Aprendizado Probabilístico: Bayes

SCC-230 – Inteligência Artificial

Prof. Thiago A. S. Pardo

1

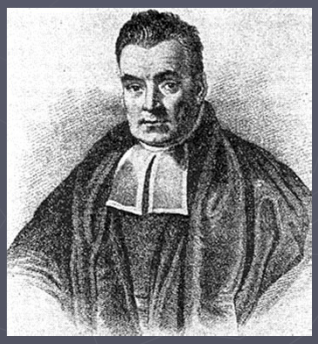
Pergunta

- O que você sabe sobre Bayes?

2

Resposta 1

- Pastor presbiteriano



3

Resposta 2

- Em 1931, teve publicado (após a morte) o livro *Benevolência divina, ou uma tentativa de provar que a principal finalidade da providência divina é a felicidade de suas criaturas*
- Em 1936, teve publicado (após a morte) o livro *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of the Analyst*

4

Resposta 3

- Criador do teorema de Bayes, publicado em 1764

5

Relembrando...

- Para dois evento A e B
  - P(A ^ B) = P(A|B)*P(B) = P(B|A)*P(A)
  - P(A v B) = P(A)+P(B)-P(A ^ B)
  - P(A|B) = P(B|A)*P(A)/P(B)
  - Se eventos $A_1 \ldots A_n$ são mutuamente exclusivos e suas probabilidades somam 1, então

$$P(B) = \sum_{i=1}^{n} P(B \mid A_i) * P(A_i)$$

6

Por que Bayes?

- P(doença|sintoma) vs. P(sintoma|doença)

7

Por que Bayes?

- P(doença|sintoma) vs. P(sintoma|doença)

P(doença|sintoma)=P(sintoma|doença)*P(doença)/P(sintoma)

8

Por que Bayes?

- P(doença|sintoma) vs. P(sintoma|doença)

P(doença|sintoma)=P(sintoma|doença)*P(doença)/P(sintoma)

- o sintoma é o que se observa, e a doença é o que se quer descobrir
  - P(doença|sintoma)
- mas quem causa o sintoma é a doença, e não o inverso
  - P(sintoma|doença)
- P(doença|sintoma) pode ser "tendencioso" e "temporal"

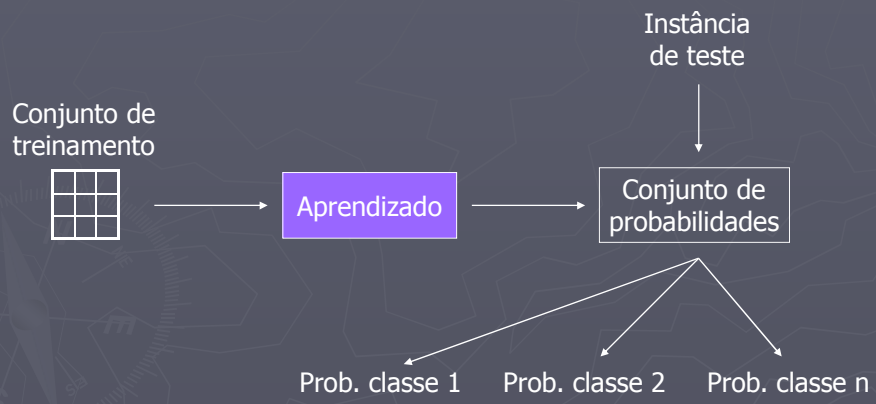
9

Aprendizado probabilístico

- Calculam-se probabilidades para as hipóteses induzidas
- Aprendizado bayesiano
  - Maior representante deste paradigma
  - Simplicidade, elegância e, mais importante, bons resultados

10

Aprendizado probabilístico: esquema geral



11

Características

- Cada exemplo de treinamento pode incrementar ou decrementar a probabilidade de uma hipótese
  - Mais flexibilidade na classificação
- Conhecimento a priori pode ser combinado com os dados observados para determinar a probabilidade final de uma hipótese
- Pode-se acomodar hipóteses que fazem predições probabilísticas
  - Por exemplo, o paciente tem 93% de chance de cura

