APRENDIZADO BAYESIANO

SCC0276 – APRENDIZADO DE MÁQUINA

PROFA. Roseli Ap. Francelin Romero

Fórmulas Básicas para Probabilidades

- Regra Produto: probabilidade $\mathbf{P(A \wedge B)}$ de uma conjunção de dois eventos **A** e **B:**

$$\mathbf{P(A \wedge B) = P(A \mid B)\, P(B) = P(B \mid A)\, P(A)}$$

- Regra Soma: probabilidade $\mathbf{P(A \vee B)}$ de uma união de dois eventos **A** e **B**:

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

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

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

Aprendizado Bayesiano

CLASSIFICADORES BAYESIANO

Aprendizado Supervisionado de Classificadores Bayesiano

Aprendizado Não Supervisionado de Classificadores Bayesiano

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?

Considerando dados observados

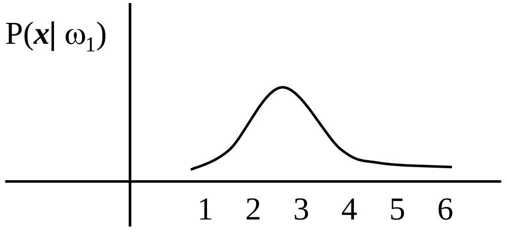
Suponha que se deseja construir um SISTEMA AUTOMÁTICO para apanhar ***batatas.*** Toda vez que um objeto toca o sensor debaixo do trator ele deve decidir se pertence à:

ω_1 ***batata*** com probabilidade $P(\omega_1)$

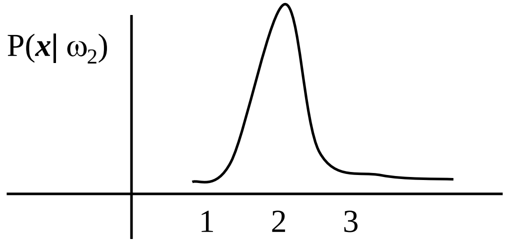
ω_2 ***pedra*** com probabilidade $P(\omega_2)$

ω_3 ***terrão*** com probabilidade $P(\omega_3)$

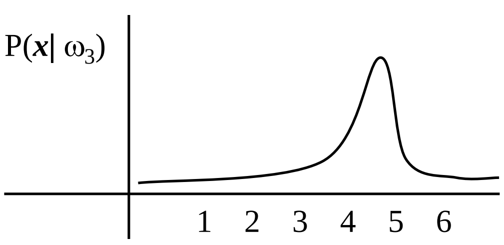
Suponha também que o sensor computa o diâmetro ***x*** do objeto e que o Instituto de Pesquisa da Batata forneceu as distribuições condicionais de ***x*** para cada classe.



BATATA



PEDRA



TERRÃO

DECISÃO

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

I - Máxima Probabilidade

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

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 ***x***.

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

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

Este é o Classificador Ótimo de Bayes.

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

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

Hipótese Comum:

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

Três Casos:

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

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

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

Caso 1: Todas componentes são independentes

$P(\underset{\sim}{\boldsymbol{x}}|\omega_i)$ tem média $\underset{\sim}{\mu}_i$. Cada componente de $\boldsymbol{x}$ é independente de outras componentes e tem variância σ^2

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

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

$$= \arg_i \max\left\{k \exp\left(\frac{-1}{2\sigma^2} \sum_j (\boldsymbol{x_j} - \mu_{i\,j})^2\right) . P(\omega_i) \right\} =$$

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

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

$$= \arg_i \min \sum_j (\boldsymbol{x_j} - \mu_{i\,j})^2 - 2\sigma^2 \log P(\omega_i)$$

- Caso duas classes

$= \arg_i \min \left((\underset{\sim}{x} - \underset{\sim}{\mu}_i)^2 - 2\sigma^2 \log P(\omega_i) \right) =$

$= \arg_i \min \left(\underset{\sim}{x}\, \underset{\sim}{x} - 2\, \underset{\sim}{x}\, \underset{\sim}{\mu}_i + \underset{\sim}{\mu}_i\, \underset{\sim}{\mu}_i - 2\sigma^2 \log P(\omega_i) \right) =$

