Aprendizado de Máquina

Métodos probabilísticos

Prof. Dr. André C. P. L. F. de Carvalho
Posdoutorando: Isvani Frias-Blanco
ICMC-USP

Principais tópicos

- Métodos baseados em probabilidade
- Métodos discriminativos
  - Regressão Logística
- Métodos generativos
  - Teoria das probabilidades
  - Teorema de Bayes
  - *Naive Bayes*

Introdução

- Muitos problemas de classificação são não determinísticos
  - Relação entre atributos de entrada e classe é probabilística
    - Ruído nos dados
    - Algumas informações importantes não são capturadas pelos atributos preditivos usados
    - Informações capturadas pelos atributos preditivos usados são incompletas ou imprecisas

Exemplo

- Predizer se uma pessoa terá problemas cardíacos
  - Atributos preditivos: peso e frequência de exercício
  - Ignora outras possíveis causas:
    - Bebida
    - Hereditariedade
    - Fumo
    - Stress
    - ...

Métodos probabilísticos

- Em várias aplicações é importante estimar a probabilidade de um exemplo pertencer a uma classe
- Modelam relacionamento probabilístico entre atributos preditivos e atributo alvo
- Tipos de modelos induzidos:
  - Modelos discriminativos
  - Modelos generativos

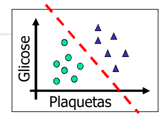
Métodos discriminativos

- Modelam a distribuição de probabilidade a posteriori (condicional) P(Y/X)
- Dado X, retornam a distribuição de probabilidade para Y
  - Ex.: Regressão logística

Métodos generativos

- Modelam a distribuição de probabilidade conjunta P(X,Y)
  - Com a distribuição conjunta é possível derivar qualquer distribuição condicional
- Induzidos por algoritmos baseados no teorema de Bayes
  - Métodos Bayesianos
  - Ex.: *Naive Bayes*

Discriminante linear



- Induz função linear
  - Função discriminante
  - Ajusta parâmetros da função discriminante

$$f(x) = w_0 + w_1x_1 + w_2x_2 + \ldots$$

  - Valor de *f(x)*
    - Distância de *x* à fronteira
      - Chance de *x* pertencer à classe + (-)
  - Semelhante à rede Perceptron

Discriminante linear

- Distância de exemplos a fronteira de decisão definida por uma função linear
- Problema:
  - Distância: $-\infty < f(x) < +\infty$
  - Modelos probabilísticos:
    - Probabilidade: $0 < f(x) < 1$
- Solução:
  - Regressão logística

Regressão logística

- Apesar do nome, é usada para tarefas de classificação
- Estima probabilidade que um exemplo pertence a uma dada classe
  - Ajusta uma função logística a um conjunto de dados
    - Gera um hiperplano de separação
    - Utiliza um conjunto de treinamento

Regressão logística

Probabilidade (P_+)	Chance ($P_+/(1-P_+)$)	Log(Chance)
0,5	50:50 = 1	0,00
0,9	90:10 (9:1) = 9	2,19
0,999	999:1 = 999	6,91
0,01	1:99 = 0,0101	-4,60
0,001	1:999 = 0,001001	-6,91

- Encontrar *f(x)* que modela *log(Chance)*
  - Permite estimar probabilidade usando modelo gerado por discriminante linear

Regressão logística

- Probabilidade de exemplo pertencer a classe positiva
  - Evento ocorreu

Função logit

$$\log\left(\frac{p_+(x)}{1-p_+(x)}\right) = f(x) = w_0 + w_1x_1 + w_2x_2 + \ldots$$

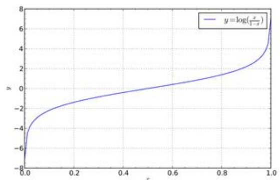
$$p_+(x) = \frac{1}{1+e^{-f(x)}}$$

$$g(x,w) = \begin{cases} p_+(\mathrm{x}) & \text{se x é} + \\ 1-p_+(\mathrm{x}) & \text{se x é} - \end{cases}$$

Função objetivo para ajuste dos pesos

Regressão logística

- Função logit
  - Inversa da função sigmoidal



$$\log\left(\frac{p_+(x)}{1-p_+(x)}\right)$$

Treinamento

- Encontrar valores de w_j que minimizem erro no conjunto de treinamento
  - Aproximação numérica da máxima verossimilhança
  - Gradiente descendente estocástica
    - Para grandes conjuntos de dados
  - Exemplo para 1 atributo preditivo
    - w_0 : posição da função sigmoidal
    - w_1 : inclinação da função sigmoidal

Teoria das probabilidades

- Espaço amostral (Ω) : todos as possíveis observações de um experimento
- Evento (A): subconjunto de possíveis observações em Ω
- Ex.: Jogar um dado 8 vezes
  - Ω = {1, 3, 3, 4, 2, 5, 1, 6}
  - A = valor do dado < 4 = {1, 3, 3, 2, 1}
  - P(A): probabilidade de um evento ocorrer