12

Dificuldades

- Requer cálculo de muitas probabilidades
- Problemas para estimar as hipóteses "ótimas"
  - Não se faz isso, em geral

13

Aprendizado bayesiano: naive-bayes

- *Naive* ou *naive*: ingênuo
  - Apesar de ingênuo, o classificador naive-bayes tem desempenho muito bom
- Aplicável em tarefas de aprendizado em que
  - Cada instância x é descrita por um conjunto de atributos e valores
  - A classe pertence a um conjunto finito de valores de V
    - Jogo de tênis!

14

Aprendizado bayesiano: naive-bayes

- Tarefa de atribuir a uma nova instância a classe mais provável v (*maximum a posteriori* - map) dados os valores de atributos <a1, a2, ..., an> que descrevem a instância

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

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

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

Teorema de Bayes

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

15

Aprendizado bayesiano: naive-bayes

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

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

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

• Calcular $P(a1,...,an|v_j)$ é praticamente 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

16

Aprendizado bayesiano: 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,...,an> corresponde à multiplicação das probabilidades de cada atributo

$$P(a1,...,an \mid v_j) = \prod_i P(ai \mid v_j)$$

17

Aprendizado bayesiano: naive-bayes

- Reescrevendo a equação anterior

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

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

Agora, o número de $P(ai|v_j)$ 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,...,an|vj)

18

Aprendizado bayesiano: naive-bayes

- O processo de aprendizado consiste em estimar $P(v_j)$ e $P(ai|v_j)$ a partir do conjunto de treinamento
- O conjunto de estimativas (probabilidades) consiste na hipótese aprendida, usada para classificar novos exemplos
- Quando a independência condicional se observa nos dados, então o naive-bayes induz uma hipótese ótima
  - Qual a frequência disso?

19

Naive-bayes: exemplo

- Considerando 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**

20

Naive-bayes: exemplo

- Desejamos classificar a instância abaixo como sim (jogar tênis) ou não

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

$$v = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(ai \mid v_j)$$

$$v = \underset{v_j \in \{sim,não\}}{\operatorname{argmax}} P(v_j) * P(\text{aparência} = \text{ensolarado} \mid v_j) * P(\text{temperatura} = \text{moderada} \mid v_j) * P(\text{umidade} = \text{alta} \mid v_j) * P(\text{vento} = \text{forte} \mid v_j)$$

Instanci-ando a equação

21

Naive-bayes: exemplo

- Para calcular a melhor classe, precisamos apenas de 10 probabilidades
  - Probabilidade $P(v_j)$ das classes: P(sim) e P(não)
  - Cada probabilidade $P(ai|v_j)$ para as duas possíveis classes
    - P(aparência=ensolarado | sim) e P(aparência=ensolarado | não)
    - P(temperatura=moderada | sim) e P(temperatura=moderada | não)
    - P(umidade=alta | sim) e P(umidade=alta | não)
    - P(vento=forte | sim) e P(vento=forte | não)

$$v = \underset{v_j \in \{sim,não\}}{\operatorname{argmax}} P(v_j) * P(\text{aparência} = \text{ensolarado} \mid v_j) * P(\text{temperatura} = \text{moderada} \mid v_j) * P(\text{umidade} = \text{alta} \mid v_j) * P(\text{vento} = \text{forte} \mid v_j)$$

22

Naive-bayes: exemplo

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**
		ente	normal	fraco	**sim**
		erada	alta	forte	**não**

P(sim)=9/14

23

Naive-bayes: exemplo

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				**sim**
D5	chuva				**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**

