Aprendizado de Máquina

Aula 5: Algoritmos baseados em probabilidade

André C. P. L. F de Carvalho
ICMC/USP
andre@icmc.usp.br

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
    - Alimentação
    - ...

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

- Discriminam um objeto entre os possíveis rótulos (classes)
  - Qual a probabilidade de um dado objeto ter um dado rótulo
- Modelam a distribuição de probabilidade a posteriori (condicional) P(Y/X)
- Dado X, retornam a probabilidade de Y ocorrer
  - X: atributo(s) preditivo(s)
  - Y: atributo alvo
  - Ex.: Regressão logística

Métodos generativos

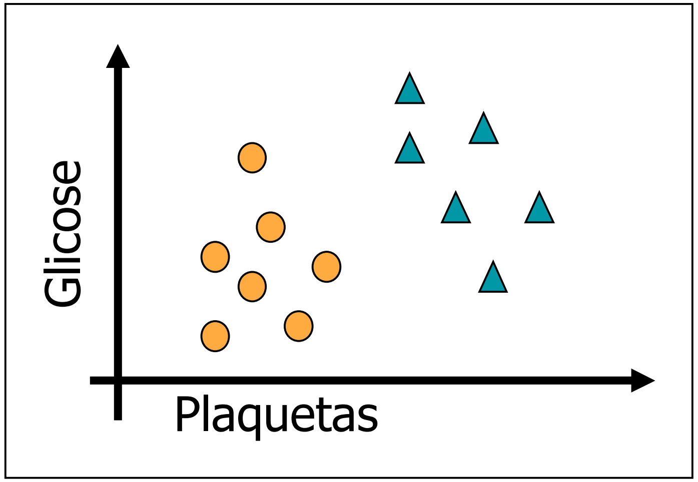
- Conhecem a distribuição dos dados
  - Sabe qual a probabilidade de existir um objeto X da classe Y
  - Podem gerar novos objetos
- Modelam a distribuição de probabilidade conjunta P(X,Y) (ou P(X) se não existirem rótulos)
  - Com a distribuição conjunta é possível derivar qualquer distribuição condicional
- Induzidos por algoritmos baseados no teorema de Bayes
  - Algoritmos Bayesianos
    - Ex.: *Naive Bayes*

Discriminativos x Generativos

Modelos Discriminativos	Modelos Generativos
Modelam a fronteira de decisão entre as classes	Modelam a distribuição de cada classe
Sabem se é provável que um objeto x tenha o rótulo y	Sabem se é provável que exista um objeto (x,y)
P(y/x)	P(y ,x)
Discriminam um objeto entre possíveis rótulos (classes)	Podem gerar novos objetos, pois conhecem a distribuição de cada classe

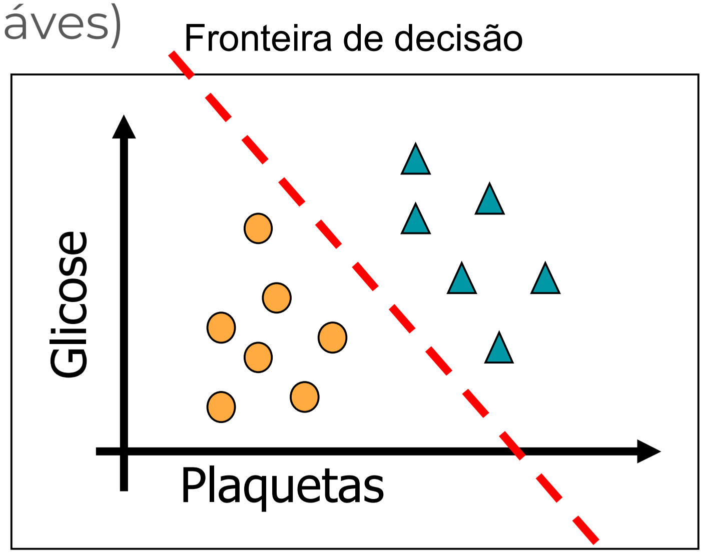
Fronteira de decisão

- Supor um conjunto de objetos, cada um representado por dois atributos preditivos (características, variáves)
    - Nível de glicose
    - Número de plaquetas no sangue
- Objetos podem ser representados em um espaço de atributos
    - Como classificar novos objetos?
    - Construir uma fronteira de decisão
    - Fácil ver em até 3 dimensões



