Curso 2 – CD, AM e DM

CLASSIFICADOR BAYESIANO
NAIVE-BAYES CLASSIFIER

PROFA. ROSELI AP. FRANCELIN ROMERO

SCC – ICMC - USP

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

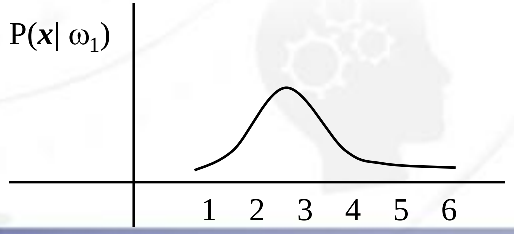
Suponha que se deseja construir um SISTEMA AUTOMÁTICO para apanhar ***batatas.*** Toda vez que um objeto toca o sensor debaixo do trator ele deve decidir se pertence à:

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

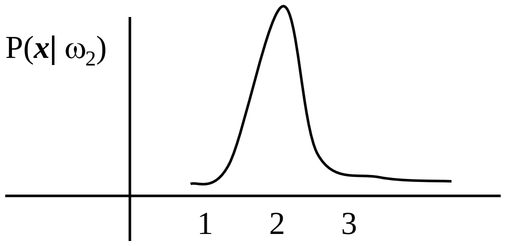
ω_2 ***pedra*** com probabilidade $P(\omega_2)$

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

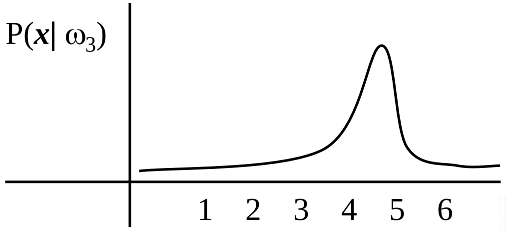
Suponha também que o sensor computa o diâmetro ***x*** do objeto e que o Instituto de Pesquisa da Batata forneceu as distribuições condicionais de ***x*** para cada classe.



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 $\boldsymbol{x}$.

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

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

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 ***x*** observado

Bayesiano Ótimo = $\arg_i \max P(\tilde{\mathbf{x}} / \omega_i) \, . \, P(\omega_i)$

Hipótese Comum:

Cada $P(\mathbf{x} / \omega_i)$ segue distribuição Gaussiana.

Três Casos:

$P(\tilde{\mathbf{x}} / \omega_i)$ - Média μ_i , variância σ^2

$P(\tilde{\mathbf{x}} / \omega_i)$ - Média μ_i , covariância Σ, arbitrária

$P(\tilde{\mathbf{x}} / \omega_i)$ - Média μ_i , covariância Σ_i , diferente para classes diferentes

Função Gaussiana

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

$$P(\tilde{\boldsymbol{x}} / \omega_i) = k \exp\left(-\frac{1}{2\sigma^2} \sum (\boldsymbol{x}_j - \mu_{i\,j})^2\right)$$

$$\text{Bayesiano Ótimo} = \arg_i \max P(\tilde{\boldsymbol{x}} / \omega_i) \,.\, P(\omega_i) =$$

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

Classificação de Padrões

- Suponha agora que voce nao conhece

  $P(w_1)$ $P(w_2)$ … $P(w_N)$, μ_1, μ_2 ... μ_N

Mas, voce deseja estimar estes parametros dos dados.

$\tilde{x}_1^{(1)}$ $\tilde{x}_2^{(1)}$ ... $\tilde{x}^{(1)}_N$ **Classe w_1**

$\tilde{x}_1^{(2)}$ $\tilde{x}_2^{(2)}$ ... $\tilde{x}^{(2)}_N$ **Classe w_2**

..

$\tilde{x}_1^{(M)}$ $\tilde{x}_2^{(M)}$ ... $\tilde{x}^{(M)}_N$ **Classe 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.

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, mas pode ser reduzido significantemente.

TEOREMA DE BAYES

Em problemas de AM estamos interessados em **P(h|D):** probabilidade posteriori, probabilidade 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** ∊ **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.

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

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

$\longrightarrow$ **É independente de h**

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

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

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

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.

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

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

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

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

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

$h_{MAP} = \neg\text{câncer}$

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)** ?

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

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

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

Hipótese 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:

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

EXEMPLO:

Considere o exemplo “Play Tennis” e a instância:

<Outlook = sol,Temp=fria,Um=alta,vento=forte>

Queremos:

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

Classificador Bayesiano Naive

$\Rightarrow$P(sim) P(sol|sim) P(frio|sim) P(alta|sim) P(forte|sim)=

= 0.0053

$\Rightarrow$P(não) P(sol|não) P(frio|não) P(alta|não) P(forte|não)=

= 0.0206

$\rightarrow V_{NB} = n$

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

REFERÊNCIAS

- MITCHELL, Tom M. Machine Learning: McGraw-Hill, 1997.
- WEKA, Weka 3: Data Mining Software in Java, Disponível em www.cs.waikato.ac.nz/ml/weka/, Acesso em: mar. 2010.