LABIC

SISTEMAS INTELIGENTES

Profa. Roseli Ap. Francelin Romero

LABIC

Fórmulas Básicas para Probabilidades

- Regra Produto: probabilidade **P(A∧B)** de uma conjunção de dois eventos **A** e **B:**

$$P(A \wedge B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)$$

- Regra Soma: probabilidade **P(A∨B)** de uma união de dois eventos **A** e **B**:

$$P(A \vee B) = P(A) + P(B) - P(A \wedge B)$$

- Teorema da probabilidade total: se eventos $A_1, \ldots, A_n$ são mutualmente exclusivos com $\sum_{i=1}^{n} P(A_i) = 1$, então:

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

LABIC

Aprendizado Bayesiano

CLASSIFICADORES BAYESIANO

- **Aprendizado Supervisionado de Classificadores Bayesiano**
- **Aprendizado Não Supervisionado de Classificadores Bayesiano**

LABIC

Classificação de Padrões

- Suponha que você está para testemunhar um evento.
- O evento pertencerá à:
  - classe ω_1 com probabilidade $P(\omega_1)$
  - classe ω_2 com probabilidade $P(\omega_2)$
  - classe ω_n com probabilidade $P(\omega_n)$
- Suponha que você deve prever a classe
- Você paga R$ 1,00 se você estiver errado
- Você não paga nada se estiver certo.

Questões:

- Qual deve ser sua estratégia ótima?
- Qual será o seu custo esperado?

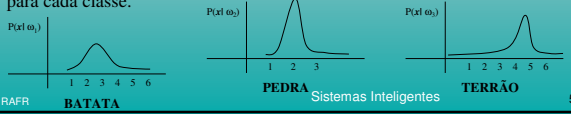
LABIC

Considerando dados observados

Suponha que se deseja construir um SISTEMA AUTOMÁTICO para apanhar ***batatas.*** Toda vez que um objeto toca o sensor debaixo do trator ele deve decidir se pertence à:

ω_1 ***batata*** com probabilidade $P(\omega_1)$
ω_2 ***pedra*** com probabilidade $P(\omega_2)$
ω_3 ***terrão*** com probabilidade $P(\omega_3)$

Suponha também que o sensor computa o diâmetro ***x*** do objeto e que o Instituto de Pesquisa da Batata forneceu as distribuições condicionais de ***x*** para cada classe.



LABIC

DECISÃO

- Conhece-se $P(\omega_1)$, $P(\omega_2)$, $P(\omega_3)$ mais as distribuições $P(x|\ \omega_1)$, $P(x|\ \omega_2)$, $P(x|\ \omega_3)$.
- Observa-se ***x.***
- Qual a classe de objetos escolhida?

I - Máxima Probabilidade

- Escolher a classe ω_i que maximiza $P(x\ |\omega_i)$.
- Fácil de calcular.
- Qual é a objeção? (pode ocorrer erro! Porque se toma a probabilidade partindo-se de uma certa classe).

LABIC

DECISÃO

II - Classificador Bayesiano Ótimo

O que devemos fazer para minimizar a chance de cometermos um erro?

- Escolher a classe ω_i que tem a maior probabilidade dada $\boldsymbol{x}$.

Escolha = $\arg_i \max P(\omega_i \mid \boldsymbol{x})$.

Bayesiano Ótimo = $\boldsymbol{\arg_i \max\ P(x \mid \omega_i).P(\omega_i)}$

Este é o Classificador Ótimo de Bayes.

LABIC

Batatas Multivariado

Suponha que temos 3 sensores $\begin{cases} x_1 \text{ - diâmetro} \\ x_2 \text{ - altura} \\ x_3 \text{ - massa} \end{cases}$

e que temos um vetor $\boldsymbol{x}$ observado

Bayesiano Ótimo = $\arg_i \max P(\underset{\sim}{x} \mid \omega_i) \cdot P(\omega_i)$

Hipótese Comum:

Cada $P(\underset{\sim}{x} \mid \omega_i)$ segue distribuição Gaussiana.