$= \arg_i \min \left(- 2\, \underset{\sim}{x}\, \underset{\sim}{\mu}_i + c_i \right)$

Se $- 2\, \underset{\sim}{x}\, \underset{\sim}{\mu}_1 + c_1 < - 2\, \underset{\sim}{x}\, \underset{\sim}{\mu}_2 + c_2 \rightarrow$ **Escolha** $\omega_1 \Leftrightarrow$

⇔ Se $\mathbf{c_1 - c_2} < \boldsymbol{2\,(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)\,\underset{\sim}{x}} \rightarrow$ **Escolha** $\boldsymbol{\omega_1}$

⇔ **A regra de decisão é:**

“ **Se** $\boldsymbol{\underset{\sim}{\omega}\,\underset{\sim}{x} > \mathit{threshold}}$ “ onde $\boldsymbol{\underset{\sim}{\omega} = 2\,(\underset{\sim}{\mu}_1 - \underset{\sim}{\mu}_2)}$

e $\boldsymbol{\mathit{threshold} = c_1 - c_2}$

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.

Hipótese mais fraca

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

$$\boxed{\text{Dist}\,(\underset{\sim}{\boldsymbol{x}}, \underset{\sim}{\boldsymbol{\mu}}_{\mathbf{i}}) = (\underset{\sim}{\boldsymbol{x}} - \underset{\sim}{\boldsymbol{\mu}}_{\mathbf{i}})^T \Sigma^{-1} (\underset{\sim}{\boldsymbol{x}} - \underset{\sim}{\boldsymbol{\mu}}_{\mathbf{i}})}$$

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

Ainda usa regiões de decisão linear.

Hipótese ainda mais fraca

$P(\underset{\sim}{\boldsymbol{x}} \mid \omega_i)$ gaussiana, média $\underset{\sim}{\boldsymbol{\mu}}_{\mathbf{i}}$ e covariância Σ_i → para diferentes classes a variância pode ser diferente.

Ainda é fácil calcular a decisão ótima

$$\boxed{\arg_i \max(P(\omega_i \mid \underset{\sim}{\boldsymbol{x}}))}$$

mas as regiões de decisão não são mais lineares.

Classificacao de Padroes

- Suponha agora que voce nao conhece

  $P(w_1)$ $P(w_2)$ … $P(w_N)$, $\underset{\sim}{\boldsymbol{\mu}}_{\mathbf{1}}$, $\underset{\sim}{\boldsymbol{\mu}}_{\mathbf{2}}$ **...** $\underset{\sim}{\boldsymbol{\mu}}_{\mathbf{N}}$

Mas, voce deseja estimar estes parametros dos dados.

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

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

..

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

Classificacao de Padroes

- Estimar $P(w_i) = \dfrac{\text{numero de dados da classe } w_i}{\text{numero total de dados}}$

- Estima a media $\boldsymbol{\mu}_\mathbf{i}$ **=** media de todos os pontos da classe w_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 vale **h** dado o conjunto de treinamento observado **D.**

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

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 a hipótese mais provável, se existirem várias).

TEOREMA DE BAYES

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

$$\mathbf{h_{MAP} = arg_{h \in H} max\ P(h|D) =}$$

$$\mathbf{= arg_{h \in H} max \frac{P(D|h)\ P(h)}{P(D)} =}$$

É independente de h

$$\mathbf{= arg_{h \in H} max\ P(D|h)\ P(h)}$$

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

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

TEOREMA DE BAYES

$$\mathbf{h_{ML} = arg_{h \in H} max\ P(D|h)}$$

Maximum likelihood (Probabilidade Maxima)

No enfoque de ML

D - exemplos de treinamento de alguma função alvo.

H - como 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.

$P(\text{câncer}) = 0.008$ **$P(\neg\text{câncer}) = 0.992$**

$P(+|\text{câncer}) = 0.98$ **$P(-|\text{câncer}) = 0.02$**

$P(+|\neg\text{câncer}) = 0.03$ **$P(-|\neg\text{câncer}) = 0.97$**

$P(+|\text{câncer}) \cdot P(\text{câncer}) = (0.98) \cdot (0.008) = 0.0078$

$P(+|\neg\text{câncer}) \cdot P(\neg\text{câncer}) = (0.03) \cdot (0.992) = 0.0298$

$h_{MAP} = \neg\text{câncer}$

Classificação mais Provável de Novas Instâncias

Até agora nós buscamos a mais provável hipótese 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** ?