Fronteira de decisão

- Supor um conjunto de objetos, cada um representado por dois atributos preditivos (características, variáves)
    - Nível de glicose
    - Número de plaquetas no sangue
- Objetos podem ser representados em um espaço de atributos
    - Como classificar novos objetos?
    - Construir uma fronteira de decisão
    - Fácil ver em até 3 dimensões



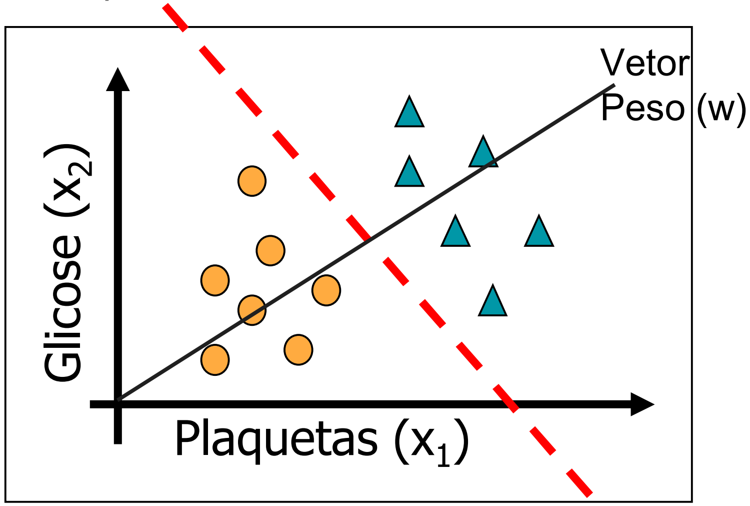
y = ax + b
Glicose = aPlaquetas + b

Fronteira de decisão

- Como construir a fronteira de decisão?
- Adicionar um novo vetor de peso, W
  - Cuja orientação será usada para definir a fronteira de decisão

Fronteira de decisão

- Supor um conjunto de objetos, cada um representado por dois atributos preditivos (características, variáves)
    - Nível de glicose
    - Número de plaquetas no sangue
- Objetos podem ser representados em um espaço de atributos
    - Como classificar novos objetos?
    - Construir uma fronteira de decisão
    - Fácil ver em até 3 dimensões



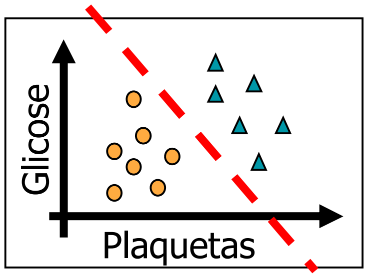
$f(x) = \sum_{i=1}^{n} w_i \, x_i - \theta$

$f(x) = w_1 x_1 + w_2 x_2 - \theta$

$f(x) = w_1 x_1 + w_2 x_2 + w_0$

Discriminante linear

- Fronteira de decisão pode ser representada por uma função linear
  - Função discriminante
    - Discrimina se um novo objeto pertence à classe positiva ou negativa
  - Ajusta parâmetros da função $f(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots$
  - Valor de *f(x)* $f(x) > 0$ *classe positiva* $f(x) < 0$ *classe negativa*
    - Distância de *x* à fronteira de decisão
      - Estima chance de *x* pertencer à classe positiva (negativa)



Fronteira de decisão

- Quando $f(x) = 0$
- Quando $w_1x_1 + w_2x_2 + w_0 = 0$

$$x_2 = -w_1/w_2\ x_1 - w_0/w_2$$
$$y = ax + b$$

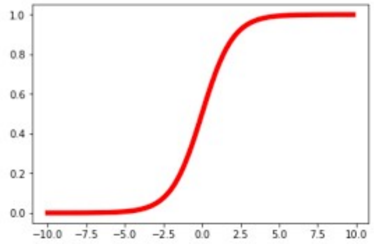
- Se temos os valores de w_2, w_1 e w_0, temos uma fronteira de decisão
- Problema: encontrar valores de w_2, w_1 e w_0

