Tema 04

O Classificador Bayesiano

Professora:
Ariane Machado Lima

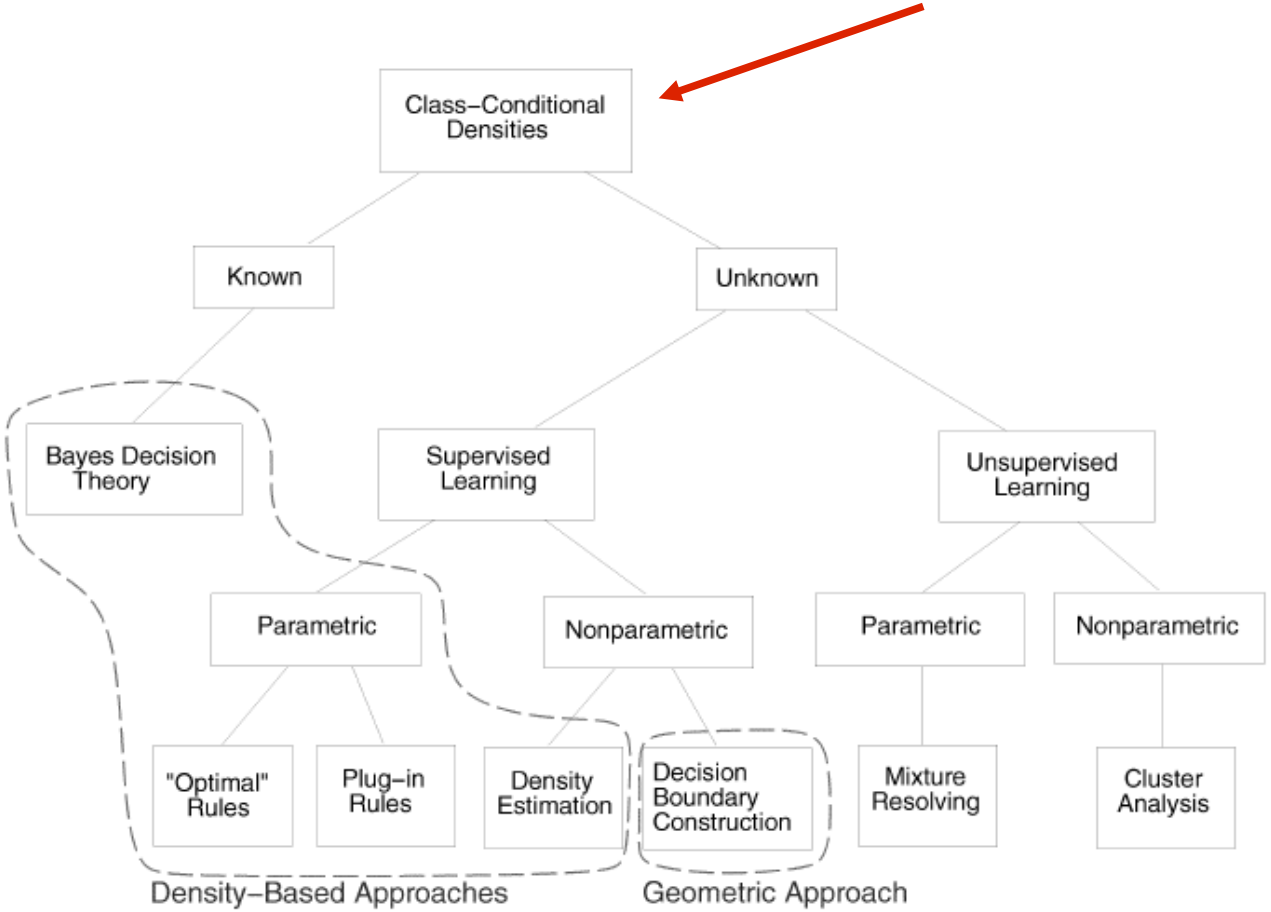
Classificação

Dado um conjunto de elementos, queremos separá-los por classes.

