CLASSIFICADORES BAEYSIANOS

Prof. André Backes | @progdescomplicada

Teorema de Bayes

- Frequentemente, uma informação é apresentada na forma de probabilidade condicional
  - Probabilidade de um evento ocorrer dada uma condição
    - A probabilidade de ocorrer um evento *B*, na condição de que outro evento *A* já tenha ocorrido
- Esse tipo de problema é tratado usando o Teorema de Bayes

Teorema de Bayes

- O Teorema de Bayes relaciona as probabilidades de *A* e *B* com suas respectivas probabilidades condicionadas

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}, para\ P(B) > 0$$

- Onde
  - *P(A)* e *P(B)*: probabilidades a **priori** de *A* e *B*
  - *P(B|A)* e *P(A|B)*: probabilidades a **posteriori** de *B* condicional a *A* e de *A* condicional a *B*, respectivamente.

Teorema de Bayes

- Probabilidade a priori
  - Probabilidade dada sem conhecimento de qualquer outro evento
    - Probabilidade de tirar um número par num dado: 1/2
- Probabilidade a posteriori
  - É a probabilidade condicional que é atribuída quando um evento relevante é considerado
    - Ao lançarmos um dado *N* vezes, teremos *a posteriori* que a distribuição dos valores tendem ao previsto *a priori*

Teorema de Bayes

- O Teorema de Bayes nos permite calcular a probabilidade a posteriori para um determinado padrão pertencente a uma determinada classe
- Em resumo

$$Prob\ Posteriori = \frac{Prob\ Priori\ * Distrib\ Prob}{Evidencia}$$

Teorema de Bayes | Exemplo

- Um médico sabe que a meningite causa torcicolo 50% das vezes
  - Probabilidade a priori de qualquer paciente ter meningite: 1/50000
  - Probabilidade a priori de qualquer paciente ter rigidez de nuca: 1/20
- Se um paciente tem rigidez de nuca (evidência), qual é a probabilidade *a posteriori* de ele ter meningite?

Teorema de Bayes | Exemplo

- *Dado*
  - *M*: meningite
  - *R*: rigidez no pescoço

$$P(M|R) = \frac{P(R|M)P(M)}{P(R)} = \frac{0{,}5 * 1/50000}{1/20} = 0{,}0002 = 0{,}02\%$$

Teorema de Bayes

- Notações básicas
  - x: vetor atributo, $x \in R^d$
  - R^d: espaço (Euclidiano) de atributos
  - $\{c_1, \ldots, c_n\}$: conjunto (finito) de n classes

Teorema de Bayes

- Vamos considerar a classificação de uma amostra *x* em uma classe c_i

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{P(x)}$$

- Temos que
  - $P(c_i)$: probabilidade a priori da classe c_i
  - $P(c_i|x)$: probabilidade a posteriori de c_i ser a classe de *x*
  - $P(x|c_i)$: função densidade de probabilidade da classe c_i
  - $P(x)$: densidade de probabilidade

Teorema de Bayes

- A densidade de probabilidade *P(x)* é definida como

$$P(x) = \sum_i P(c_i)P(x|c_i)$$

  - Note que a densidade de probabilidade é a soma das probabilidade de cada classe multiplicada por sua função densidade de probabilidade
  - Pode ser suprimida dependendo da aplicação

Classificação Bayesiana

- Escolhe a classe mais provável, dado um vetor de atributos *x*
  - *maximizar* $P(c_i|x)$
  - É sempre a escolha ótima!
- Problema:
  - Como estimar $P(c_i|x)$ *?*
    - Não conheço a classe que contém *x*
    - Porém eu conheço como se comportam os dados da classe c_i

Classificação Bayesiana

- Teorema de Bayes

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{P(x)}$$

$$P(x) = \sum_i P(c_i)P(x|c_i)$$

  - Dado um vetor *x*, *P(x)* é constante. A classe pode ser escolhida maximizando
    - A probabilidade à posteriori
    - A verossimilhança

Classificador de máximo à posteriori

- Do inglês *maximum a posteriori (MAP)*
  - Busca maximizar a probabilidade $P(c_i|x)$
  - Ignora $P(x)$
    - Valor é constante
    - Simplificação dos cálculos
  - Admite que as classes possuem probabilidades distintas de ocorrerem

$$P(c_i) \neq P(c_j)$$

Classificador de máximo à posteriori

- Como calcular?
  - Simplificamos o classificador MAP (*maximum a posteriori*) para

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{P(x)} \propto P(c_i)P(x|c_i)$$

  - Na prática, esse é um bom critério!
    - Como estimar $P(x|c_i)$? Veremos mais adiante...

Classificador de máxima verossimilhança

- Do inglês *Maximum Likelihood (ML)*
  - Intuitivamente verdadeiro
  - A priori, podemos admitir que todas as classes são equiprováveis (mesma probabilidade de ocorrerem)
  - Para um dado *x*, a classe mais provável é a que com maior probabilidade gera esse dado!

Classificador de máxima verossimilhança

- Como calcular?
  - Simplificamos o classificador ML (*maximum likelihood*) para

$$P(c_i|x) = \frac{P(c_i)P(x|c_i)}{P(x)} \propto P(x|c_i)$$

  - Na prática, esse é um bom critério!
    - Como estimar $P(x|c_i)$?

Como estimar $P(x,c_i)$?

- De modo geral, desconhece-se a forma analítica de $P(x|c_i)$, mas podemos estimá-lo a partir dos dados
  - Problema central em classificação
  - Podemos estimá-lo de duas formas:
    - Estimação paramétrica
    - Estimação não-paramétrica

Como estimar $P(x,c_i)$?

- Estimação paramétrica
  - Podemos assumir que $P(x|c_i)$ tem uma distribuição “conhecida”
    - Gaussiana
    - Uniforme
    - etc.
  - Temos então que estimar os parâmetros dessa distribuição

Como estimar $P(x,c_i)$?

- Estimação não-paramétrica
  - Temos que calcular $P(x|c_i)$ diretamente a partir dos dados de treinamento
    - Calcular as probabilidades individuais para cada combinação de valores discretos

Exemplo de Classificação

- Jogar tênis: sim ou não?

