Objectivos

- Adquirir a noção de aprendizagem bayesiana
- Explicar o Teorema de Bayes para as probabilidades condicionais
- Descrever o algoritmo Naïve Bayes

- Sumário:
  - Classificadores bayesianos
    - Introdução
    - Teorema de Bayes
    - Algoritmo Naïve Bayes

Classificadores probabilísticos (1/2)

- Implementam a inferência desde um ponto de vista probabilístico
- Se baseiam no pressuposto de que as variáveis de interesse são governadas por distribuições de probabilidades e que é possível tomar decisões óptimas com base nas referidas probabilidades e nos dados observados

Classificadores probabilísticos (2/2)

- A base teórica é estabelecida pelo *Teorema de Bayes* (Reverendo Thomas Bayes, 1701 – 1761)
- Designados por isso *métodos de aprendizagem bayesianos*
- São relevantes dentro do contexto da AA
  - Se encontram entre as abordagens mais práticas a certos tipos de problemas
  - Desde o ponto de vista teórico, proporcionam uma perspectiva útil para a análise de vários métodos de aprendizagem

Teorema de Bayes (1/7)

- Constitui a pedra angular da *aprendizagem bayesiana*
- Dado um espaço de hipóteses (*H*), estabelece um método para o cálculo da probabilidade associada a uma hipótese ($h \in H$), dado um conjunto de exemplos (*E*) mais qualquer conhecimento inicial sobre as probabilidades associadas a priori a cada hipótese existente em *H*

Teorema de Bayes (2/7)

- Considerando um espaço de hipóteses *H*, um conjunto de dados *E* e uma hipótese $h \in H$,

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

Teorema de Bayes (3/7)

- *P(h)* : denota a probabilidade inicial de que a hipótese *h* seja correcta, antes de observar-mos os dados de treino
  - Reflecte qualquer conhecimento básico que possuímos acerca da possibilidade de que a hipótese *h* seja correcta
  - Denominada *probabilidade a priori*
  - Se não possuímos nenhum conhecimento a priori, se atribui a mesma *probabilidade a priori* a todas as hipóteses candidatas

Teorema de Bayes (4/7)

- *P(E)* : denota a probabilidade de que os dados de treino *E* sejam observados
  - Representa a probabilidade associada à observação dos dados *E*, sem nenhum conhecimento acerca de que hipótese $h \in H$ é a correcta

Teorema de Bayes (5/7)

- *P(E|h)* : denota a probabilidade associada à observação dos dados *E* caso a hipótese *h* seja a correcta
  - Denominada *verossimilhança* do conjunto de dados *E*, dada a hipótese *h*

Teorema de Bayes (6/7)

- *P(h|E)* : denota a probabilidade de que a hipótese *h* seja correcta uma vez observados os dados de treino *E*
  - Denominada *probabilidade a posteriori* de *h*
  - Reflecte a confianza em que *h* é uma hipótese correcta depois de haver sido observados os dados de treino *E*
  - Reflecte a influência dos dados de treino *E*, a diferença da *probabilidade a priori, P(h)*, que é independente de *E*

Teorema de Bayes (7/7)

- Num cenário de aprendizagem, geralmente se considera um conjunto de hipóteses candidatas *H* e se trata de buscar a hipótese mais provável $h \in H$, dados os exemplos de treino *E*
  - É denominada hipótese *máxima a posteriori* (MAP)

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

$$= \arg\max_{h \in H} P(E \mid h)P(h)$$

Teorema de Bayes (exemplo)

- Seja um problema de diagnóstico médico com duas possibilidades:
  - O paciente tem uma forma particular de câncer
  - O paciente não tem câncer
- Os dados existentes são relativos a um teste de laboratório com dois possíveis resultados: Pos (positivo) ou Neg (negativo).
- Existe conhecimento a priori de que no conjunto completo da população, apenas uma proporção de 0,008 possui esta doença

Exemplo Teorema de Bayes (1/3)

- Por outro lado, o teste é um indicador imperfeito da existência da doença. O mesmo retorna um resultado positivo apenas em 98% dos casos em que está presente a doença e um resultado negativo correcto em 97% dos casos em que não está presente a doença.
- Suponhamos que o teste *é positivo* para um paciente dado. Devemos diagnosticar como *tendo câncer* ou *não*?

