e $p(y_j)$ pode
  - $p'(v_i \mid y_j)$ ← estimativa de $p(v_i \mid y_j)$
  - $p'(y_j)$ ← estimativa de $p(y_j)$
- Classificador de novas instancias(x)

$$h* = \underset{y_j}{argmax}\, p'(y_j) \prod_{v_i} p'(v_i|y_j)$$

- Idealmente, h* aproxima f

Classificador Bayesiano Ingênuo (5)

Dia	Tempo	Temp.	Humid.	Vento	Jogar
D1	Sol	Quente	Alta	Fraco	Não
D2	Sol	Quente	Alta	Forte	Não
D3	Coberto	Quente	Alta	Fraco	Sim
D4	Chuva	Normal	Alta	Fraco	Sim
D5	Chuva	Frio	Normal	Fraco	Não
D6	Chuva	Frio	Normal	Forte	Não
D7	Coberto	Frio	Normal	Forte	Sim
D8	Sol	Normal	Alta	Fraco	Não
D9	Sol	Frio	Normal	Fraco	Sim
D10	Chuva	Normal	Normal	Fraco	Sim
D11	Sol	Frio	Alta	Forte	?

Classificador Bayesiano Ingênuo (6)

p(Sim) = 5/10 = 0.5
p(Não) = 5/10 = 0.5
p(Sol/Sim) = 1/5 = 0.2
p(Sol/Não) = 3/5 = 0.6
p(Frio/Sim) = 2/5 = 0.4
p(Frio/Não) = 2/5 = 0.4
p(Alta/Sim) = 2/5 = 0.4
p(Alta/Não) = 3/5 = 0.6
p(Forte/Sim) = 1/5 = 0.2
p(Forte/Não) = 2/5 = 0.4

P(Sim)P(Sol/Sim) P(Frio/Sim) P(Alta/Sim) P(Forte/Sim) =
= 0.0032

P(Não)P(Sol/Não)P(Frio/Não) P(Alta/Não) P(Forte/Não) =
= 0.0288

⇒ Jogar_Tenis (D11) = Não

Exercício

- Faça um classificador bayesiano ingênuo para:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	sim	sim	pequenas	não	saudável
José	sim	não	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

- (Luis, não, não, pequenas, sim)?
- (Laura, sim, sim, grandes, sim)?

Probabilidades nulas

- Novas instâncias podem conter valores de atributos não vistos em treinamento
  - Ex.: ceu = “nublado” na classe negativa não ocorreu
- Esses valores zeram os scores do produtório para a nova instância
  - É possível que todas as possíveis classes ficam com score 0

Probabilidades nulas (2)

- Solução intuitiva: adicione 1 em todos os nominadores

```
p(Sol/Sim)  = 1/5 →  2/5
p(Sol/Não)  = 3/5 →  4/5
p(Frio/Sim) = 2/5 →  3/5
p(Frio/Não) = 2/5 →  3/5
…
```

  - Pode alterar o resultado da classe mais provável
- Solução utilizada: suavização de Laplace, de LidStone, entre outras

Suavização de Laplace

- Suavização de Laplace usa um pseudocontador
- No nominador, somar 1
- No denominador, somar o número de valores que um atributo pode assumir
- É uma forma de **data augmentation**: cria instâncias virtuais

Suavização de Laplace (2)

- Suponha que sol tem céu tenha 3 valores e que temperatura tenha 3

```
p(Sol/Sim)   = 1/5 →   (1+1)/(5+3)
p(Sol/Não)   = 3/5 →   (1+1)/(5+3)
p(Frio/Sim)  = 2/5 →   (2+1)/(5+3)
p(Frio/Não)  = 2/5 →   (3+1)/(5+3)
```

Outra Suavização: Text Mining

- Suponha que desejamos classificar textos em humorístico, esportivo ou biografia
- Os atributos representam as palavras que podem ocorrer no texto (bag of words)
- Os valores representam quantas vezes cada palavra ocorreu
- Problema: nem todas as palavras aparecem nos documentos de cada classe
- Isso leva a probabilidades nulas

Outra Suavização: Text Mining (2)

- Existem duas formas de calcular probabilidades da palavra x_i dada a classe y_j
- Opção 1: número de documentos da classe y_j que possuem a palavra x_i pelo número de total de documentos na classe y_j
- Opção 2: total de ocorrências da palavra x_i em documentos da classe y_j pelo total geral de palavras de todos os documentos da classe y_j

Outra Suavização: Text Mining (3)

- O segundo caso ajuda a criar uma intuição sobre o processo de suavização.
  - Intuitivamente, colocamos as documentos de cada classe em uma pasta física
  - Depois, anexamos o dicionário Aurélio a cada pasta
  - Aí procedemos com a contagem
  - É garantido que cada palavra ocorrerá pelo menos uma vez em cada classe

Outra Suavização: Text Mining (4)

- $p'(y_j) = \dfrac{\text{número de documentos da classe } y_j}{\text{número total de documentos}}$
- $p'(v_i|y_j) = \dfrac{n_{ij} + 1}{n_j + |\text{Vocabulário}|}$
- n_j é o número **total** de palavras de **todos** os documentos da classe y_j
- n_{ij} é o número **total** de ocorrências da palavra x_i em **todos** os documentos da classe y_j

Log da probabilidade

- É comum usar logaritmo natural (**log-likelihood**) em implementações do classificador bayesiano ingênuo
  - Enquanto probabilidades são multiplicadas, logs são somados (mais eficiente)
  - Evita erros de arredondamento
    - Produtos de probabilidade convergem rapidamente a zero
    - A precisão dos tipos de ponto flutuante começa a deixar a deixar

Log da probabilidade (2)