Tempo	Temperatura	Umidade	Vento	Jogo
Sol	Quente	Alta	Não	Não
Sol	Quente	Alta	Sim	Não
Nublado	Quente	Alta	Não	Sim
Chuva	Suave	Alta	Não	Sim
Chuva	Frio	Normal	Não	Sim
Chuva	Frio	Normal	Sim	Não
Nublado	Frio	Normal	Sim	Sim
Sol	Suave	Alta	Não	Não
Sol	Frio	Normal	Não	Sim
Nublado	Suave	Normal	Não	Sim
Sol	Suave	Normal	Sim	Sim
Nublado	Suave	Alta	Sim	Sim
Nublado	Quente	Normal	Não	Sim
Chuva	Suave	Alta	Sim	Não

Exemplo de Classificação

- Sabendo apenas a variável “tempo”
  - Teremos “jogo” se tivermos “Sol”?
    - P(jogo|tempo = “Sol”)
  - Vamos experimentar os dois classificadores
    - Classificador MAP
    - Classificador ML

Exemplo de Classificação

- Classificador MAP: **não joga!**
  - P(jogo|tempo) ∝ P(tempo|jogo)P(jogo)
    - P(jogo=sim) = 9/14 = 0,64
    - P(jogo=não) = 5/14 = 0,36

    - P(tempo="Sol"|jogo=sim) = 2/9 = 0,22
    - P(tempo="Sol"|jogo=não) = 3/5 = 0,60

    - P(jogo=sim|tempo="Sol") ∝ 0.22 x 0.64 = 0,14
    - P(jogo=não|tempo="Sol") ∝ 0.60 x 0.36 = 0,22

Exemplo de Classificação

- Classificador ML: **não joga!**
  - P(jogo|tempo) ∝ P(tempo|jogo)
    - P(tempo="Sol"|jogo=sim) = 2/9 = 0,22
    - P(tempo="Sol"|jogo=não) = 3/5 = 0,60

Trabalhando com vários atributos

- No exemplo anterior, trabalhamos a classificação com apenas 1 atributo
  - *x* pode ser um vetor contendo vários atributos
- Se o vetor *x* for muito grande (muitos atributos), se torna difícil calcular $P(x,c_i)$
  - Maldição da dimensionalidade

Trabalhando com vários atributos

- Uma solução seria assumir a independência entre os vários atributos
- Relembrando *eventos independentes*
  - A ocorrência do evento *A* em nada interfere na probabilidade de ocorrência do outro evento, *B*
  - A probabilidade de que ambos ocorrerem é igual ao produto de suas probabilidades
    - P(A e B) = P(A ∩ B) = P(A) * P(B)

Trabalhando com vários atributos

- ***Classificador naive Bayes*** faz isso.
  - Utiliza o Teorema de Bayes
  - Trabalha com a hipótese de independência entre atributos
    - Atributos não correlacionados

Classificador *naive* Bayes

- Considerando os atributos *x(1),...,x(p)* independentes entre si, temos

  - Independência entre atributos
    - $P(x(1),...,x(p)) = P(x(1))* P(x(2))* ... *P(x(p))$

  - Independência condicional
    - $P(x(1),...,x(p)|c_i) = P(x(1)|c_i)* P(x(2)|c_i)* ... * P(x(p)|c_i)$

Classificador *naive* Bayes

- Consequentemente, o Teorema de Bayes passa a ser descrito como sendo

$$P(c_i|x(1),\dots,x(p)) = \frac{P(c_i)\prod_{j=1}^{p} P(x(j)|c_i)}{\prod_{j=1}^{p} P(x(j))}$$

Classificador *naive* Bayes

- Porque assumir independência entre os atributos *x(1),...,x(p)*?
  - Estimar as probabilidades conjuntas exige uma quantidade mínima de exemplos para cada combinação possível
    - *P(x(1),...,x(p))*
    - *P(x(1),...,x(p)|c_i)*
  - Isso é impraticável para muitos atributos

Classificador *naive* Bayes

- Porque assumir independência entre os atributos *x(1),...,x(p)*?
  - Essa hipótese (independência entre atributos) é quase sempre violada
  - Mas, na prática, o classificador *naive* Bayes se mostra bastante competitivo

Exemplo de Classificação

- Estimar a probabilidade de cada valor da classe (ou atributo meta) dados os valores dos demais atributos

Tempo			Temperatura			Umidade			Vento			Jogo	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3	Quente	2	2	Alta	3	4	Não	6	2	9	5
Nublado	5	0	Suave	4	2	Normal	6	1	Sim	3	3		
Chuva	2	2	Frio	3	1								
Sol	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Não	6/9	2/5	9/14	5/14
Nublado	5/9	0/5	Suave	4/9	2/5	Normal	6/9	1/5	Sim	3/9	3/5		
Chuva	2/9	2/5	Frio	3/9	1/5								

Tempo	Temperatura	Umidade	Vento	Jogo
Sol	Quente	Alta	Não	Não
Sol	Quente	Alta	Sim	Não
Nublado	Quente	Alta	Não	Sim
Chuva	Suave	Alta	Não	Sim
Chuva	Frio	Normal	Não	Sim
Chuva	Frio	Normal	Sim	Não
Nublado	Frio	Normal	Sim	Sim
Sol	Suave	Alta	Não	Não
Sol	Frio	Normal	Não	Sim
Nublado	Suave	Normal	Não	Sim
Sol	Suave	Normal	Sim	Sim
Nublado	Suave	Alta	Sim	Sim
Nublado	Quente	Normal	Não	Sim
Chuva	Suave	Alta	Sim	Não

Exemplo de Classificação

- Para um novo dia, teremos jogo?

Tempo	Temperatura	Umidade	Vento	Jogo
Sol	Frio	Alta	Sim	?

Tempo			Temperatura			Umidade			Vento			Jogo	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3	Quente	2	2	Alta	3	4	Não	6	2	9	5
Nublado	5	0	Suave	4	2	Normal	6	1	Sim	3	3		
Chuva	2	2	Frio	3	1								
Sol	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Não	6/9	2/5	9/14	5/14
Nublado	5/9	0/5	Suave	4/9	2/5	Normal	6/9	1/5	Sim	3/9	3/5		
Chuva	2/9	2/5	Frio	3/9	1/5								

Exemplo de Classificação

- Para um novo dia, teremos jogo?
  - Resposta: **Não!**

P(Sim|Sol,Frio,Alta,Sim) = (2/9 * 3/9 * 3/9 * 3/9 * 9/14) / P(Sol,Frio,Alta,Sim)

P(Sim|Sol,Frio,Alta,Sim) = 0,0053 / P(Sol,Frio,Alta,Sim)

P(Não|Sol,Frio,Alta,Sim) = (3/5 * 1/5 * 4/5 * 3/5 * 5/14) / P(Sol,Frio,Alta,Sim)

P(Não|Sol,Frio,Alta,Sim) = **0,0206** / P(Sol,Frio,Alta,Sim)