Discriminante linear

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

Regressão logística

- Apesar do nome, é usada para tarefas de classificação
- Estima probabilidade que um exemplo x pertencer a uma dada classe y: P(y/x)
  - Ajusta uma função logística a um conjunto de dados
    - Utiliza um conjunto de treinamento
    - Gera uma curve sigmoide
    - Produz uma fronteira de decisão
      - Hiperplano de separação



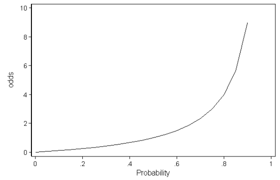
Regressão logística

- Chance de sucesso em relação ao fracasso (odds)
- Ex.: Suponha as seguintes probabilidades:
  - $\text{Probabilidade}_{\text{Sucesso}} = 0,8$
  - $\text{Probabilidade}_{\text{Fracasso}} = 1,0 - 0,8 = 0,2$
  - $\text{Chance}_{\text{Sucesso}} = \frac{P_{Sucesso}}{P_{Fracasso}} = 0,8/0,2 = 4:1$
  - $\text{Chance}_{\text{Positiva}} = \frac{P_{Positiva}}{P_{Negativa}} = 4:1$

Regressão logística

P_+ : Probabilidade de sucesso

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

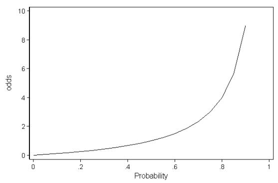


Probabilidade (P_+)	Chance (P_+:(1-P_+))	Log(Chance)
0,5	50:50 (1:1) = 1	0,00
0,9	90:10 (9:1) = 9	2,19
0,999	999:1 = 999	6,91
0,01	1:99 = 0,0101	-4,60
0,001	1:999 = 0,001001	-6,91

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

Regressão logística

P_+ : Probabilidade de sucesso



Probabilidade (P_+)	Chance (P_+: P_-)	log(Chance)
0,5	50:50 (1:1) = 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 consiga modelar *log(Chance)*
    - Permite estimar probabilidade usando modelo gerado por discriminante linear

Regressão logística

- Probabilidade de exemplo pertencer a classe positiva
  - Evento ocorreu

$$\text{Função logit} \longrightarrow \log\left(\frac{p_+(x)}{1-p_+(x)}\right) = f(x) = w_0 + w_1 x_1 + w_2 x_2 + \ldots$$

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

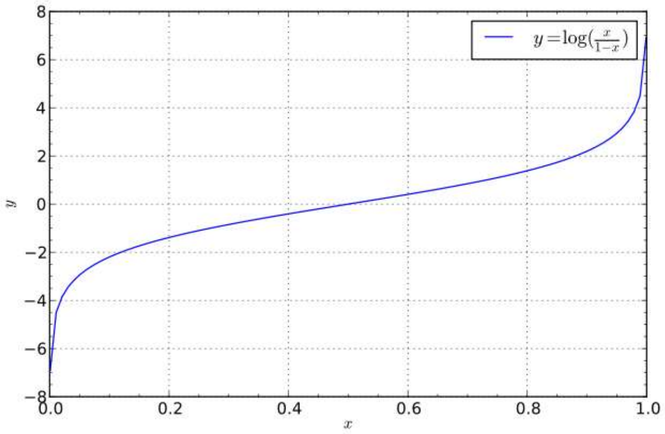
  - Treinamento

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

Função objetivo para ajuste dos pesos

Regressão logística

- Função logit
  - Inversa da função logística



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

Treinamento

- Encontrar valores de w_i que minimizem erro no conjunto de treinamento
  - Faz aproximação numérica utilizando método de máxima verossimilhança
  - Gradiente descendente estocástica
    - Para grandes conjuntos de dados
  - Exemplo para 1 atributo preditivo
    - w_0 : posição da função logística
    - w_1 : inclinação da função logística

Teoria das probabilidades

- Espaço amostral (Ω) : todos as possíveis observações de um experimento
- Evento (A): subconjunto de possíveis observações em Ω
- Ex.: Valores de um dado de 6 faces
    - $\Omega = \{1, 2, 3, 4, 5, 6\}$
    - A = valor do dado $< 3 = \{1, 2\}$
    - A = valor do dado é par= $\{2, 4, 6\}$
    - 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
      - $(A \cap B) = \varnothing$
      - $P(A \cap B) = 0$
      - $P(A \cup B) = P(A) + P(B)$