Três Casos:

$P(\underset{\sim}{x} \mid \omega_i)$ - Média μ_i, variância σ^2

$P(\underset{\sim}{x} \mid \omega_i)$ - Média μ_i, covariância $\sum$, arbitrária

$P(\underset{\sim}{x} \mid \omega_i)$ - Média μ_i, covariância $\sum_i$, diferente para classes diferentes

LABIC

Caso 1

Caso 1: Todas componentes são independentes $P(\underset{\sim}{x} \mid \omega_i)$ tem média μ_i. Cada componente de $\boldsymbol{x}$ é independente de outras componentes e tem variância σ^2

$$P(\underset{\sim}{x} \mid \omega_i) = k \exp\left(\frac{-1}{2\sigma^2} \sum_j (x_j - \mu_{ij})^2\right)$$

Bayesiano Ótimo = $\arg_i \max P(\underset{\sim}{x} \mid \omega_i) \cdot P(\omega_i) =$

$= \arg_i \max\{k \exp\left(\frac{-1}{2\sigma^2} \sum_j (x_j - \mu_{ij})^2\right) \cdot P(\omega_i)\} =$

LABIC

Caso 1

$$= \arg_i \max \frac{-1}{2\sigma^2} \sum_j (x_j - \mu_{ij})^2 + \log P(\omega_i) =$$

$$= \arg_i \min \frac{\sum_j (x_j - \mu_{ij})^2 - 2\sigma^2 \log P(\omega_i)}{2\sigma^2} =$$

$$= \arg_i \min \sum_j (x_j - \mu_{ij})^2 - 2\sigma^2 \log P(\omega_i)$$

LABIC

Caso 1

- Caso duas classes

$= \arg_i \min\ (\sum (\underset{\sim}{x} - \underset{\sim}{\mu_i})^2 - 2\sigma^2 \log P(\omega_i)) =$

$= \arg_i \min\ (\sum \underset{\sim}{x}\,\underset{\sim}{x} - 2\,\underset{\sim}{x}\,\underset{\sim}{\mu_i} + \underset{\sim}{\mu_i}\,\underset{\sim}{\mu_i} - 2\sigma^2 \log P(\omega_i)) =$

$= \arg_i \min\ (\sum - 2\,\underset{\sim}{x}\,\underset{\sim}{\mu_i} + c_i)$

Se $-2\,\underset{\sim}{x}\,\underset{\sim}{\mu_1} + c_1 < -2\,\underset{\sim}{x}\,\underset{\sim}{\mu_2} + c_2 \rightarrow$ **Escolha** $\omega_1 \Leftrightarrow$

LABIC

Caso 1

$\Leftrightarrow$ Se $c_1 - c_2 < 2(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})\,\underset{\sim}{x} \rightarrow$ **Escolha** ω_1

$\Leftrightarrow$ **A regra de decisão é:**

" Se $\underset{\sim}{\omega}\,\underset{\sim}{x} >$ *threshold* **"** onde $\underset{\sim}{\omega} = 2(\underset{\sim}{\mu_1} - \underset{\sim}{\mu_2})$

e *threshold* $= c_1 - c_2$

Portanto a decisão ótima é de um CLASSIFICADOR LINEAR! Perceptrons são corretos!

OBS.: A regra do Perceptron pode ser obtida do classificador ótimo de Bayes.

Caso 2 - Hipótese mais fraca

Agora, $P(\underset{\sim}{x}|\omega_i)$ gaussiana, média $\underset{\sim}{\mu_i}$ e covariância arbitrária Σ . Temos que a mesma regra ocorre, mas numa medida de distancia diferente:

$$Dist(\underset{\sim}{x}, \underset{\sim}{\mu_i}) = (\underset{\sim}{x} - \underset{\sim}{\mu_i})^T \Sigma^{-1} (\underset{\sim}{x} - \underset{\sim}{\mu_i})$$

Se todos os $P(\omega_i)_S$ são iguais $\Rightarrow$ método do vizinho mais próximo (KNN)

Ainda usa regiões de decisão linear.

Caso 3 - Hipótese ainda mais fraca

$P(\underset{\sim}{x}|\omega_i)$ gaussiana, média $\underset{\sim}{\mu_i}$ e covariância $\Sigma_i \rightarrow$ para diferentes classes a variância pode ser diferente.

Ainda é fácil calcular a decisão ótima

$$arg_i max(P(\omega_i | \underset{\sim}{x}))$$

mas as regiões de decisão não são mais lineares.

Classificacao de Padroes

- Suponha agora que voce não conhece $P(w_1)\ P(w_2) \ldots P(w_N)$, $\underset{\sim}{\mu_1}, \underset{\sim}{\mu_2} \ \ldots \ \underset{\sim}{\mu_N}$

Mas, voce deseja estimar estes parametros dos dados.

$x_1^{(1)}\ x_2^{(1)} \ldots\ x^{(1)}_N$ **Classe w_1**

$x_1^{(2)}\ x_2^{(2)} \ldots\ x^{(2)}_N$ **Classe w_2**

..