Tempo			Temperatura			Umidade			Vento			Jogo	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3	Quente	2	2	Alta	3	4	Não	6	2	9	5
Nublado	5	0	Suave	4	2	Normal	6	1	Sim	3	3		
Chuva	2	2	Frio	3	1								
Sol	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Não	6/9	2/5	9/14	5/14
Nublado	5/9	0/5	Suave	4/9	2/5	Normal	6/9	1/5	Sim	3/9	3/5		
Chuva	2/9	2/5	Frio	3/9	1/5								

Problema da Frequência Zero

- Determinado valor não aparece no treinamento, mas aparece no teste?
  - Exemplo: Tempo = “Nublado” para a classe “Não”
    - Probabilidade correspondente será zero
    - P(Nublado | “Não”) = 0

Tempo	Temperatura	Umidade	Vento	Jogo
Sol	Quente	Alta	Não	Não
Sol	Quente	Alta	Sim	Não
Nublado	Quente	Alta	Não	Sim
Chuva	Suave	Alta	Não	Sim
Chuva	Frio	Normal	Não	Sim
Chuva	Frio	Normal	Sim	Não
Nublado	Frio	Normal	Sim	Sim
Sol	Suave	Alta	Não	Não
Sol	Frio	Normal	Não	Sim
Nublado	Suave	Normal	Não	Sim
Sol	Suave	Normal	Sim	Sim
Nublado	Suave	Alta	Sim	Sim
Nublado	Quente	Normal	Não	Sim
Chuva	Suave	Alta	Sim	Não

Problema da Frequência Zero

- Nesse caso, a probabilidade a posteriori também será zero
  - P(“Não” | Nublado, ...) = 0
    - Multiplicação das probabilidades
    - Não importa as probabilidades dos outros atributos
- Zerar a probabilidade a posteriori é muito radical
  - A base de treinamento pode não ser totalmente representativa
    - Classes minoritárias podem ter valores raros

Problema da Frequência Zero

- Estimador de Laplace é uma solução possível
  - Ideia: adicionar 1 unidade fictícia para cada combinação de valor-classe
  - Resultados
    - Valores sem exemplos de treinamento passam a conter 1 exemplo
    - As probabilidades nunca serão zero

Problema da Frequência Zero

- Exemplo: Tempo = “Nublado” e Classe = “Não”
  - Somar 1 para cada combinação valor-classe
  - Somar 3 na base (3 combinações valor-classe)
    - Sol: $\frac{3}{5} \rightarrow \frac{3+1}{5+3}$
    - Nublado: $\frac{0}{5} \rightarrow \frac{0+1}{5+3}$
    - Chuva: $\frac{2}{5} \rightarrow \frac{2+1}{5+3}$
  - Isso deve ser feito para todas as classes
    - Do contrário, estamos inserindo viés nas probabilidades de apenas uma classe

Problema da Frequência Zero

- **Estimativa m**: solução mais geral
  - Idéia: adicionar **múltiplas** unidades fictícias para cada combinação de valor-classe
  - Exemplo: Tempo = “Nublado” e Classe = “Não”
    - Sol: $\frac{3}{5} \rightarrow \frac{3+\frac{m}{3}}{5+m}$
    - Nublado: $\frac{0}{5} \rightarrow \frac{0+\frac{m}{3}}{5+m}$
    - Chuva: $\frac{2}{5} \rightarrow \frac{2+\frac{m}{3}}{5+m}$

Valores ausentes

- O que fazer se uma amostra não tiver o valor de um atributo?

Tempo	Temperatura	Umidade	Vento	Jogo
??	Frio	Alta	Sim	???

  - No treinamento
    - Devemos excluir a amostra do conjunto de treinamento
  - Na classificação
    - Devemos considerar apenas os demais atributos da amostra

Valores ausentes | Exemplo

Tempo	Temperatura	Umidade	Vento	Jogo
??	Frio	Alta	Sim	???

P(Sim|Frio,Alta,Sim) = (3/9 * 3/9 * 3/9 * 9/14) = 0,0238

P(Não|Frio,Alta,Sim) = (1/5 * 4/5 * 3/5 * 5/14) = 0,0343

Pobabilidade(Sim) = 0,0238 / (0,0238 + 0,0343) = 41%

Pobabilidade(Não) = 0,0343 / (0,0238 + 0,0343) = **59%**

Tempo			Temperatura			Umidade			Vento			Jogo	
	Sim	Não		Sim	Não		Sim	Não		Sim	Não	Sim	Não
Sol	2	3	Quente	2	2	Alta	3	4	Não	6	2	9	5
Nublado	5	0	Suave	4	2	Normal	6	1	Sim	3	3		
Chuva	2	2	Frio	3	1								
Sol	2/9	3/5	Quente	2/9	2/5	Alta	3/9	4/5	Não	6/9	2/5	9/14	5/14
Nublado	5/9	0/5	Suave	4/9	2/5	Normal	6/9	1/5	Sim	3/9	3/5		
Chuva	2/9	2/5	Frio	3/9	1/5								

Trabalhando com atributos contínuos

- São atributos que assumem valores dentro de intervalos de números reais
  - Peso das pessoas em uma sala
  - Altura das pessoas em uma sala
  - Distância entre cidades
- Alternativa 1
  - Discretizar os dados
    - Muita informação pode vir a ser perdida

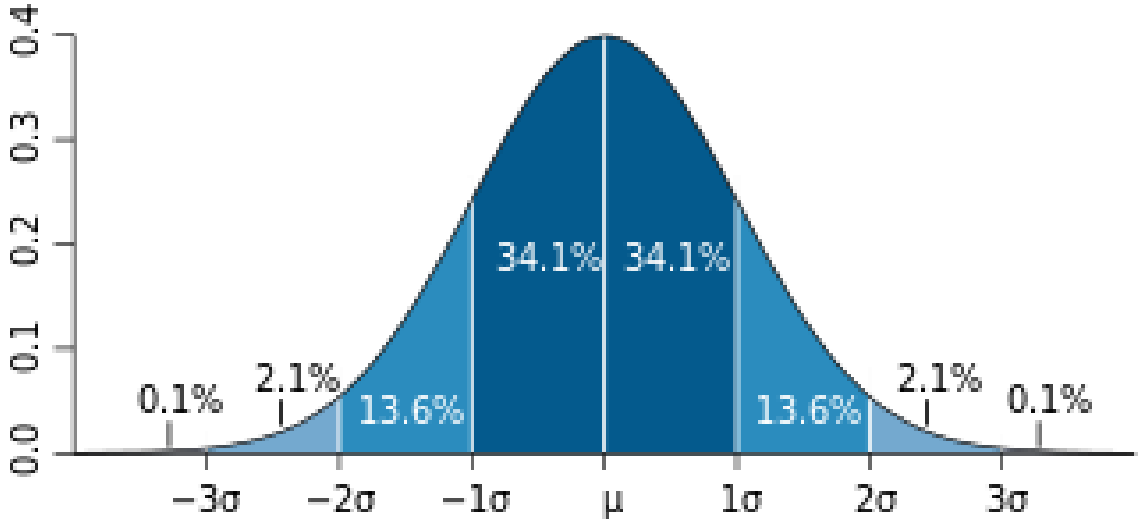
Trabalhando com atributos contínuos

- Alternativa 2
  - Considerar uma função de densidade de probabilidade
    - Isto nos permite estimar as probabilidades
  - Geralmente, usa-se a distribuição Normal ou Gaussiana