Exemplo Teorema de Bayes (2/3)

$$P(câncer) = 0{,}008 \qquad P(\neg câncer) = 0{,}992$$

$$P(Pos \mid câncer) = 0{,}98 \qquad P(Neg \mid câncer) = 0{,}02$$

$$P(Pos \mid \neg câncer) = 0{,}03 \qquad P(Neg \mid \neg câncer) = 0{,}97$$

$$P(câncer \mid Pos) = ? \qquad P(\neg câncer \mid Pos) = ?$$

Exemplo Teorema de Bayes (3/3)

$$P(Pos \mid câncer)P(câncer) = 0{,}98 \,.\, 0{,}008 = 0{,}0078$$

$$P(Pos \mid \neg câncer)P(\neg câncer) = 0{,}03 \,.\, 0{,}992 = 0{,}298$$

$$\Rightarrow h_{MAP} = \neg câncer$$

$$P(câncer \mid Pos) = \frac{0{,}0078}{0{,}0078 + 0{,}0298} = 0{,}21$$

$$P(\neg câncer \mid Pos) = \frac{0{,}0298}{0{,}0078 + 0{,}0298} = 0{,}79$$

Teorema de Bayes (considerações)

- A aplicação prática dos métodos *bayesianos* requer um conhecimento inicial de várias probabilidades
  - Geralmente são estimadas a partir do conhecimento de base sobre o problema, dos dados disponíveis e fazendo suposições acerca da forma das distribuições de probabilidades
- No caso geral, a determinação da hipótese óptima segundo Bayes tem um custo computacional significativo

Classificador Naïve Bayes (1/6)

- Método de aprendizagem *bayesiano* amplamente utilizado
- Em muitos domínios apresenta rendimentos ao nível dos métodos de top em matéria de classificação

Classificador Naïve Bayes (2/6)

- Aplica-se a problemas de classificação nos quais se pode assumir que os valores dos atributos são independentes uns dos outros
  - Designação de "naïve" proveniente desta consideração de independência
- Dado um conjunto de treino e uma nova instância, descrita por um vector de valores dos atributos, o algoritmo trata de predizer a classe correspondente à nova instância

Classificador Naïve Bayes (3/6)

- O enfoque *bayesiano* consiste em atribuir à nova instância a classe mais provável, c_{MAP}, dado o vector de atributos que descreve a mesma

$$c_{MAP} = \arg\max_{c_i \in C} P(c_i \mid a_1, a_2, ..., a_n)$$

- Para tal se utiliza o teorema de Bayes

$$c_{MAP} = \arg\max_{c_i \in C} \frac{P(a_1, a_2, ..., a_n \mid c_i) P(c_i)}{P(a_1, a_2, ..., a_n)} = \arg\max_{c_i \in C} P(a_1, a_2, ..., a_n \mid c_i) P(c_i)$$

Classificador Naïve Bayes (4/6)

- Os dois termos na equação são estimados a partir dos dados
- O termo $P(c_i)$ se estima facilmente contando a frequência com que cada classe está presente nos dados de treino

Classificador Naïve Bayes (5/6)

- O termo $P(a_1, a_2, \ldots, a_n \mid c_i)$ se estima assumindo que os valores dos atributos são condicionalmente independentes, dado o valor da função objectivo
- A probabilidade de observar a conjunção de atributos a1, a2… se pode calcular como o produto das probabilidades de observar cada atributo individualmente

$$P(a_1, a_2, \ldots, a_n \mid c_i) = \prod_j P(a_j \mid c_i)$$

Classificador Naïve Bayes (6/6)

- O classificador Naïve Bayes proporciona à saída a classe $c_{NB} = \arg\max_{c_i \in C} P(c_i) \prod_j P(a_j \mid c_i)$
- A aprendizagem consiste em estimar os valores $P(c_i)$ e $P(a_j/c_i)$ a partir das correspondentes frequências no conjunto de dados
- O conjunto de valores estimados corresponde à hipótese aprendida e é utilizado para classificar as novas instâncias

Exemplo classificador Naïve Bayes (1/2)

- Dado o problema do tempo, utilizar Naïve Bayes para classificar a seguinte instância **x** = <Aspecto = sol, Temperatura = fresco, Humidade = alta, Ventoso = sim>

Céu	Temperatura	Humidade	Ventoso	Jogar
sol	quente	alta	falso	não
sol	quente	alta	verd.	não
nublado	quente	alta	falso	sim
chuva	temperado	alta	falso	sim
chuva	fresco	normal	falso	sim
chuva	fresco	normal	verd.	não
nublado	fresco	normal	verd.	sim
sol	temperado	alta	falso	não
sol	fresco	normal	falso	sim
chuva	temperado	normal	falso	sim
sol	temperado	normal	verd.	sim
nublado	temperado	alta	verd.	sim
nublado	quente	normal	falso	sim
chuva	temperado	alta	verd.	não

Exemplo classificador Naïve Bayes (2/2)

$$c_{NB} = \arg\max_{c_i \in C} P(c_i) \prod_j P(a_j \mid c_i)$$

$$= \arg\max_{c_i \in \{sim, não\}} P(c_i)\ P(Céu = sol \mid c_i) P(Temperatura = fresco \mid c_i) P(Humidade = alta \mid c_i) P(Ventoso = sim \mid c_i)$$