$x_1^{(M)}\ x_2^{(M)} \ldots\ x^{(M)}_N$ **Classe w_N**

Classificacao de Padroes

- Estimar $P(w_i) = \frac{\text{numero de dados da classe } w_i}{\text{numero total de dados}}$
- Estima a media μ_i = media de todos os pontos da classe w_i

$$arg_i max\ P(\underset{\sim}{x}|\omega_i).P(\omega_i)$$

Métodos de Aprendizado Bayesiano

- Calculam explicitamente probabilidades para hipóteses (Naïve Bayes Classificador).

Mitchie et al. (1994) comparou o classificador Naïve Bayes com RN e DT.

- Eles fornecem uma perspectiva útil para compreensão dos algoritmos de aprendizado que não explicitamente manipulam probabilidades.

Características dos Métodos de Aprendizado Bayesiano

- Cada exemplo observado pode incrementalmente diminuir ou aumentar a probabilidade estimada que uma hipótese está correta.
- Conhecimento "priori" pode ser combinado com o dado observado para determinar a probabilidade final de uma hipótese. Em Aprendizado Bayesiano, conhecimento a prior, pode ser fornecido:
  - Dando uma probabilidade "a priori" para cada hipótese candidata.
  - Distribuição de probabilidade sobre os dados para cada hipótese possível.

Características dos Métodos de Aprendizado Bayesiano

- Métodos Bayesiano podem acomodar hipóteses que contém previsões probabilísticas, tais como:

"este paciente, com pneumonia, tem 93% de chance de cura".

- Novas instâncias podem ser classificadas combinando as previsões de múltiplas hipóteses, ponderadas por ***"suas probabilidades".***

- Em métodos computacionais igualmente intratáveis, eles podem fornecer um padrão de tomada de decisão ótima.

Características dos Métodos de Aprendizado Bayesiano

- Dificuldade 1:

  Requerem o conhecimento de muitas probabilidades. Quando estas probabilidades não são conhecidas "a priori" elas são estimadas baseadas no: **conhecimento do problema, dados previamente disponíveis e hipóteses sobre a forma da distribuição fundamental dos dados.**

- Dificuldade 2:

  Custo computacional requerido pode ser reduzido significantemente.

TEOREMA DE BAYES

Em problemas de ML estamos interessados em **P(h|D):** probabilidade a posteriori, probabilidade que vale **h** dado o conjunto de treinamento observado **D.**

Teorema de Bayes: $P(h|D) = \frac{P(D|h)\,P(h)}{P(D)}$

Em muitos casos o aprendiz considera algum conjunto de hipóteses candidatas **H** e está interessado em encontrar a hipótese mais provável $h \in H$ dado o conjunto de dados observado **D** (ou no mínimo uma hipótese mais provável, se existirem várias).

TEOREMA DE BAYES

Tal hipótese é chamada uma Maximum A Posteriori (MAP) hipótese.

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

É independente de h

Em alguns casos, assumiremos que toda hipótese em **H** é igualmente provável, isto é:

$P(h_i) = P(h_j)$ para todos h_i e h_j em **H** então a equação anterior fica:

TEOREMA DE BAYES

$$h_{ML} = \arg\max_{h \in H} P(D|h)$$

Maximum likelihood (Probabilidade Maxima)

No enfoque de ML

D - exemplos de treinamento de alguma função alvo.

H - o espaço das funções alvo candidatas.

EXEMPLO

Paciente tem câncer ou não?

Um paciente faz um teste de laboratório e o resultado volta positivo. O teste devolve um resultado positivo correto em só 98% dos casos nos quais a doença está realmente presente, e um resultado negativo correto em 97% dos casos nos quais a doença não está presente. Além disso, 0.008 da população inteira tem este câncer.

P(câncer) = 0.008 **P(¬câncer) = 0.992**

P(+|câncer) = 0.98 **P(-|câncer) = 0.02**

P(+| ¬ câncer) = 0.03 **P(-| ¬ câncer) = 0.97**

P(+|câncer) . P(câncer) = (0.98) . (0.008) = 0.0078

P(+| ¬ câncer) . P(¬ câncer) = (0.03) . (0.992) = 0.0298

h_{MAP} **= ¬ câncer**

Classificação mais Provável de Novas Instâncias

Até agora nós buscamos a hipótese mais provável dado o conjunto **D** (i.e. $\mathbf{h_{MAP}}$)

Dado nova instância **x**, qual é a sua classificação mais provável?

$\mathbf{h_{MAP}(x)}$ não é a classificação mais provável.