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

Trabalhando com atributos contínuos

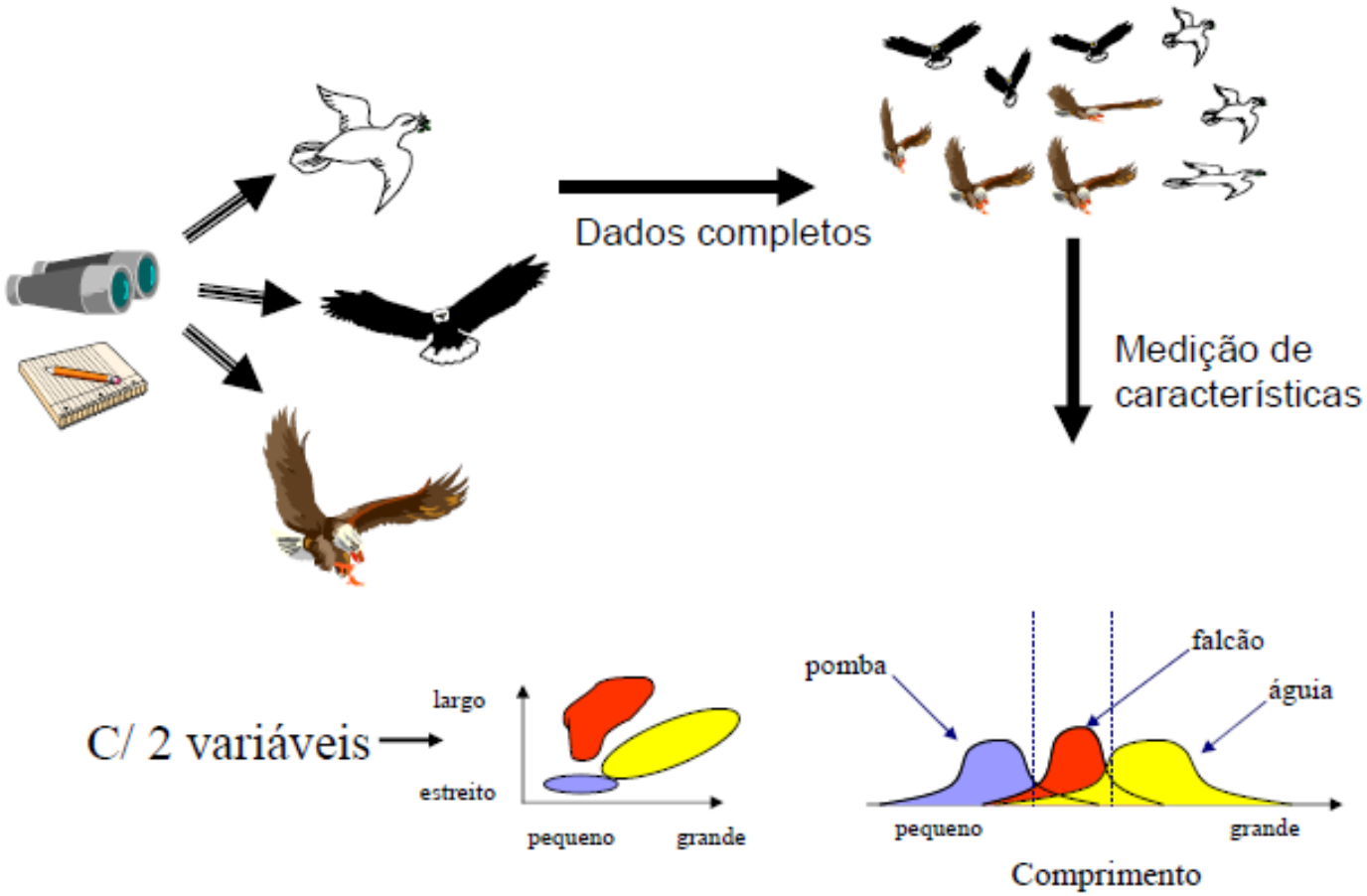
- Distribuição Normal ou Gaussiana
  - Inteiramente descrita pela média e desvio padrão dos dados
  - O desvio padrão define a área sob a curva, e para cada valor de desvio padrão corresponde uma proporção de casos da população



Trabalhando com atributos contínuos

- Convenientemente, o Teorema de Bayes permite usar o valor da densidade de probabilidade para estimar a probabilidade de um valor pontual

Trabalhando com atributos contínuos



Trabalhando com atributos contínuos

- Valor da densidade
  - Temperatura = 66 e jogo = “Sim”

Tempo			Temperatura		Umidade		Vento			Jogo	
	Sim	Não	Sim	Não	Sim	Não		Sim	Não	Sim	Não
Sol	2	3	64, 68,	65, 71,	65, 70,	70, 85,	Não	6	2	9	5
Nublado	4	0	69, 70,	72, 80,	70, 75,	90, 91,	Sim	3	3		
Chuva	3	2	72, ...	85, ...	80, ...	95, ...					
Sol	2/9	3/5	μ= 73	μ= 75	μ= 79	μ= 86	Não	6/9	2/5	9/14	5/14
Nublado	4/9	0/5	σ= 6,2	σ= 7,9	σ= 10,2	σ= 9,7	Sim	3/9	3/5		
Chuva	3/9	2/5									

$$f(temperatura = 66, \text{jogo} = sim) = \frac{1}{\sqrt{2\pi(6{,}2)^2}} e^{\frac{-(66-73)^2}{2(6{,}2)^2}}$$

$$f(temperatura = 66, \text{jogo} = sim) = 0{,}0340$$

Por que usar o classificador naive Bayes

- Principais características
  - É robusto a ruídos isolados (*outliers*)
    - *Outliers* afetam pouco o cálculo das probabilidades
  - É robusto a atributos irrelevantes
    - Esses atributos afetam pouco as probabilidades relativas entre classes

Por que usar o classificador naive Bayes

- Principais características
  - Capaz de classificar amostras com valores ausentes
  - Considera os atributos como igualmente importantes
  - Complexidade computacional linear em todas as variáveis do problema

Por que usar o classificador naive Bayes

- Principais características
  - Desempenho pode ser afetado pela presença de atributos correlacionados
    - Mas muitas vezes não é
- Desempenho ruim?
  - Presença muito significativa de atributos correlacionados
    - Atributos redundantes: fazer seleção de atributos
    - Tentar outra abordagem (Redes Bayesianas?)