- Revisão rápida de logaritmos (base 2):
  - $8 * 32 = 256$
    $2^3 * 2^5 = 2^8$
    $\log(2^3 * 2^5) = \log 2^8$
    $\log 2^3 + \log 2^5 = \log 2^8$
    $3 + 5 = 8$
  - $8 * 0.5 = 4$
    $2^3 * 2^x = 2^2$
    $x = -1$

Log da probabilidade (3)

- Seja **p** uma probabilidade e **q** seu logaritmo natural
  - Para recuperar p, é só fazer:
    $p = e^q$
- Assim, a fórmula será revisada para:

$$h* = \underset{y_j}{argmax}\left(\log p'(y_j) + \sum_{v_i} \log p'(v_i|y_j)\right)$$

Atributos contínuos

- Opção 1: discretizar
- Opção 2: considerar que os dados seguem uma distribuição normal
  - Usar a função de densidade de probabilidade normal para estimar a probabilidade

Redes Bayesianas

- Alternativa ao Bayesiano ingênuo
- Busca encontrar e incluir probabilidades dependentes no cálculo ao invés de recorrer à hipótese ingênua
- É um grafo acíclico e dirigido: cada nó da rede representa uma variável aleatória
  - Um conjunto de ligações ou arcos dirigidos conectam pares de nós

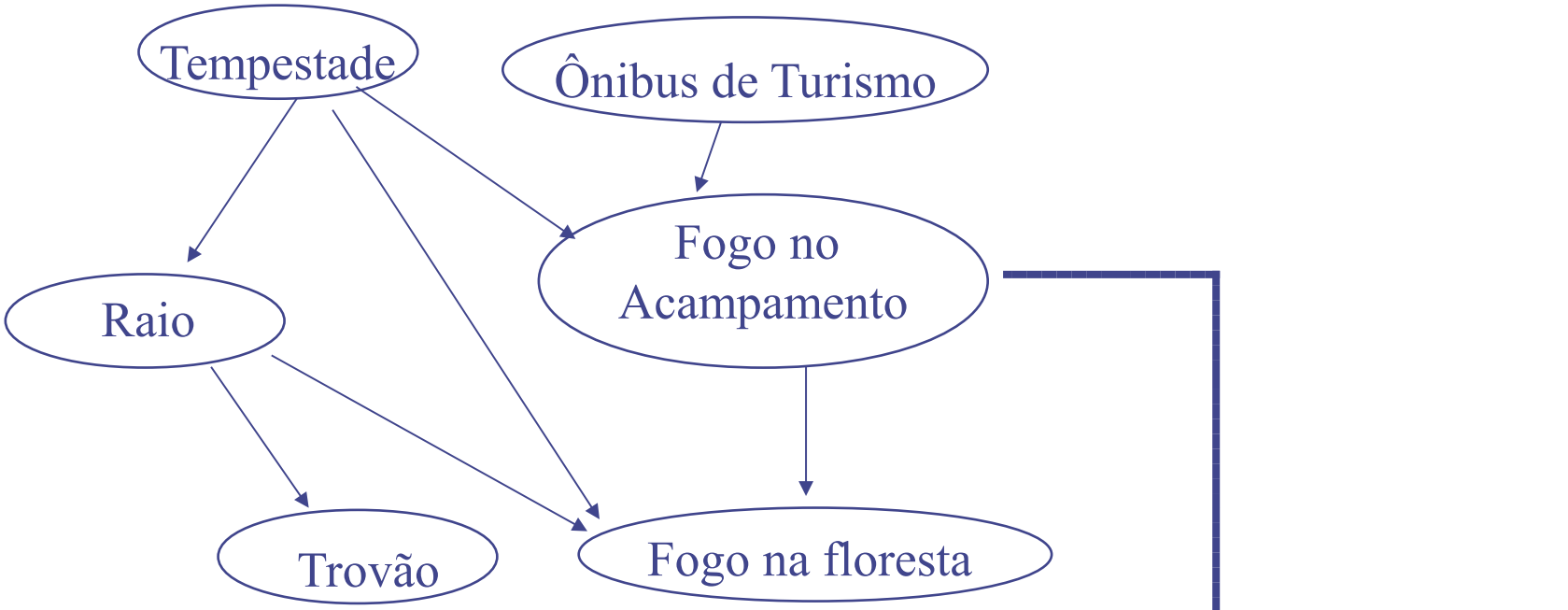
Redes Bayesianas (2)

- Cada nó recebe arcos dos nós que tem influência direta sobre ele (nós pais).
- Cada nó possui uma tabela de probabilidade condicional associada que quantifica os efeitos que os pais têm sobre ele
- Variações do algoritmo:
  - Projetista informa dependências entre atributos
  - Algoritmo encontra sozinho dependências entre atriubtos

Redes Bayesianas (3)

- Como rede Bayesiana busca encontrar probabilidades conjuntas, é mais adequada para datasets maiores
  - Datasets grandes tem mais chances de conter probabilidades conjuntas
- Exemplo a seguir:
  - Rede para encontrar fogo na floresta
  - Para simplificar, pensar no caso fogo no acampamento

Redes Bayesianas (4)



	T,O	T,¬O	¬T, O	¬T, ¬O
FA	0.4	0.1	0.8	0.2
¬FA	0.6	0.9	0.2	0.8

Fogo no Acampamento

Distribuição de Probabilidade:
$P(FA/T,O)$

Redes Bayesianas (4)

- No exemplo:
  - Estamos interessados em saber se há fogo no acampamento de posse das demais informações
  - Ônibus de turismo pode causar fogo no acampamento
  - Tempestade pode apagá-lo
  - Ou ainda, causar relâmpagos que causam fogo

Créditos

- Baseado em material da UFPE, disponível em www.cin.ufpe.br/~if684/aulas/Aprend_Bayestbl.ppt