Algumas questões:

- você sabe quantas classes?
- você conhece exemplos dessas classes?

Técnicas de Classificação



[JAIN et al, 2000]

O que são essas densidades condicionais da classe?

- O que quer dizer densidade neste contexto?
- Função densidade de probabilidade

Revisão (rápida) de probabilidade e estatística

- X é uma variável aleatória (cujo valor pode ser visto como o resultado de um experimento) que assume valores sobre Ω (espaço amostral)
- Função de probabilidade: $p(x) = P(X = x)$
- Função de distribuição de probabilidade: $F(x) = P(X <= x)$

Revisão (rápida) de probabilidade e estatística

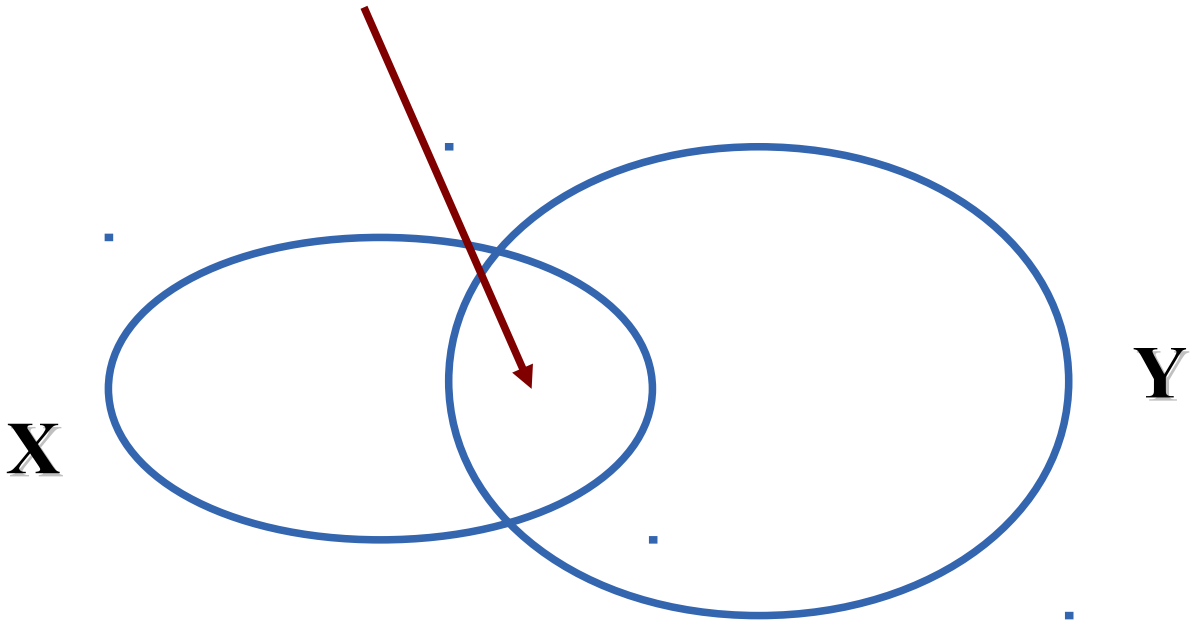
- Se X é discreta (assume apenas valores inteiros):
  - $F(x) = \sum_{i:x_i<=x} P(X=x_i)$
- Se X é contínua (pode assumir valores reais não inteiros):
  - $F(X) = \int_{-\infty}^{x} f(t)\,dt$, sendo f a função densidade de probabilidade de X
  - Ou seja, a densidade só vira probabilidade quando integrado em um intervalo

Revisão (rápida) de probabilidade e estatística

- Exemplos de distribuições discretas:
  - Bernoulli (lançamento de uma moeda)
  - Binomial (vários Bernoulli)
  - Multinomial (generalização da Binomial para k possíveis resultados)
  - Poisson (vista como uma Binomial de eventos raros)
- Exemplos de distribuições contínuas
  - Normal
  - Exponencial (intervalo entre dois sucessos consecutivos de uma Poisson)
  - Gumbel (EVD)

Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y

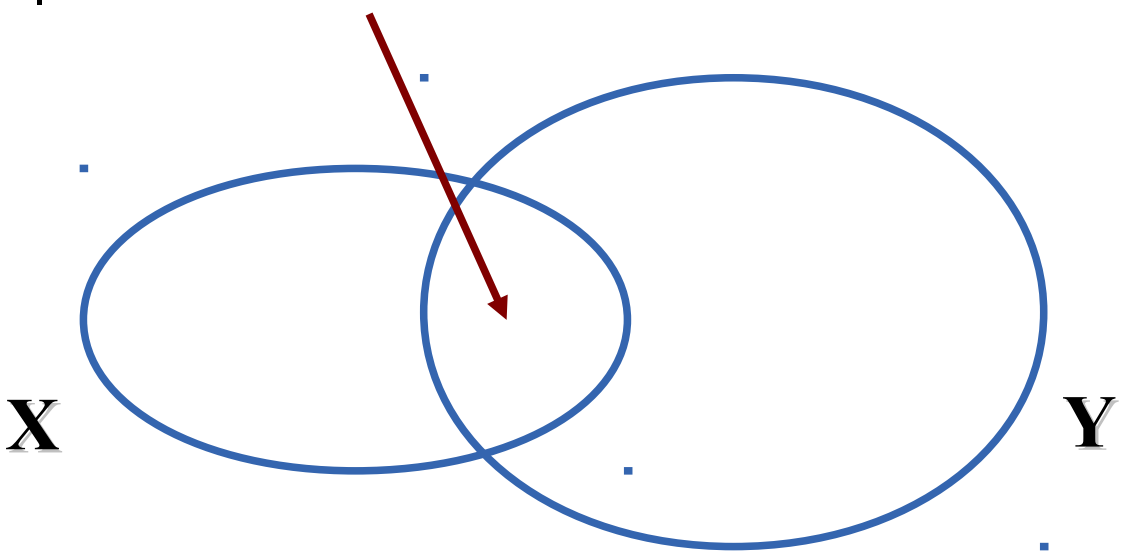


Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (a ∈ X | a ∈ Y) = P ([a ∈ X] ∩ [a ∈ Y]) / P (a ∈ Y)

Revisão (rápida) de probabilidade e estatística

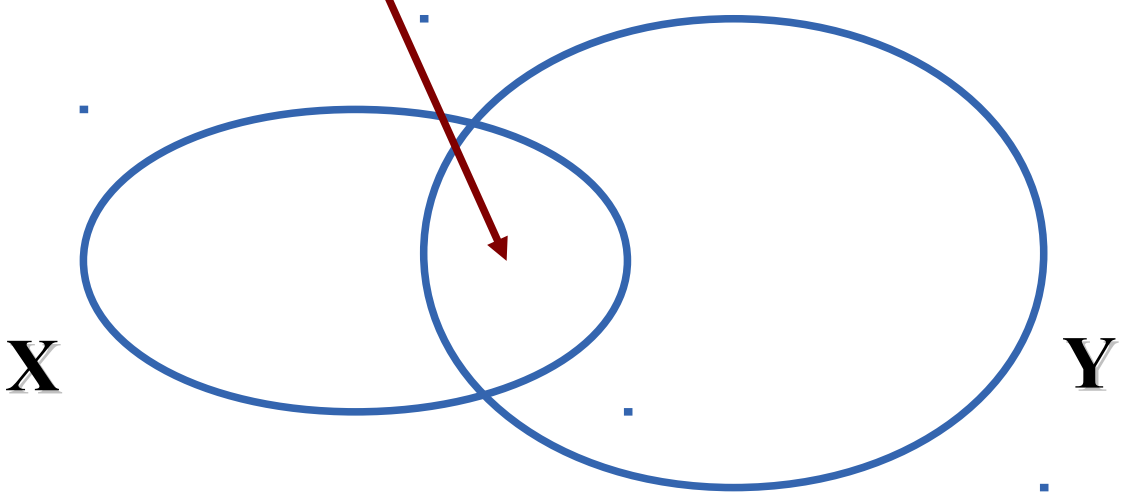
- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (X | Y) = P (X ∩ Y) / P(Y)