P(não)=5/14

24

Naive-bayes: exemplo

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**

P(vento=forte | sim)=3/9
P(vento=forte | não)=3/5

25

Naive-bayes: exemplo

- Depois de fazer todos os cálculos...
  - P(sim) * P(aparência=ensolarado | sim) * P(temperatura=moderada | sim) * P(umidade=alta | sim) * P(vento=forte | sim) = 0.0053
  - P(não) * P(aparência=ensolarado | não) * P(temperatura=moderada | não) * P(umidade=alta | não) * P(vento=forte | não) = 0.0206

a classe escolhida é **não** jogar tênis

26

Naive-bayes: exemplo

- Normalizando-se os resultados
  - P(sim) * P(aparência=ensolarado | sim) * P(temperatura=moderada | sim) * P(umidade=alta | sim) * P(vento=forte | sim) = 0.0053/(0.0053+0.0206) = 0.205
  - P(não) * P(aparência=ensolarado | não) * P(temperatura=moderada | não) * P(umidade=alta | não) * P(vento=forte | não) = 0.0206/(0.0053+0.0206) = 0.795

27

Exercício: joga ou não tênis?

- <aparência=chuva, temperatura=fria, umidade=alta, vento=forte>

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**

Exercício: joga ou não tênis?

- <aparência=chuva, temperatura=fria, umidade=alta, vento=?>

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**

Aprendizado bayesiano: naive-bayes

- Como lidar com probabilidades baixas?
  - Por exemplo, P(vento=forte|sim)=1/1000=0.001?
  - Uma probabilidade 0 ou próxima de 0 pode zerar o resultado do cálculo de probabilidades anteriores (pois é uma multiplicação de vários valores)
    - Acontece com dados esparsos ou pouco representativos do problema
- Uma forma de lidar com isso: estimativa-m
  - Seja nc o número de exemplos com um determinado atributo para uma classe em particular
  - Seja n o número de exemplos com uma classe em particular
    - P(vento=forte|sim)=nc/n=1/1000
  - Solução: $\frac{nc + m \times p}{n + m}$

30

Aprendizado bayesiano: naive-bayes

- Como lidar com probabilidades baixas?
  - Por exemplo, P(vento=forte|sim)=1/1000=0.001?
  - Uma probabilidade 0 ou próxima de 0 pode zerar o resultado do cálculo de probabilidades anteriores (pois é uma multiplicação de vários valores)
    - Acontece com dados esparsos ou pouco representativos do problema
- Uma forma de lidar com iss
  - Seja nc o número de exemplo uma classe em particular
  - Seja n o número de exemplos
    - P(vento=forte|sim)=nc/n=1/1000
  - Solução:

$$\frac{nc + m \times p}{n + m}$$

p: estimativa a priori da probabilidade desejada; em geral, é calculada como 1/k, em que k é o número de possíveis valores do atributo em questão; no caso do vento, k=2 e, portanto, p=0.5

m: *equivalent sample size*; aumenta n com um número adicional de m amostras virtuais

31

Naive-bayes

- Lida facilmente com valores ausentes para atributos
  - Como?
- Atributos redundantes são ruins e podem deixar o processo tendencioso
  - Por que?
- Qual é a hipótese (ou as hipóteses) produzida por esse método?

32

Aplicação: classificação textual

(Mitchell, 1997)

- Classificação textual: tarefa de classificar textos segundo algum critério
  - Textos sobre um assunto X
  - Textos que um usuário Y goste
- Uma vez aprendida uma hipótese, os textos podem ser filtrados para quaisquer finalidades
- O método naive-bayes está entre os melhores métodos para classificação de textos
  - Muitos sistemas utilizam

33

Aplicação: classificação textual

- Situação
  - Espaço de instâncias X consistindo de todos os possíveis textos
  - Cada instância classificada como um valor de V
    - Classes: uma pessoa julga o texto interessante (*gosta*) ou não (*não_gosta*)
  - Tarefa: dado um novo texto, decidir se ele é interessante ou não

34

Aplicação: classificação textual

- Questões principais
  - Como representar um texto qualquer em termos de atributos e valores?
  - Como estimar as probabilidades necessários para o naive-bayes?

