APRENDIZADO BAYESIANO

André Câmara

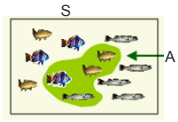
Inteligência Artificial

REVISÃO DE PROBABILIDADE

Conceitos Básicos

- Estamos realizando um evento aleatório (pegar um peixe no mar)
- Espaço amostral *S*
  - Conjunto de todas a possibilidades
- Um evento *A*
  - Um subconjunto de *S*
- Lei da probabilidade
  - Regra que atribui uma probabilidade aos eventos de um experimento

$A \rightarrow P(A)$



Probabilidade *a priori*

- Grau de crença acordado para a proposição na ausência de quaisquer outras informações.
  - Também conhecida como **probabilidade incondicional**.
  - Ex. Se de uma amostra de 10 pacientes apenas 1 apresentar Cárie, a seguinte probabilidade *a Priori* deve ser considerada.

P(Cárie = verdadeiro) = 0.1 ou P(Cárie = falso) = 0.9

- Seu uso deve ser restrito apenas a casos onde não haja a presença de nenhuma outra informação relacionada à variável aleatória.

Probabilidade a priori

- Probabilidade a priori (nenhuma outra informação é conhecida)
  - Ex.: P(A) = 0.43
- Podemos também representar as probabilidades de todos os valores possíveis de uma dada variável como um vetor

$$P(Weather = sunny) = 0.7$$
$$P(Weather = rain) = 0.2$$
$$P(Weather = cloudy) = 0.08$$
$$P(Weather = snow) = 0.02 .$$

$$\mathbf{P}(Weather) = \langle 0.7, 0.2, 0.08, 0.02 \rangle$$

Também chamado de **distribuição de probabilidade** da variável aleatória *Weather*

Axiomas Básicos da Probabilidade

- P(A) >= 0
- P(S) = 1
- P(A U B) = P(A)+P(B) se A e B forem mutuamente exclusivos
  - Eventos que não ocorrem simultaneamente, ou seja A ∩ B = ø
  - **Regra da soma**
  - Ex: P(weather='sunny' U weather='rain') = 0.7+0.2 = 0.9
- Caso contrário
  - P(A U B) = P(A) + P(B) – P(A ∩ B)
  - Principio da Inclusão-exclusão
- P(A) + P(~A)= 1

Probabilidade Condicional

- Dado que o agente obtém alguma evidência relativa às variáveis aleatórias, o uso das probabilidades *a priori* (por si só) não é mais recomendado.
  - Ao invés disso, usa-se as probabilidades condicionais ou ***a posteriori***.
  - P(a|b) – probabilidade de o evento a ocorrer dado que o evento b ocorreu.
    - a hipótese, b evidência
  - P(cárie|dordedente) = 0,8
- Probabilidades condicionais podem ser definidas em termos de probabilidades incondicionais
  - P(a|b) = P(a ∩ b) / P(b)

Probabilidade Condicional

- A probabilidade de ocorrer um evento, na condição de que outro evento já tenha ocorrido.

$$P(A \mid B) = \frac{P(A \text{ ú } B)}{P(B)}$$

- Considere o seguinte exemplo:
  - 250 alunos estão matriculados no primeiro ano
    - 100 homens e 150 mulheres
    - 110 cursam física e 140 química

Sexo\Discip.	Física	Química	Total
H	40	60	100
M	70	80	150
Total	110	140	250

Probabilidade Condicional

- Um aluno é sorteado ao acaso. Qual a probabilidade de que esteja cursando química dado que seja mulher.

$$P(Q \mid M) = \frac{P(Q \text{ ú } M)}{P(M)} = \frac{\frac{80}{250}}{\frac{150}{250}} = \frac{80}{150} = 0.53$$

Sexo\Discip.	Física	Química	Total
H	40	60	100
M	70	80	150
Total	110	140	250

Probabilidade Condicional

- Ex: a (cárie = verdadeiro), b (dor de dente = verdadeiro)

amostra	cárie	Dor de Dente
1	verdadeiro	verdadeiro
2	verdadeiro	falso
3	falso	falso
4	falso	falso
5	falso	verdadeiro
6	verdadeiro	verdadeiro