Wrapper Naive Bayes

- Wrapper
  - Processo de seleção de atributos via *algoritmo guloso*
  - Esse processo de seleção atua junto com o *naive Bayes*
- Funcionamento
  - Para um único atributo, seleciona o melhor classificador *naive Bayes*
    - Avalia todos os atributos em um conjunto de dados de teste

Wrapper Naive Bayes

- Funcionamento
  - Enquanto houver melhora no desempenho do classificador, faça
    - Selecione o melhor classificador *naive Bayes* com os atributos já selecionados anteriormente adicionados a um dentre os atributos ainda não selecionados

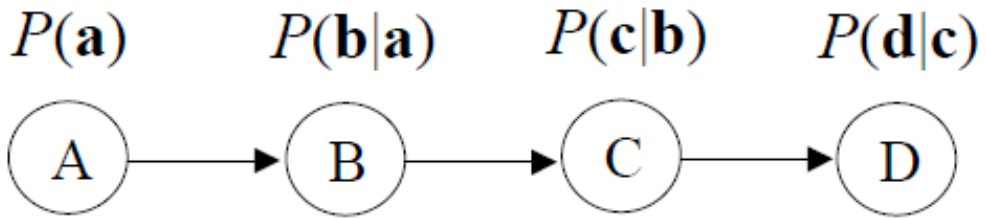
Wrapper Naive Bayes

- Seleção de atributos via *wrapper*
  - Em geral, torna o método muito custoso em termos computacionais
  - Apesar disso, o *Wrapper Naive Bayes* é relativamente rápido
    - Isso se deve a simplicidade e eficiência computacional do *naive Bayes*

Redes Bayesianas

- Usa um modelo de probabilidades na forma de um grafo
  - Explora a informação estrutural no raciocínio com variáveis
  - Usa relações probabilísticas entre variáveis
  - Assume relações causais
- Teoricamente, são capazes de solucionar o Calcanhar de Aquiles do *naive Bayes*

Redes Bayesianas



- Definição formal
  - Grafo acíclico
  - nó: variável aleatória (atributo)
  - Vértice (ou arco): efeito, causa
    - A afeta B→B condicionado a A
  - Cada nó condicionalmente independente dos não descendentes
  - Representa probabilidade conjunta das variáveis

$$P(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d}) = P(\mathbf{a})P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{c}\,|\,\mathbf{b})P(\mathbf{d}\,|\,\mathbf{c})$$

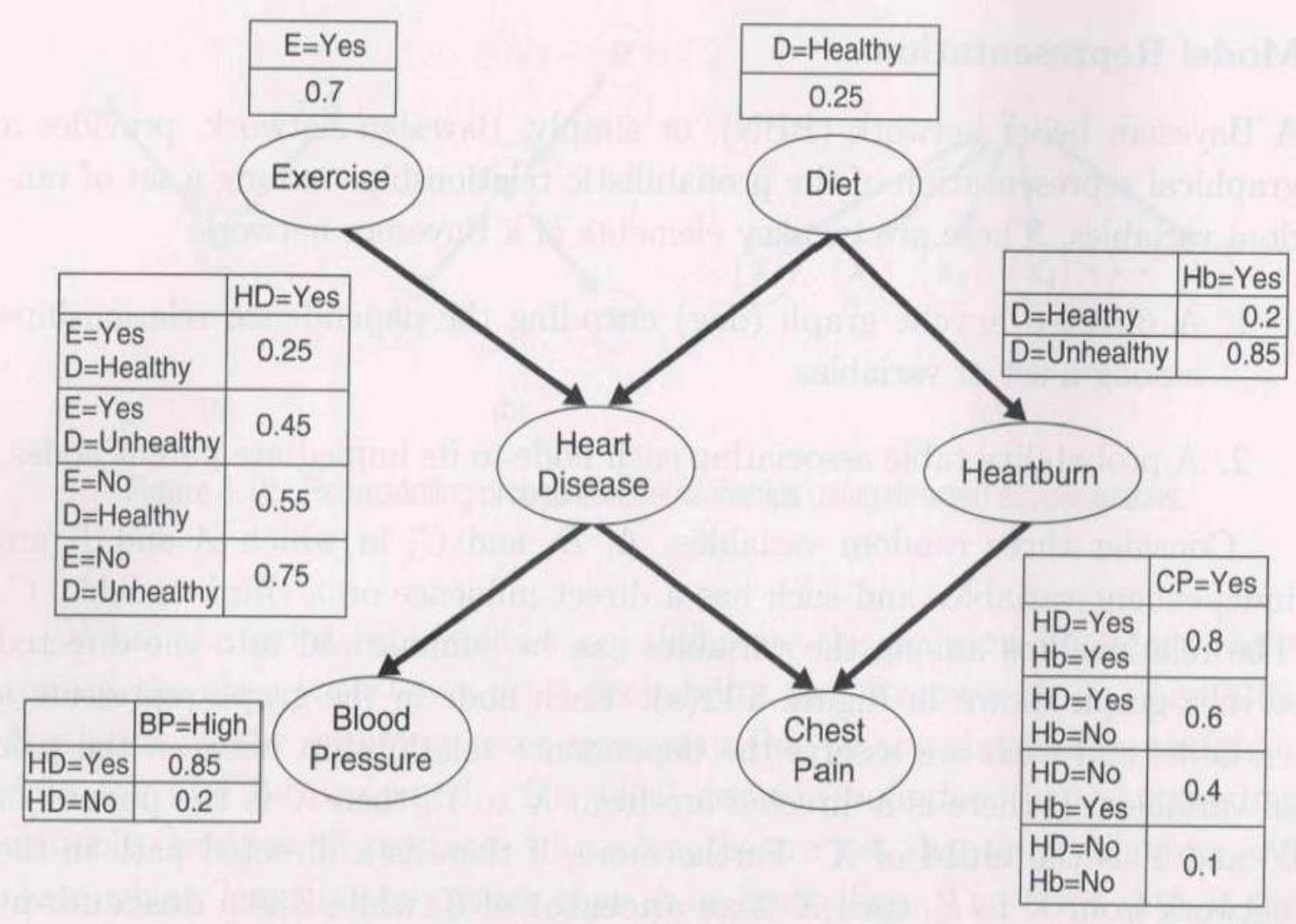
$$P(\mathbf{d}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a},\mathbf{b},\mathbf{c},\mathbf{d})$$