Considere por exemplo:

- três hipóteses: $\mathbf{P(h_1|D)=0.4,\ P(h_2|D)=0.3,\ P(h_3|D)=0.3}$
- Dado a nova instância **x** : $\mathbf{h_1(x)= +\ ,\ h_2(x)= -\ ,\ h_3(x)= -}$
- Qual é a mais provável classificação de **x** ?

$\mathbf{p_+(x)= 0.4\ ,\quad p_-(x)= 0.6}$, portanto é mais provável que **x** seja -

Neste caso, é diferente da classificação gerada pela $\mathbf{h_{MAP}}$

Classificador Bayesiano Ótimo

$$\mathbf{arg_{v_j \in V} max \sum_{h_i \in H} P(v_j \mid h_i)\ .\ P(h_i \mid D)}$$

EXEMPLO:

$\mathbf{P(h_1 \mid D) = 0.4,\quad P(- \mid h_1) = 0,\quad P(+ \mid h_1) = 1,}$

$\mathbf{P(h_2 \mid D) = 0.3,\quad P(- \mid h_2) = 1,\quad P(+ \mid h_2) = 0,}$

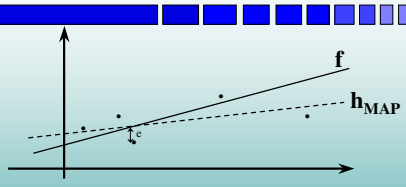
$\mathbf{P(h_3 \mid D) = 0.3,\quad P(- \mid h_3) = 1,\quad P(+ \mid h_3) = 0,}$

Portanto, $\mathbf{\sum_{h_i \in H} P(+ \mid h_i)\ .\ P(h_i \mid D) = 0.4}$

$\mathbf{\sum_{h_i \in H} P(- \mid h_i)\ .\ P(h_i \mid D) = 0.6}$

Portanto, $\mathbf{arg_{v_j \in V} max \sum_{h_i \in H} P(v_j \mid h_i)\ .\ P(h_i \mid D) = -}$

Aprendizado de uma Função Real



Considere exemplos de treinamento $\mathbf{<x_i,d_i>}$, onde $\mathbf{d_i}$ é o ruído dado por:

$$\mathbf{d_i = f(x_i) + e_i}$$

onde $\mathbf{e_i}$ é uma variável aleatória, independente para

Aprendizado de uma Função Real

cada x_i de acordo com alguma distribuição Gaussiana com média = 0. Então,

$$\mathbf{h_{ML} = arg_{h \in H} min \sum_{i=1}^{m} (d_i - h(x_i))^2}$$

Demonstração:

$$\mathbf{h_{ML} = arg_{h \in H} max\ p(D|h) = arg_{h \in H} max \prod_{i=1} p(d_i|h) =}$$

$$\mathbf{= arg_{h \in H} max \prod_{i=1}^{m} \frac{1}{\sqrt{2\sigma^2}}\ e^{-\frac{1}{2}((d_i - h(x_i))/\sigma)^2} =}$$

Maximizando o logaritmo natural:

$$\mathbf{h_{ML} = arg_{h \in H} max \sum_{i=1}^{m} -\tfrac{1}{2}((d_i - h(x_i))/\sigma)^2 =}$$

Aprendizado de uma Função Real

$$\mathbf{= arg_{h \in H} max \sum_{i=1}^{m} -(d_i - h(x_i))^2 =}$$

$$\mathbf{= arg_{h \in H} min \sum_{i=1}^{m} (d_i - h(x_i))^2}$$

Classificador Bayesiano Naive

Está entre um dos melhores classificadores (árvores de decisão, NN, KNN)

Quando usar:

- Conjunto de treinamento grande.
- Atributos são condicionalmente independentes.

Aplicações bem sucedidas:

- Diagnósticos
- Classificação de textos em documentos

LABIC

Classificador Bayesiano Naive

Seja: $\mathbf{f: X \rightarrow V}$

$\mathbf{x = < a_1, a_2, \dots , a_n >}$

Qual é o mais provável valor de **f(x)** ?

$$\mathbf{v_{MAP} = \arg_{v_j \in V} \max P(v_j | a_1, a_2, \dots , a_n)}$$

$$\mathbf{v_{MAP} = \arg_{v_j \in V} \max \frac{P(a_1, a_2, \dots , a_n | v_j) P(v_j)}{P(a_1, a_2, \dots , a_n)}}$$

$$\mathbf{v_{MAP} = \arg_{v_j \in V} \max P(a_1, a_2, \dots , a_n | v_j) P(v_j)}$$