- P(a ∩ b) = 2/6
- P(b) = 3/6
- P(a | b) = (2/6) / (3/6)= 0.66

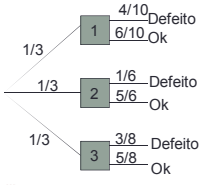
Probabilidade Total

- Outra forma de especificar a probabilidade condicional
- Uma sequência finita de experimentos na qual cada experimento tem um número finito de resultados com uma determinada probabilidade é chamada de processo estocástico finito.
- Árvore Bayesiana é uma boa ferramenta para visualização do problema.
- A probabilidade final é calculada pela lei da probabilidade final.

$$P(A) = \sum_{k=1}^{n} P(A \mid B_k) P(B_k)$$

Probabilidade Total: Exemplo

- Considere 3 caixas
  - **Caixa 1** tem 10 lampadas, das quais 4 com defeito
  - **Caixa 2** tem 6 lâmpadas, das quais 1 com defeito
  - **Caixa 3** tem 8 lâmpadas, das quais 3 com defeito.
- O problema consiste em saber a probabilidade de uma lâmpada ser defeituosa P(A), ao selecionar uma caixa aleatoriamente e depois selecionar uma lâmpada aleatoriamente.



$$P(A) = \sum_{k=1}^{n} P(A \mid B_k) P(B_k)$$

Baseado no conceito de probabilidade total, temos como probabilidade P(A)

$$P(A) = \left(\frac{1}{3} \times \frac{4}{10}\right) + \left(\frac{1}{3} \times \frac{1}{6}\right) + \left(\frac{1}{3} \times \frac{3}{8}\right) = 0.31$$

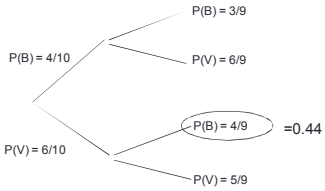
Eventos Independentes

- Dois eventos são ditos independentes se P(A∩B) = P(A) * P(B)
- Logo, pela regra da probabilidade condicional, se A e B são independentes,

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A) \times P(B)}{P(B)} = P(A)$$

Exemplo

- Suponha que uma urna contenha 4 bolas brancas e 6 vermelhas. Vamos sortear duas bolas (sem reposição) em momentos distintos. Qual a probabilidade de sair uma bola branca seguida de uma vermelha
  - Eventos não independentes



Exemplo (cont)

- Agora considere o exemplo anterior com reposição, ou seja, eventos independentes.
- A probabilidade de sair uma bola branca seguida de uma vermelha
  - P(B,V) = P(B) X P(V) = 4/10 X 6/10 = 0.24

Teorema de Bayes

- Pode ser usado para calcular a probabilidade de que um evento venha a ocorrer dado que já conhecemos um fragmento relacionado de informação.

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

- Deduzindo:

$$P(B \cap A) = P(B \mid A)P(A)$$
$$P(B \cap A) = P(A \cap B) = P(A \mid B)P(B)$$
$$P(B \mid A)P(A) = P(A \mid B)P(B)$$

Teorema de Bayes

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

- Duas formas básicas de ser utilizada:
  - Para aprender uma hipótese h a partir do conjunto de dados D:

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

  - P(h): probabilidade *a priori* de que h está correta
  - P(h|D): probabilidade *a posteriori* de que h está correta
  - P(D): probabilidade *a priori* de D
  - P(D|h): probabilidade de observar D dado que h aconteceu
  - Classificar um exemplo *e*:

P(classe | *e*) = P(*e* | classe)P(classe) / P(*e*)

Escolha das Hipóteses

- Geralmente, existe um espaço de hipóteses (H), e deseja-se a **hipótese (h∈ H) mais provável**, observados os dados de treinamento (D)
  - Uma aproximação do valor real
- Hipótese de **máxima a posteriori** h_{MAP}

$$\begin{aligned} h_{MAP} &= \arg\max_{h \in H} P(h / D) \\ &= \arg\max_{h \in H} \frac{P(D / h)P(h)}{P(D)} \\ &= \arg\max_{h \in H} P(D / h)P(h) \end{aligned}$$

Teorema de Bayes: Exemplo

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Um médico sabe que a meningite causa torcicolo em 50% dos casos. Porém, o médico sabe que a meningite atinge 1/50.000 e também que a probabilidade de se ter torcicolo é de 1/20.
- Usando Bayes pra saber a probabilidade de uma pessoa ter meningite dado que ela está com torcicolo

P(T|M) = 0.5
P(M) = 1/50000
P(T) = 1/20

$$P(M \mid T) = \frac{P(M) \times P(T \mid M)}{P(T)} = \frac{1/50000 \times 0.5}{1/20} = 0.0002$$