$$= \sum_{\mathbf{a},\mathbf{b},\mathbf{c}} P(\mathbf{a})P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{c}\,|\,\mathbf{b})P(\mathbf{d}\,|\,\mathbf{c})$$

$$= \sum_{\mathbf{c}} P(\mathbf{d}\,|\,\mathbf{c}) \underbrace{\sum_{\mathbf{b}} P(\mathbf{c}\,|\,\mathbf{b}) \underbrace{\sum_{\mathbf{a}} P(\mathbf{b}\,|\,\mathbf{a})P(\mathbf{a})}_{P(\mathbf{b})}}_{P(\mathbf{c})}$$

$$P(\mathbf{d})$$

Redes Bayesianas



Redes Bayesianas

- A obtenção da rede não é um problema trivial
  - Como determinar a topologia do grafo?
    - Não é uma tarefa simples de ser sistematizada
    - Especialmente sem conhecimento de domínio do problema

Redes Bayesianas

- A obtenção da rede não é um problema trivial
  - Somente podemos determinar as probabilidades (como no *naive Bayes*) se todos os atributos forem observáveis
    - Do contrário, é preciso um método de otimização para estimar essas probabilidades a partir dos atributos observáveis

Redes Bayesianas

- Além disso, a inferência também não é trivial
  - Teoricamente, as probabilidades de qualquer subconjunto das variáveis da rede podem ser obtidas (inferidas) a partir dos valores observados e/ou probabilidades das demais variáveis

Redes Bayesianas

- Além disso, a inferência também não é trivial
  - A inferência exata é um problema **NP-hard (ou NP-complexo)**
    - Só pode ser resolvido por um algoritmo não-determinístico polinomial
    - Ex.: Problema do caixeiro viajante
  - Até mesmo a obtenção de boas aproximações pode ser um problema complexo
    - Apesar disso, seu poder e flexibilidade têm motivado o interesse e pesquisa crescentes nesses modelos

Funções Discriminantes

- O que é uma função discriminante?
  - Consiste de uma combinação linear de características observadas e que apresente maior poder de discriminação entre populações
    - Busca minimizar as probabilidades de má classificação, quando as populações são normalmente distribuídas com média e variância conhecidas

Funções Discriminantes

- A função discriminante permite decidir entre as possíveis classes de uma amostra
  - $P(c_i)P(x|c_i) > P(c_j)P(x|c_j)$
  - Permite dividir o espaço em diferentes regiões, cada uma associada a uma classe

Funções Discriminantes

- Funções discriminação para a distribuição normal
  - Definem como as características são combinadas
  - Naive Bayes
    - Hipótese de independência entre atributos
  - Análise Discriminante Lineares (LDA) e Análise Discriminante Quadráticos (QDA)
    - Atributos correlacionados

Funções Discriminantes

- Distribuição Normal ou Gaussiana

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

- Forma multivariada da distribuição Normal ou Gaussiana
  - *p* atributos

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)}$$

Funções Discriminantes

- Considerando a seguinte função discriminante

$$p(c_i|x) > p(c_j|x)$$

$$p(c_i|x) = p(x|c_i)P(c_i)$$

$$p(c_i|x) = \frac{1}{(2\pi)^{p/2}} \frac{1}{|\Sigma_i|^{1/2}} e^{-\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i)} P(c_i)$$

$$p(c_i|x) = -\frac{1}{2}\ln|\Sigma_i| - \frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) + \ln P(c_i)$$

$$p(c_i|x) = \ln|\Sigma_i| + (x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - 2\ln P(c_i)$$

Funções Discriminantes

- A função discriminante baseada na distribuição de Gauss

$$p(c_i|x) = \ln|\Sigma_i| + (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - 2 \ln P(c_i)$$

- Pode ser utilizada de formas distintas
  - Naive Bayes
  - Análise Discriminante Lineares (LDA)
  - Análise Discriminante Quadráticos (QDA)

Funções Discriminantes

- Naive Bayes
  - Hipótese de independência entre atributos
  - Matriz de covariâncias: diagonal principal (variância de cada atributo)
    - normalização dos dados por score-z

$$\Sigma_{p\times p} = \begin{bmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & \sigma_p^2 \end{bmatrix}$$

$$p(c_i|x) = \ln|\Sigma_\mathrm{i}| + (x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - 2\ln P(c_i)$$

Funções Discriminantes

- Análise Discriminante Lineares (LDA)
  - Fisher Discriminant Analysis
  - Matriz de covariâncias é igual para todas as classes

$$p(c_i|x) = \ln|\Sigma_i| + (x - \mu_i)^T \Sigma_i^{-1} (x - \mu_i) - 2 \ln P(c_i)$$