Teoria das probabilidades

- Probabilidade conjunta
  - Probabilidade de dois eventos ocorrerem simultaneamente
  - $P(A \cap B)$ ou $P(A,B)$
  - Se eventos 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

- Tarefa: dado o resultado de um exame, prever se paciente está doente
- Atributo preditivo
  - Resultado de exame
- Atributo alvo
  - Diagnóstico do paciente
  - Predição (classificação)

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 condicional

- Probabilidade de ocorrência de um evento dada a ocorrência de outro
    - P(A/B)
        - Probabilidade de ocorrência de um evento A dada a ocorrência de um evento B
            - Ex.: Probabilidade de estar doente (A) dado que um exame (B) deu positivo
        - Se os atributos (eventos) forem independentes: P(A/B) = P(A)
        - Caso não sejam, usar lei de probbilidade condicional

Probabilidade e AM

- P(A): probabilidade a priori do paciente está doente
- P(B): distribuição da variável preditiva B ser verdade (exame deu resultado positivo)
  - Evidência
- P(B/A): probabilidade de verosimilhança
  - Para um valor fixo de A, define verosimilhança (plausibilidade) de cada um dos possíveis valores de B
  - Verossímil: similar a verdade, provável
  - Verosimilhança: possui a qualidade de ser verossímil

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 da variável preditiva e probabilidade *a priori do atributo alvo* podem ser estimadas 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 da variável preditiva e probabilidade *a priori do atributo alvo* podem ser estimadas pela frequência

P(negativo) = 4/10 = 0,4
P(positivo) = 6/10 = 0,6
P(presente) = 6/10 = 0,6
P(ausente) = 4/10 = 0,4

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

De forma similar, é possível estimar a probabilidade de que um evento oorra para cada classe (probabilidade de verossimilhança)

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

De forma similar, é possível estimar a probabilidade de que um evento oorra para cada classe (probabilidade de verossimilhança)

P(negativo/presente) = 2/6 = 0.33
P(positivo/presente) = 4/6 = 0,66
P(negativo/ausente) = 2/4 = 0,5
P(positivo/ausente) = 2/4 = 0,5

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

Probabilidade a posteriori

- Fácil estimar pela frequência das probabilidades *a priori*
  - P(B): probabilidade do resultado do exame ser positivo
  - P(A): probabilidade do do paciente estar doente
  - P(B/A): probabilidade do resultado do exame ser positivo dado que o paciente está 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 a posteriori

- 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 a posteriori

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

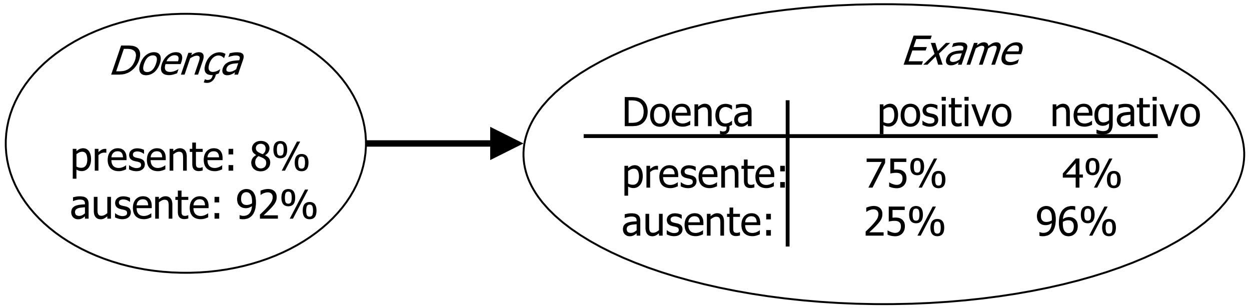
$$P(B) = P(B \cap A_1) + P(B \cap A_2)$$
$$P(B) = P(B / A_1)P(A_1) + P(B / A_2)P(A_2)$$

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

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

Modelo probabilístico gráfico

- Mostra os valores das probabilidades a priori e os valores das probabilidades condicionais
    - Modelo qualitativo: grafo cujos nós representam variáveis
    - Modelo quantitativo: tabelas com a distribuição das variáveis



Probabilidade a posteriori

De acordo com experiências passadas

Doença

presente: 8%
ausente: 92%

Exame

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

P(Exame/Doença) = 0,75
P(¬Exame/ ¬ Doença) = 0,96
P(Exame) = P(Exame/Doença)P(Doença) + P(Exame/¬ Doença)P(¬ Doença)
P(Exame) = 0,75x0,08 + 0,25x0,92 = 0,29
...
P(Doença/Exame) = ?