Teorema de Bayes: Exercício

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Considere o sistema de classificação de peixes visto anteriormente. Para essa época do ano, sabe-se que a probabilidade de pescar salmão é maior que pescar robalo, P(salmão) = 0.82 e P(robalo) = 0.18.
- Suponha que a única característica que você pode contar é a intensidade do peixe ou seja, se ele é claro ou escuro. Sabe-se que 49.5% dos salmões tem intensidade clara e que 85% dos robalos tem intensidade clara.
- Calcule **a probabilidade de ser salmão dado que o peixe amostrado tem intensidade clara**.

$$P(S \mid C) = \frac{P(S) \times P(C \mid S)}{P(C)} = \frac{0.82 \times 0.495}{\underbrace{0.82 \times 0.495 + 0.18 \times 0.85}} = 0.726$$

Probabilidade total

Aprendizagem de máquina

- P(h): probabilidade a priori de que h está correta
- E se não tivermos preferência por nenhum h?
  - Assumir P(h)=P(h') para todo h, h' ∈ H
- Nesses casos, h_{MAP} é chamado de hipótese de máximo verossimilhança (*maximum likelihood hypothesis*) h_{ML}

$$h_{ML} = \arg\max_{h_i \in H} P(D / h_i)$$

Aprendizado pelo método da força bruta

- Para cada hipótese h ∈ H, calcule a probabilidade a posteriori

$$P(h \mid D) = \frac{P(D \mid h)P(h)}{P(D)}$$

- Escolha a hipótese h_{MAP} de maior probabilidade a posteriori

$$h_{MAP} = \arg\max_{h \in H} P(h \mid D)$$

Aprendizado pelo método da força bruta (cont.)

- Suponha que D é o conjunto de exemplos
  D = {(x_1, $f(x_1)$), … (x_m, $f(x_m)$)}
- Cálculo de P(D|h)
  - P(D|h) = 1, se h é consistente com D (ou seja $f(x_i) = h(x_i)$, $\forall (x_i, f(x_i)) \in D$)
  - P(D|h) = 0, caso contrário

Aprendizado pelo método da força bruta (cont.)

- Assumir que todas as hipóteses tem a mesma probabilidade

$$P(h) = \frac{1}{|H|}$$

$$P(D) = \sum P(D \mid h_i)P(h_i) = \frac{|VS_{H,D}|}{|H|}$$

- onde $VS_{H,D}$ é subconjunto de hipóteses de H consistentes com D

Aprendizagem pelo Método da Força Bruta (cont.)

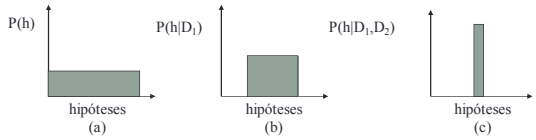
- Uma vez definidas as probabilidades a *priori*, podemos voltar para o primeiro passo:
- Cálculo da probabilidade *a posteriori*:
  - Se h é inconsistente com D

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)} = \frac{0 * P(h)}{P(D)} = 0$$

  - Se h é consistente com D

$$P(h/D) = \frac{P(D/h)P(h)}{P(D)} = \frac{1 * \frac{1}{|H|}}{\frac{|VS_{H,D}|}{|H|}} = \frac{1}{|VS_{H,D}|}$$

Evolução das Probabilidades a Posteriori



- em (a) todas as hipóteses tem a mesma probabilidade
- em (b) e (c) a medida que novos exemplos são adquiridos, a probabilidade a posteriori das hipóteses inconsistentes se tornam nulas, enquanto que a probabilidade a posteriori das hipóteses que restaram no espaço de versões aumenta

Método da Força Bruta – Observações

- Na prática:
  - só funciona quando o conceito verdadeiro está contido no espaço de hipóteses
  - funciona com dados **sem ruído**
- No cálculo da probabilidade P(h|D) pode se levar em consideração
  - erro obtido pela hipótese h no conjunto D
  - o tamanho da hipótese

Classificador Bayesiano Ingênuo (Naive Bayes)

- Método de classificação simples e popular
- Baseado na regra de Bayes, associada a suposição de independência condicional
  - Raramente ocorre na prática
  - Entretanto, normalmente funciona bem
- Aplicado com sucesso:
  - Classificação de documentos de texto
  - Diagnóstico

Classificador Bayesiano Ingênuo (Naive Bayes)

- Dados os valores dos atributos, qual o valor mais provável do atributo de classe?