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

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

Classificador Bayesiano Ótimo

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

EXEMPLO:

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

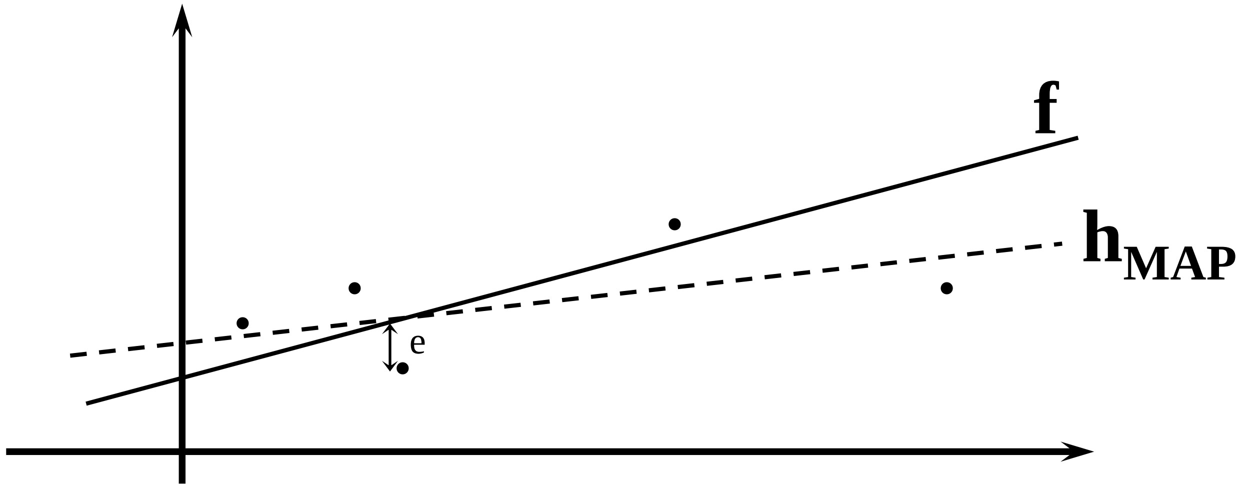
$P(h_2 \mid D) = 0.3,\quad P(- \mid h_2) = 1,\quad P(+ \mid h_2) = 0,$

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

Portanto,
$$\sum_{h_i \in H} P(+ \mid h_i) \, . \, P(h_i \mid D) = 0.4$$
$$\sum_{h_i \in H} P(- \mid h_i) \, . \, P(h_i \mid D) = 0.6$$

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

Aprendizado de uma Função Real



Considere exemplos de treinamento $\langle \mathbf{x_i}, \mathbf{d_i} \rangle$, 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,

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

Demonstração:

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

$$\mathbf{= arg_{h \in H} max \prod_{i=1}^{m} \frac{1}{\sqrt{2\sigma^2}}\ e^{-½((d_i - h(x_i))|\sigma)2} =}$$

Maximizando o logaritmo natural:

$$\mathbf{h_{ML} = arg_{h \in H} max \sum_{i=1}^{m} -½((d_i - h(x_i))/\sigma)^2 =}$$

Aprendizado de uma Função Real

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

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

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

Classificador Bayesiano Naive

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

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

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

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

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

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

Hipotese de Naïve Bayes: $\mathbf{P(a_1, a_2, \ldots , a_n \mid v_j) = \pi_i\ P(a_i \mid v_j)}$

Classificador Bayesiano Naive

Classificador Bayesiano Naïve:

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

EXEMPLO:

Considere o exemplo “Play Tennis” e a instância:

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

Queremos:

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

Classificador Bayesiano Naive

$\Rightarrow$P(yes) P(sunny|yes) P(cool|yes) P(high|yes) P(strong|yes)=

= 0.0053

$\Rightarrow$P(no) P(sunny|no) P(cool|no) P(high|no) P(strong|no)=

= 0.0206

$\rightarrow$ $\mathbf{V_{NB} = n}$

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