Probabilidade condicional

De acordo com com experiências passadas

Doença

presente: 8%
ausente: 92%

Exame

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

P(Exame/Doença) = 0,75
P(¬Exame/ ¬ Doença) = 0,96
P(Exame) = P(Exame/Doença)P(Doença) + P(Exame/¬ Doença)P(¬ Doença)
P(Exame) = 0,75x0,08 + 0,25x0,92 = 0,29
...
P(Doença/Exame) = P(Exame/Doença)P(Doença)/P(Exame) = 0.75x0,8/0,29
=

Classificação Bayesiana

- Sejam y_i, i = 1, 2, ..., m, as possíveis classes
  - Novo exemplo pertence à classe que tiver probabilidade *a posteriori* máxima
    - $Y_{MAP} = \arg\max_{i} P(y_i/X)$ (maior valor obtido variando *i*)
- Definição de $P(y_i/X)$
    - $P(y_i/X) = P(X/y_i)\,P(y_i) / P(X)$ (Teorema de Bayes)

Classificação Bayesiana

- Expressão $P(X/y_i)\ P(y_i) / P(X)$ 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 de classificação
    - 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

Naive Bayes

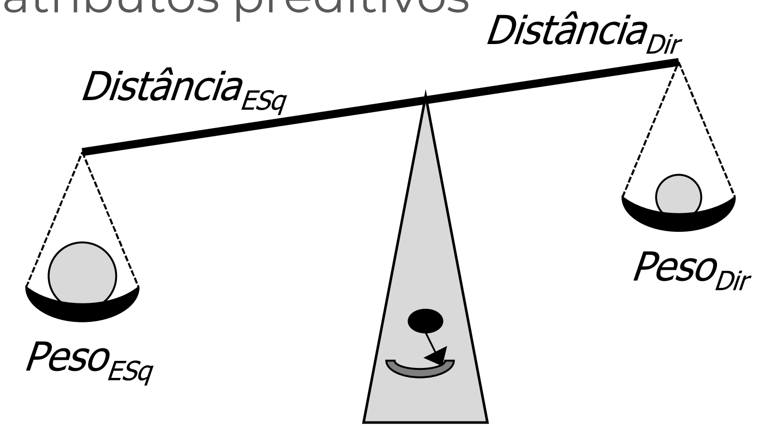
- Como é implementado
  - Todas as probabilidades necessárias são calculadas a partir do conjunto de dados de treinamento
  - Cálculo da probabilidade a priori de cada classe
    - Usar um contador para cada classe
  - Cálculo da probabilidade condicional de observar um valor de um atributo, dado que o exemplo pertence a uma dada classe
    - Necessário distinguir entre atributos nominais e atributos contínuos

Naive Bayes

- Cálculo da probabilidade condicional
  - Atributos preditivos nominais
    - Usar um contador para cada valor
  - Atributos preditivos contínuos (número de possíveis valores é infinito)
    - Assumir uma distribuição de probabilidade para os valores do atributo
      - Em geral é assumida a distribuição normal
    - Discretizar o atributo em uma fase de pré-processamento
      - Geralmente produz resultados melhores

Exemplo

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



Exemplo

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

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

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

P(Distancia$_{Esq}$/Equilíbrio) P(Peso$_{Esq}$/Equilíbrio) P(Distancia$_{Dir}$/Equilíbrio)...
P(Distancia$_{Esq}$/Esquerda) P(Peso$_{Esq}$/Esquerda) P(Distancia$_{Dir}$/Esquerda)...
...