35

Aplicação: classificação textual

- Abordagem inicial
  - Atributos e valores em um texto
    - Atributos: as posições de cada palavra no texto
    - Valores: as próprias palavras
  - Exemplo: "Esses slides são sobre..."
    - <1,esses>, <2,slides>, <3,são>, etc.

Note que, quanto maior o texto, mais atributos haverá.

36

Aplicação: classificação textual

- Supondo que
  - Temos 700 textos da classe não_gosta e 300 textos da classe gosta no conjunto de treinamento
  - Um dos textos de teste que queremos classificar é o que começa com a sentença "Esses slides são sobre..."

$$v = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(ai \mid v_j)$$

$$v = \underset{v_j \in \{gosta, não\_gosta\}}{\operatorname{argmax}} P(v_j) * P(a1 = "esses" \mid v_j) * P(a2 = "slides" \mid v_j) * P(a3 = "são" \mid v_j) * ...$$

37

Aplicação: classificação textual

- Portanto, determinação da classe que maximiza a probabilidade de se observar as palavras do texto
- Questão: suposição da independência condicional se verifica?

  Não, pois as palavras do texto não são escolhidas e alocadas ao acaso

38

Aplicação: classificação textual

- Precisam ser calculadas as probabilidades
  - $P(v_j)$
  - $P(ai=palavra_i|v_j)$
- $P(v_j)$ com base na fração de cada classe no conjunto de treinamento
  - P(gosta)=300/1000=0.3
  - P(não_gosta)=700/1000=1-P(gosta)=0.7
- Calcular $P(ai=palavra_i|v_j)$ é mais difícil, pois necessitamos saber todas as palavras que ocorrem em cada posição para todos os textos para cada uma das classes
  - Inviável

39

Aplicação: classificação textual

- Outra simplificação para lidar com o problema de calcular $P(ai=palavra_i|v_j)$
  - Assumimos que a probabilidade de encontrar uma palavra qualquer em uma posição específica é a mesma de encontrá-la em uma outra posição
    - $P(ai=palavra_i|v_j)$ ➔ $P(palavra_i|v_j)$
      - Diminui o número de cálculos necessários
      - Ameniza o possível problema dos dados serem esparsos

40

Aplicação: classificação textual

- Aplicação real para classificação de textos jornalísticos de acordo com o grupo de notícias em que apareceram
  - 20 grupos de notícias (classes) e 1.000 artigos por grupo
  - 20.000 textos: 2/3 para treino e 1/3 para teste
    - Taxa de acerto com o classificador: 89%
    - Taxa de acerto com a classe majoritária: **?**
- Pequena modificação no algoritmo anterior: as 100 palavras mais freqüentes do texto (*stopwords*) eram removidas
  - Por que se faz isso?

41

Aplicação: classificação textual

- Aplicação real para classificação de textos jornalísticos de acordo com o grupo de notícias em que apareceram
  - 20 grupos de notícias (classes) e 1.000 artigos por grupo
  - 20.000 textos: 2/3 para treino e 1/3 para teste
    - Taxa de acerto com o classificador: 89%
    - Taxa de acerto com a classe majoritária: 5%
- Pequena modificação no algoritmo anterior: as 100 palavras mais freqüentes do texto (*stopwords*) eram removidas
  - Por que se faz isso?

42

Naive-bayes

- Exercício em duplas
  - A USP lhe contrata para desenvolver um sistema que julgue a qualidade das aulas dos professores
    - Você decidiu usar aprendizado de máquina para tratar a questão, mais especificamente, o método naive-bayes
  - Modele a tarefa como um problema de aprendizado de máquina
    - Atributos, classes
    - Número de instâncias necessário, como conseguir os dados
    - Como usar o sistema
    - Etc.
  - Ao fim, responda: naive-bayes é um bom método para isso?

43