Teoria das probabilidades

- P(A) satisfaz axiomas de Kolmogorov
  - $P(A) \geq 0$
  - $P(\Omega) = 1$
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
    - Se A e B são eventos mutuamente exclusivos
      - $(A \cap B) = \varnothing$
      - $P(A \cap B) = 0$
      - $P(A \cup B) = P(A) + P(B)$

Probabilidade conjunta

- Probabilidade conjunta
  - Probabilidade de dois eventos ocorrerem simultaneamente
  - $P(A \cap B)$ ou P(A,B)
  - Se são eventos independentes
    - A ocorrência de um não afeta a probabilidade de ocorrência do outro
      - $P(A \cap B) = P(A) * P(B)$

Probabilidade e AM

- Sejam dois eventos A e B
  - A: atributo alvo (presença de uma doença)
    - Variável aleatória com dois valores: presença e ausência
  - B: atributo de entrada (resultado de um exame)
    - Variável aleatória com dois valores: positivo e negativo
  - P(A): probabilidade do evento A ocorrer (presença da doença)
    - $P(A) = 1 - P(\neg A)$
  - P(B): probabilidade do evento B ocorrer (exame positivo)
    - $P(B) = 1 - P(\neg B)$

Probabilidade e AM

Paciente	Exame	Doença
001	positivo	presente
002	negativo	presente
003	negativo	ausente
004	positivo	presente
005	positivo	ausente
006	positivo	presente
007	negativo	ausente
008	negativo	presente
009	positivo	ausente
010	positivo	presente

Probabilidade *a priori* pode ser estimada pela frequência

P(negativo) =
P(positivo) =
P(presente) =
P(ausente) =

O que se deseja em AM é a probabilidade *a posteriori*

Probabilidade e AM

Paciente	Exame	Doença
001	positivo	presente
002	negativo	presente
003	negativo	ausente
004	positivo	presente
005	positivo	ausente
006	positivo	presente
007	negativo	ausente
008	negativo	presente
009	positivo	ausente
010	positivo	presente

Probabilidade *a priori* pode ser estimada pela frequência

P(negativo) = 0,4
P(positivo) = 0,6
P(presente) = 0,6
P(ausente) = 0,4

O que se deseja em AM é a probabilidade *a posteriori*

Probabilidade

- Probabilidade *a priori x a posteriori* de um indivíduo estar doente
  - Probabilidade *a priori :*
    - Probabilidade de alguém esta cursando AM no ICMC
  - *Probabilidade a posteriori :*
    - Probabilidade de alguém esta cursando AM no ICMC dado que faz pós no ICMC
  - P(doente) ≠ P(doente/exame)

Probabilidade condicional

- Probabilidade de ocorrência de um evento depende da ocorrência de outro
  - P(A/B)
    - Probabilidade de ocorrência de um evento A depende da ocorrência de um evento B
      - Ex.: Probabilidade de estar doente (A) dado que um exame (B) deu positivo
    - Atributos (eventos) independentes: P(A/B) = P(A)

Probabilidade condicional

- Fácil estimar pela frequência as probabilidades *a priori*
  - P(B): prob. do resultado do exame ser positivo
  - P(A): prob. do resultado do paciente estar doente
  - P(B/A): prob. do resultado do exame ser positivo dado que o paciente esta doente
- Difícil estimar probabilidade *a posteriori*
  - P(A/B): probabilidade do paciente estar doente dado que seu exame deu positivo
    - Teorema (regra) de Bayes

Probabilidade condicional

- Lei da probabilidade condicional
  - $P(A/B) = P(A \cap B) / P(B)$
- Teorema de Bayes
  - Permite calcular probabilidade *a posteriori* de um evento
  - $P(A \cap B) = P(A/B)P(B) = P(B/A)P(A)$
  - $P(A/B) = P(B/A)P(A)/P(B)$
    - *Posteriori* = (verossimilhança x *priori) / evidência*
  - P(B): lei da probabilidade total

Probabilidade condicional

- Lei da probabilidade total
  - Evento A pode ter 2 possíveis resultados, A (A_1) e $\neg$A (A_2), que formam uma partição em Ω

$$P(B) = P(B \cap A_1) + P(B \cap A_2)$$
$$P(B) = P(B / A_1)P(A_1) + P(B / A_2)P(A_2)$$

  - Evento A pode ter n possíveis resultados mutuamente exclusivos, A_1, A_2, ..., A_n, que formam uma partição em Ω

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

Probabilidade condicional

De acordo com experiências passadas

Doença

presente: 8%
ausente: 92%

Teste

Doença	positivo	negativo
presente:	75%	4%
ausente:	25%	96%

P(Teste/Doença) = 0,75
P($\neg$Teste/ $\neg$ Doença) = 0,96
P(Teste) = P(Teste/Doença)P(Doença) + P(Teste/ $\neg$ Doença)P($\neg$ Doença)
P(Teste) = 0,75x0,08 + 0,25x0,92 = 0,29
...
P(Doença/Teste) = ?