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

- **Problema**: muitos dados necessários para estimar $P(a_1,\ldots,a_n|v_j)$

Classificador Bayesiano Ingênuo (Naive Bayes)

- Assumir que atributos são independentes, dada a classe
  - $P(a_1,\ldots,a_n|v_j) = P(a_1|v_j)P(a_2|v_j)\ldots P(a_n|v_j)$
- Sendo assim, v_{MAP} é dado por:

$$v_{NB} = \arg\max P(v_j) \prod_i P(a_i \mid v_j)$$

Naive Bayes - Algoritmo

```
NaiveBayesLearn(examples)
 para cada atributo alvo v_j
    P^(v_j) = estimar P(v_j)
        Para cada valor de atributo a_i de cada atributo a
            P^(a_i|v_j) = estimar P(a_i|v_j)

 ClassifyNewInstance(x)
```

$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j)$$

Naive Bayes - Estimar parâmetros

- Como estimar $P(v_j)$ e $P(a_i|v_j)$?
  - Estimação padrão estatística
  - Estimar a probabilidade a partir das amostras
    - $P(v_j) = count(v_j) / N$
    - $P(a_i|v_j) = count(a_i) / count(v_j)$
- Exemplo:
  - 100 exemplos com 70+ e 30 –
  - P(+)=0.7 e P(-) = 0.3
  - Entre os 70 exemplos positivos, 35 com a_1="Ensolarado"
  - P(a_1="Ensolarado"|+) = 35 / 70 = 0.5


$$v_{NB} = \arg\max_{v_j \in V} \hat{P}(v_j) \prod_i \hat{P}(a_i \mid v_j)$$

Classificador Bayesiano Ingênuo: Exemplo

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	?

P(Sim) = 5/10 = 0.5
P(Não) = 5/10 = 0.5
P(Tempo=Sol | Sim) = 1/5 = 0.2
P(Tempo=Sol | Não) = 3/5 = 0.6
P(Temp.=Frio | Sim) = 2/5 = 0.4
P(Temp.=Frio | Não) = 2/5 = 0.4
P(Humid.=Alta | Sim) = 2/5 = 0.4
P(Humid.=Alta | Não) = 3/5 = 0.6
P(Vento=Forte | Sim) = 1/5 = 0.2
P(Vento=Forte | Não) = 2/5 = 0.4

P(Sim) (P(Tempo=Sol | Sim) P(Temp.=Frio | Sim) P(Humid.=Alta | Sim) P(Vento = Forte/Sim)) = P(Sim | D11) = 0.0032

P(Não) (P(Tempo=Sol | Não) P(Temp.=Frio | Não) P(Humid.=Alta | Não) P(Vento = Forte/Não)) = P(Não| D11) = 0.0288

⇒ **Jogar Tenis (D11) = Não**

Naive Bayes - Dificuldades

- Suposição de independência condicional quase sempre violada, mas funciona surpreendentemente bem

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

- O que acontece se nenhuma das instancias classificadas como v_j tiver o valor a_i?

$$P'(a_i \mid v_j) = 0 \Rightarrow P'(v_j) \prod_i P'(a_i \mid v_j) = 0$$

  - Pseudocounts


Exemplo Classificando texto

- Objetivo
  - Aprender quais noticias são interessantes
  - Aprender a dizer qual a fonte de cada noticia
  - Aprender a classificar paginas web por tópico
- Naive Bayes funciona bem nessas tarefas
  - A forma de representar os exemplos é vital para o sucesso da aplicação


Representação

- Classificar documentos em duas classes
  - v_j = {'interesse', 'não-interesse}
- Variáveis a_1,..., a_n são palavras de um vocabulário e $P'(a_i \mid v_j)$ é a freqüência com que a palavra a_i aparece entre os documentos da classe v_j
- $P'(v_j)$ = número de documentos da classe v_j / número total de documentos


Algoritmo

- $P'(a_i \mid v_j) = \frac{n_{ij} + 1}{n_j + |\text{Vocabulário}|}$

onde n_j é o número total de palavras nos documentos da classe v_j e n_{ij} é o número de ocorrências da palavra a_i nos documentos da classe v_j.

Usa-se m = |Vocabulário| e
p = 1/ |Vocabulário| (assumindo que cada palavra tem a mesmo probabilidade de ocorrência)

Resultados

- Jochims 1996
- Classificar documentos de acordo com a fonte (newsgroups)

- Naive Bayes: 89% de acerto
  - 100 palavras mais frequentes removidas (the, and, of,…)
  - Palavras com menos de 3 ocorrências removidas
  - Vocabulário final com 38.500 palavras