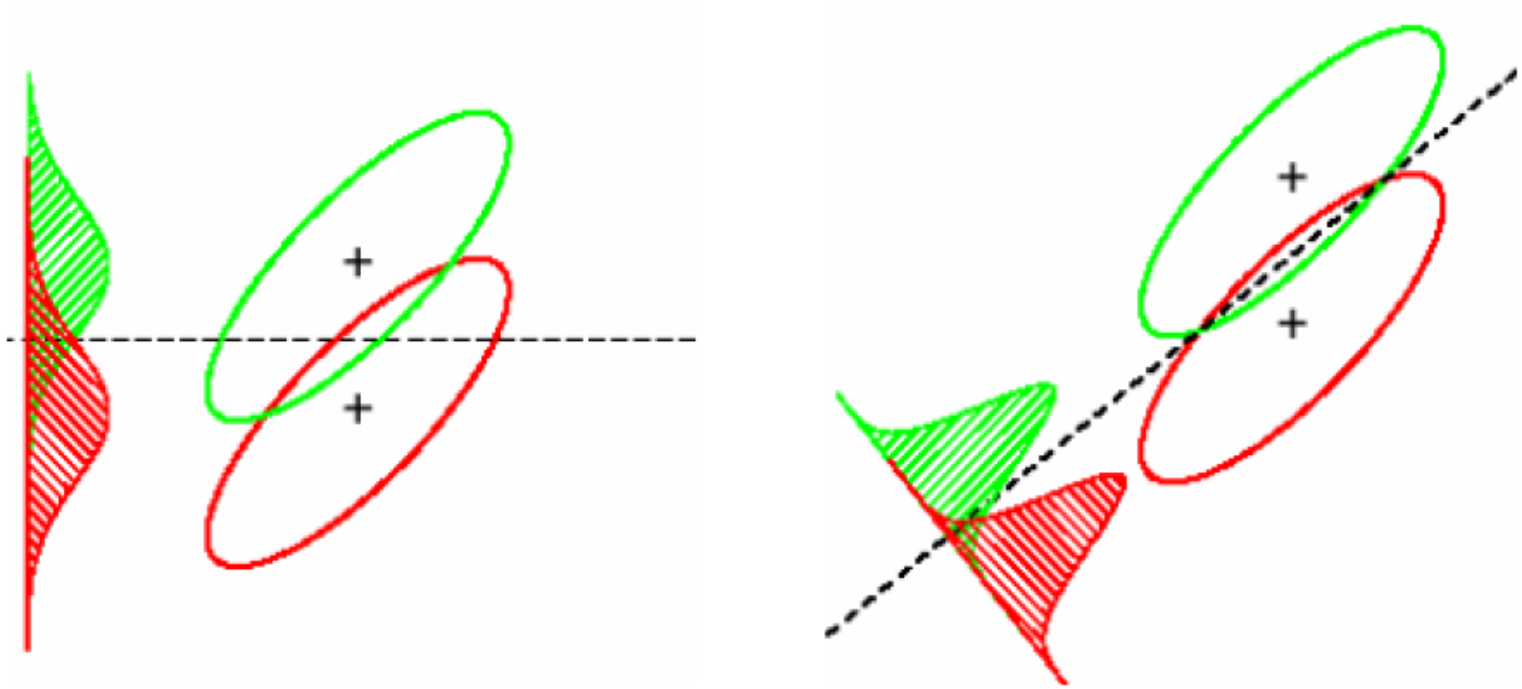
$$p(c_i|x) = (x - \mu_i)^T \Sigma^{-1} (x - \mu_i) - 2 \ln P(c_i)$$

Funções Discriminantes

- LDA: funcionamento
  - Para um problema linearmente separável, rotaciona os dados de modo a maximizar a distância entre as classes e minimizar a distância intra-classe
    - Melhor direção capaz de separar os dados projetados
    - Transformação linear que maximiza a distância entreclasses e minimiza a distância intra-classe

Funções Discriminantes

- LDA: funcionamento



Funções Discriminantes

- Análise Discriminante Quadráticos (QDA)
  - Similar a Análise Discriminante Lineares (LDA)
  - Matriz de covariâncias é diferente para cada classe

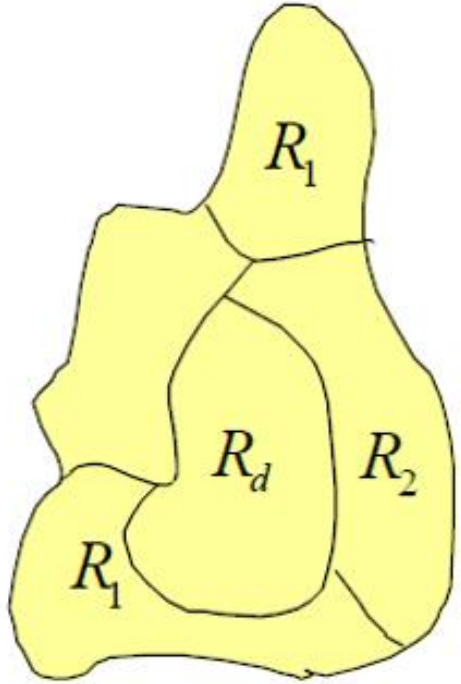
$$p(c_i|x) = \ln|\Sigma_i| + (x - \mu_i)^T \Sigma_i^{-1}(x - \mu_i) - 2\ln P(c_i)$$

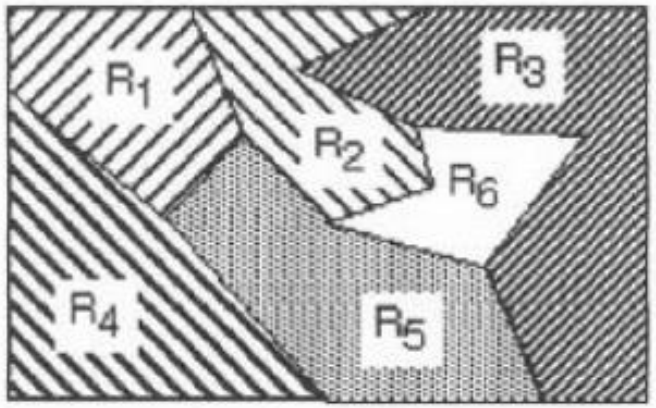
Superfície de Decisão

- Como funciona um classificador?
  - De modo geral, um classificador particiona o espaço de características em volumes designados regiões de decisão
  - Todos os vetores de atributos no interior de uma região de decisão são atribuídos à mesma categoria.
    - Pode ser simplesmente conexa, ou pode consistir em duas ou mais sub-regiões não adjacentes

Superfície de Decisão

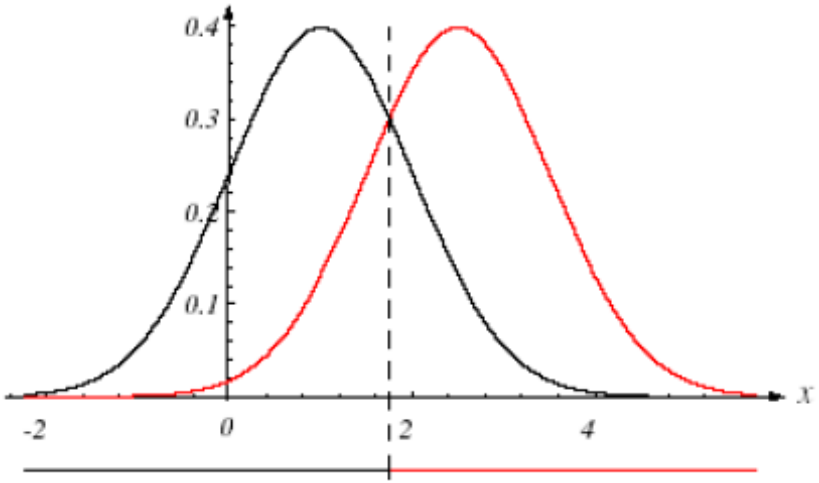
- Como funciona um classificador?
  - Uma superfície de decisão (ou de separação) separa duas dessas regiões
    - Elas representam pontos onde existem “empates” entre duas ou mais categorias