Céu			Temperatura			Humidade			Ventoso			Jogar	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3	Quente	2	2	Alta	3	4	Verdadeiro	3	3	9	5
Nublado	4	0	Temperado	4	2	Normal	6	1	Falso	6	2		
Chuva	3	2	Fresco	3	1								
Sol	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Verdadeiro	3/9	3/5	9/14	5/14
Nublado	4/9	0/5	Temperado	4/9	2/5	Normal	6/9	1/5	Falso	6/9	2/5		
Chuva	3/9	2/5	Fresco	3/9	1/5								

$$P(sim)P(sol \mid sim)P(\ fresco \mid sim)P(alta \mid sim)P(verd \mid sim) = \frac{9}{14}\frac{2}{9}\frac{3}{9}\frac{3}{9}\frac{3}{9} = 0{,}0053$$

$$P(não)P(sol \mid não)P(\ fresco \mid não)P(alta \mid não)P(verd \mid não) = \frac{5}{14}\frac{3}{5}\frac{1}{5}\frac{4}{5}\frac{3}{5} = 0{,}0206$$

$$P(nao \mid \mathbf{x}) = \frac{0{,}0206}{0{,}0053 + 0{,}0206} = 0{,}795$$

Variantes (estimação – m)

- Al utilizar o método pode ocorrer que não exista nenhum exemplo de uma classe determinada, c, que tenha para um atributo, a, um valor dado, v,
- O valor 0 para a probabilidade condicional dominaria o cálculo final das probabilidades

Variantes (estimação – m)

- Se evita que as probabilidades condicionais sejam iguais a 0 utilizando-se a chamada estimação – m

$$p(A = v \mid c) = \frac{n_c + m.p}{n + m}$$

  - n_c - # de instâncias que têm o valor v no atributo A e pertencem à classe c
  - *n* - # de instâncias que pertencem à classe c
  - *p* - estimação a priori da probabilidade que se deseja calcular; se não está disponível, assumir uma distribuição uniforme (se o atributo tem *k* valores, *p* = 1/*k*)
  - *m* – parâmetro corrector, denominado *tamanho da mostra equivalente.* Determina o peso a atribuir a *p* relativamente aos dados observados

Variantes (valores contínuos)

- Os atributos numéricos são manejados assumindo que os mesmos apresentam uma distribuição normal de probabilidades
- Para estimar as correspondentes probabilidades, calculam-se as correspondentes médias e desvios padrões e
- Se utiliza a expressão correspondente à função de densidade de probabilidade para uma distribuição normal com média μ e desvio padrão σ

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Variantes (valores contínuos)

Céu	Temperatura	Humidade	Ventoso	Jogar
sol	85	85	falso	não
sol	80	90	verd.	não
nublado	83	86	falso	sim
chuva	70	96	falso	sim
chuva	68	80	falso	sim
chuva	65	70	verd.	não
nublado	54	65	verd.	sim
sol	72	95	falso	não
sol	69	70	falso	sim
chuva	75	80	falso	sim
sol	75	70	verd.	sim
nublado	72	90	verd.	sim
nublado	81	75	falso	sim
chuva	72	91	verd.	não

Variantes (valores contínuos)

Céu			Temperatura			Humidade			Ventoso			Jogar	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3		83	85		86	85	Verdadeiro	3	3	9	5
Nublado	4	0		70	80		96	90	Falso	6	2		
				68	65		80	70					
				64	72		65	95					
				69	71		70	91					
				75			80						
				75			70						
				72			90						
Chuva	3	2		81			75						
Sol	2/9	3/5	Média	73,0	74,6	Média	79,1	86,2	Verdadeiro	3/9	3/5	9/14	5/14
Nublado	4/9	0/5	Desv Pad	6,2	7,9	Desv Pad	10,2	9,7	Falso	6/9	2/5		
Chuva	3/9	2/5											

$$f(Temperatura = 66 \mid sim) = \frac{1}{\sqrt{2\pi}.6,2^2} e^{-\frac{(66-73)^2}{2.6,2^2}} = 0,0340$$

Exercício

- Considere o seguinte conjunto de treino, o qual tem dois atributos com três valores possíveis cada (0, 1 e 2) e duas possíveis classes (x e y).

Instância	A1	A2	Classe
1	0	1	x
2	2	1	x
3	1	1	x
4	0	2	x
5	1	2	y
6	2	0	y

Exercício

- Que classe seria atribuída pelo algoritmo Naïve Bayes para a seguinte instância de teste?

Instância	A1	A2	Classe
7	2	2	?

Bibliografia

- Mitchell, pg. 154 – 158, 177 – 184
- Witten, pg. 90 – 99
- Borrajo Millán