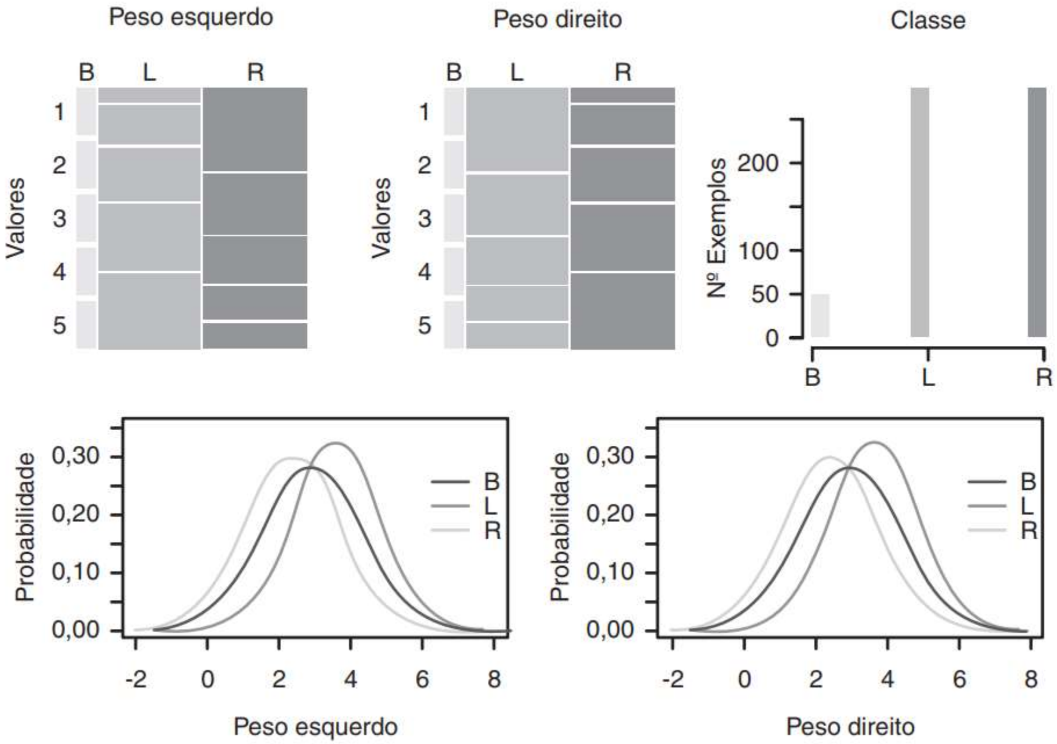
Discretização de valores

- Transformar valores numéricos em intervalos ou categorias
- Sub-tarefas
  - Definição de número de categorias
    - Geralmente feito pelo usuário
  - Definição de como mapear valores dos atributos contínuos para essas categorias
    - Definição do frequencia/largura dos intervalos
    - Geralmente feito pelo algoritmo
  - Exemplo: distribuir valores 1, 3, 4, 5, 6, 7, 9, 12 em dois intervalos
    - Por largura: {1, 3, 4, 5, 6} e {7, 9, 12}
    - Por frequência: {1, 3, 4, 5} e {6, 7, 9, 12}

Distribuição dos valores dos atributos

	Distribuição normal		Discretização				
$\boldsymbol{Peso}_{Esq}$	Média	Desvio padrão	V1	V2	V3	V4	V5
Equilibrado	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
$\boldsymbol{Distância}_{Esq}$	Média	Desvio padrão	V1	V2	V3	V4	V5
Equilibrado	2,938	1,42	10	11	9	10	9
Esquerda	3,611	1,22	17	43	63	77	88
Direita	2,399	1,33	98	71	53	38	28
$\boldsymbol{Peso}_{Dir}$	Média	Desvio padrão	V1	V2	V3	V4	V5
Equilibrado	2,938	1,42	10	11	9	10	9
Esquerda	2,399	1,33	98	71	53	38	28
Direita	3,611	1,22	17	43	63	77	88
$\boldsymbol{Distância}_{Dir}$	Média	Desvio padrão	V1	V2	V3	V4	V5
Equilibrado	2,938	1,42	10	11	9	10	9
Esquerda	2,399	1,33	98	71	53	38	28
Direita	3,611	1,22	17	43	63	77	88

Distribuição dos valores dos atributos



Conclusão

- Métodos baseados em probabilidade
- Discriminante linear
- Regressão logística
- Teorema de Bayes
- Naive Bayes

Fim do apresentação