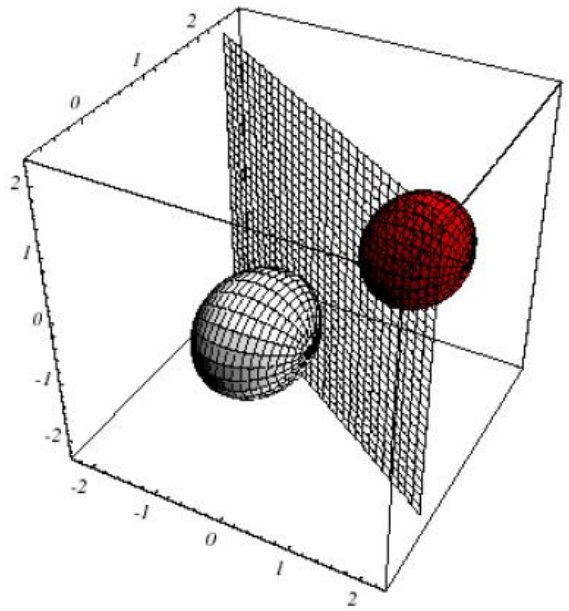


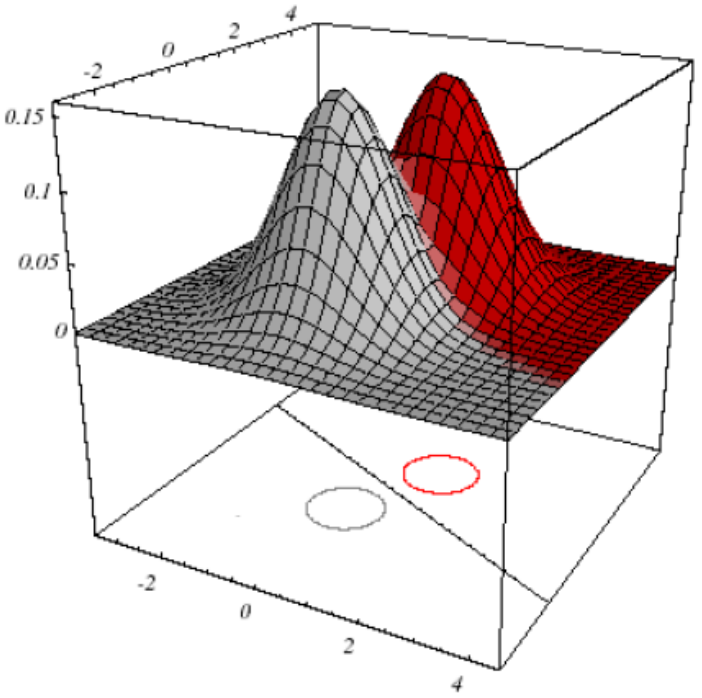


Superfície de Decisão

- Superfícies de decisão para um classificador linear: 2 classes

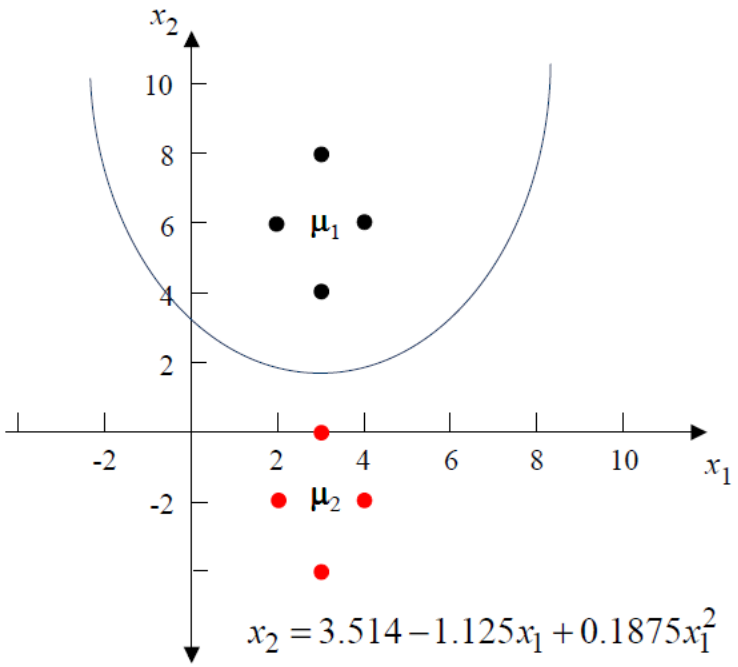




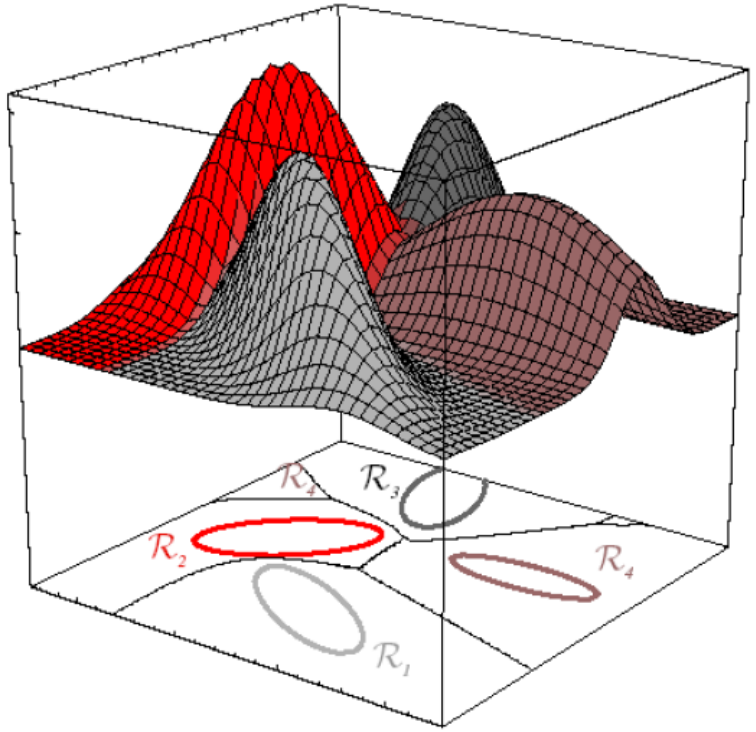


Superfície de Decisão

- Superfícies de decisão para um classificador quadrático



2 classes



4 classes