Classificação Bayesiana

- Sejam y_i, i = 1, 2, ..., m, as possíveis classes
  - Novo exemplo pertence a classe com probabilidade *a posteriori* máxima
    - $Y_{MAP} = \arg\max_i P(y_i/X)$
- Definição de $P(y_i/X)$
  - $P(y_i/X) = P(X/y_i)\, P(y_i) / P(X)$

Classificação Bayesiana

- Exp. $P(X/y_i)\, P(y_i) / P(X)$ pode ser simplificada
  - P(X) é comum a todas as classes
  - Considerar as classes equiprováveis ($P(y_i) = P(y_j)$)
- Exemplo x pertence a classe com máxima verossimilhança
  - $h_{MV} = \arg\max_i P(X/y_i)$
- Difícil calcular valores
  - Precisa de um número de exemplos muito grande

Classificação Bayesiana

- Inferência Bayesiana
  - Cálculo da probabilidade *a posteriori* a partir da probabilidade *a priori*
- Várias alternativas para estimar $P(X/y_i)$
  - Produzem diferentes funções discriminantes
    - Ex.: Classificador Naive Bayes

Naive Bayes

- Classificador Bayesiano mais simples
- Assume que os atributos são independentes
  - $P(X/y_i) = P(x_1/y_i) * ... * P(x_d/y_i)$

$$P(y_i / X) \propto P(y_i) \prod_{j=1}^{d} P(x_j / y_i)$$
$$\log P(y_i / X) \propto \log P(y_i) + \sum_{j=1}^{d} \log P(x_j / y_i)$$

Naive Bayes

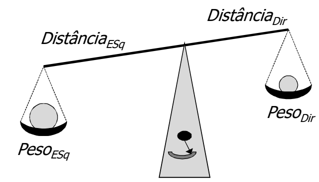
- Para duas classes

$$\log \frac{P(y_1/X)}{P(y_2/X)} \propto \log \frac{P(y_1)}{P(y_2)} + \sum_{j=1}^{d} \log \frac{P(x_j/y_1)}{P(x_j/y_2)}$$

  - Sinal do primeiro log indica a classe
  - Sinal de cada termo do somatório indica contribuição de cada atributo

Exemplo

- Conjunto de dados da UCI *Balance Scale*
  - Classe é o maior valor entre $D_{esq}.P_{esq}$ e $D_{dir}.P_{dir}$
  - 4 atributos preditivos



Y: classe
X_1: distância
X_2: peso

Exemplo

- Conjunto tem 625 exemplos em 3 classes
  - Esquerda, direita e equilíbrio
  - Domínio de valores para atributos preditivos é {1, 2, 3, 4, 5}
  - Definir P(Classe/Atribuitos)

	Equilíbrio	Esquerda	Direita
Freq(classe)	49	288	288
P(classe)	0,0784	0,4608	0,4608

P(Distancia$_{Esq1}$/Equilíbrio) P(Peso$_{Esq1}$/Equilíbrio) ...
P(Distancia$_{Esq2}$/Equilíbrio) P(Peso$_{Esq2}$/Equilíbrio) ...
...

Distribuição dos valores dos atributos

	Dist. Normal		Discretização				
Peso$_{Esq}$	**Média**	**DP**	**V1**	**V2**	**V3**	**V4**	**V5**
Equilibrada	2,938	1,42	10	11	9	10	9
Esquerda	3,611	1,23	17	43	63	77	88
Direita	2,399	1,33	98	71	53	38	28
Distância$_{Esq}$	**Média**	**DP**	**V1**	**V2**	**V3**	**V4**	**V5**
...							
Peso$_{Dir}$	**Média**	**DP**	**V1**	**V2**	**V3**	**V4**	**V5**
...							
Distância$_{Dir}$	**Média**	**DP**	**V1**	**V2**	**V3**	**V4**	**V5**
...							

Exercício

- Usar Naive Bayes para gerar um modelo probabilístico
  - Usar exemplos de treinamento (001, -1) e (110, +1)
  - Definir a classe dos exemplos: 111, 000, 100 e 011

Exercício

- Seja o seguinte cadastro de pacientes:

Nome	Febre	Enjôo	Manchas	Dores	Diagnóstico
João	sim	sim	pequenas	sim	doente
Pedro	não	não	grandes	não	saudável
Maria	não	sim	pequenas	não	saudável
José	sim	sim	grandes	sim	doente
Ana	sim	não	pequenas	sim	saudável
Leila	não	não	grandes	sim	doente

Exercício

- Utilizar Naive Bayes para induzir modelo capaz de distinguir:
  - Pacientes potencialmente saudáveis
  - Pacientes potencialmente doentes
- Testar o modelo para novos casos
  - (Luis, não, não, pequenas, sim)
  - (Laura, sim, sim, grandes, sim)

Conclusão

- Métodos baseados em probabilidade
- Teorema de Bayes
- Naive Bayes
- Classificadores Bayesianos
- Classificadores Bayesianos com k dependências

Perguntas