Hipótese Naïve Bayes: $\mathbf{P(a_1, a_2, \dots , a_n | v_j) = \pi_i P(a_i | v_j)}$

LABIC

Classificador Bayesiano Naive

Classificador Bayesiano Naïve:

$$\mathbf{V_{NB} = \arg_{v_j \in V} \max P(v_j) \pi_i P(a_i | v_j)}$$

EXEMPLO:

Considere o exemplo "Play Tennis" e a instância:

<Outlook = sunny,Temp=cool,Hum=high,wind=strong>

Queremos:

$$\mathbf{V_{NB} = \arg_{v_j \in V} \max P(v_j) \pi_i P(a_i | v_j) =}$$

LABIC

Algoritmo Naive Bayes

- Naive_Bayes_Learn(examples)
  - Para cada valor alvo v_j
  - Calcule $P'(v_j)$ ← estimativa $P(v_j)$
  - Para cada valor de atributo a_i de cada atrib. A
    - $P'(a_i | v_j)$ ← estimativa $P(a_i | v_j)$

Classificar nova instancia x

$$\mathbf{V_{NB} = \arg_{v_j \in V} \max P'(v_j) \pi P'(a_i | v_j)}$$

LABIC

Exemplos de Treinamento

DAY	OUTLOOK	TEMPERATURA	UMIDADE	VENTO	PLAYTENN
D1	SOL	QUENTE	ALTA	FRACO	NÃO
D2	SOL	QUENTE	ALTA	FORTE	NÃO
D3	NUBLADO	QUENTE	ALTA	FRACO	SIM
D4	CHUVA	AMENO	ALTA	FRACO	SIM
D5	CHUVA	FRIO	NORMAL	FRACO	SIM
D6	CHUVA	FRIO	NORMAL	FORTE	NÃO
D7	NUBLADO	FRIO	NORMAL	FORTE	SIM
D8	SOL	AMENO	ALTA	FRACO	NÃO
D9	SOL	FRIO	NORMAL	FRACO	SIM
D10	CHUVA	AMENO	NORMAL	FRACO	SIM
D11	SOL	AMENO	NORMAL	FORTE	SIM
D12	NUBLADO	AMENO	ALTA	FORTE	SIM
D13	NUBLADO	QUENTE	NORMAL	FRACO	SIM
D14	CHUVA	AMENO	ALTA	FORTE	NÃO

LABIC

EXEMPLO

Considere o exemplo "Play Tennis" e a instância:

<Outlook = sunny,Temp=cool,Hum=high,wind=strong>

Queremos:

$$\mathbf{V_{NB} = \arg_{v_j \in V} \max P(v_j) \pi P(a_i | v_j) =}$$

LABIC

Classificador Bayesiano Naive

⇒P(yes) P(sunny|yes) P(cool|yes) P(high|yes) P(strong|yes)=
9/14 * 2/9 * 3/9 */ 3/9 * 3/9 =
= 0.0053
⇒P(no) P(sunny|no) P(cool|no) P(high|no) P(strong|no)=
= 0.0206

→ V_{NB} = n

OBS: Cap.6 - T. Mitchell para ver aplicação de busca de texto em documentos da Web.

LABIC

Curiosidades

- A hipótese de independência condicional é frequentemente violada

  Mas, mesmo assim o Classificador Naive Bayes funciona muito bem.

- Não se precisa de estimativas corretas da prob. A posteriori $\mathbf{P(v_j \mid x)}$ mas apenas de:

  $\mathbf{P'(v_j)}$ **e** $\mathbf{P'(a_i \mid v_j)}$

LABIC

Curiosidades

- O que fazer se nenhum dos exemplos no conj. De dados com valor alvo $\mathbf{v_j}$ **tem valor de atributo** $\mathbf{a_i}$ **?**

  $\mathbf{P'(a_i \mid v_j) = 0}$ **e**

  $\mathbf{P'(v_j) \pi P'(a_i \mid v_j) = 0}$

  Solução é estimar $\mathbf{P'(a_i \mid v_j)}$

  $$\mathbf{P'(a_i \mid v_j) = (n_c + m) / (n+m.p)}$$

Onde: n – no. exs. de treinamento para o qual $\mathbf{v = v_j}$

$\mathbf{n_c}$ **- no. de exs. Virtuais para os quais** $\mathbf{v = v_j}$ **e** $\mathbf{a = a_i}$

p é uma estimativa da priori para $\mathbf{P'}$(ai | vj) e m é um peso dado a priori