Revisão (rápida) de probabilidade e estatística

- Probabilidades condicionais:
  - P (X | Y) = probabilidade do evento X dado que ocorreu o evento Y
  - P (X | Y) = P (X , Y) / P(Y)

Distribuição conjunta



Probabilidades condicionais em classificação

- X pode ser um vetor aleatório das características dos elementos que eu observo (que eu quero classificar)
- Y pode ser as classes possíveis: P(Y =yi) é a probabilidade de um elemento ser da classe yi
- X possui uma distribuição dentro de Ω (geral, “no mundo”), e possivelmente uma distribuição diferente “dentro de” (condicionada a) uma dada classe

Ex: * probabilidade de um aluno da USP ser mulher ou homem

* probabilidade de um aluno da USP do curso de SI ser mulher ou homem

* probabilidade de um aluno da USP do curso de pedagogia ser mulher ou homem

Como isso pode nos ajudar na classificação?

Probabilidades condicionais em classificação

- X pode ser um vetor aleatório das características dos elementos que eu observo (que eu quero classificar)
- Y pode ser as classes possíveis: P(Y =yi) é a probabilidade de um elemento ser da classe yi
- X possui uma distribuição dentro de Ω (geral, “no mundo”), e possivelmente uma distribuição diferente “dentro de” (condicionada a) uma dada classe

Ex: * probabilidade de um aluno da USP ser mulher ou homem

* probabilidade de um aluno da USP do curso de SI ser mulher ou homem

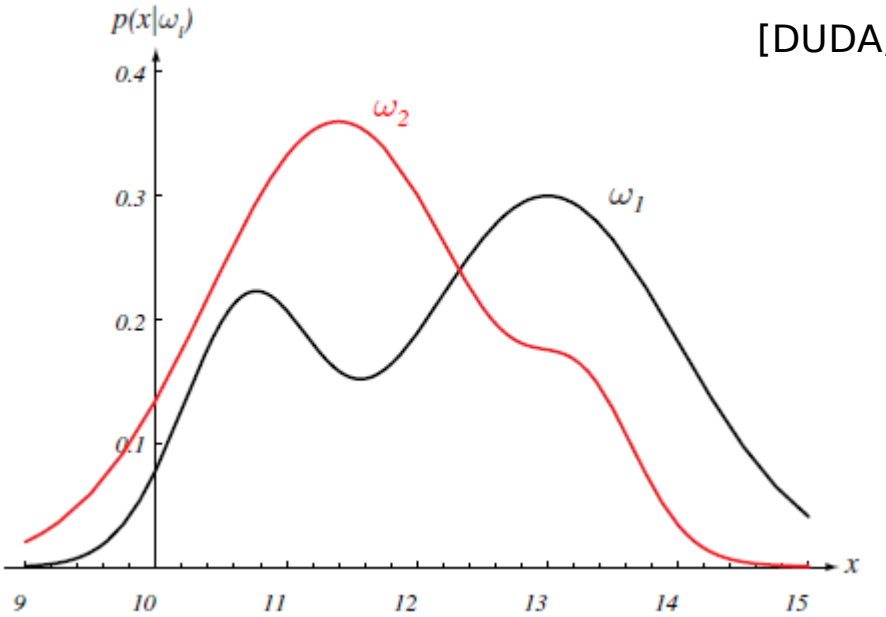
* probabilidade de um aluno da USP do curso de pedagogia ser mulher ou homem

Como isso pode nos ajudar na classificação?

Saber o sexo do aluno pode nos ajudar a prever se ele é de SI ou pedagogia

Probabilidades condicionais

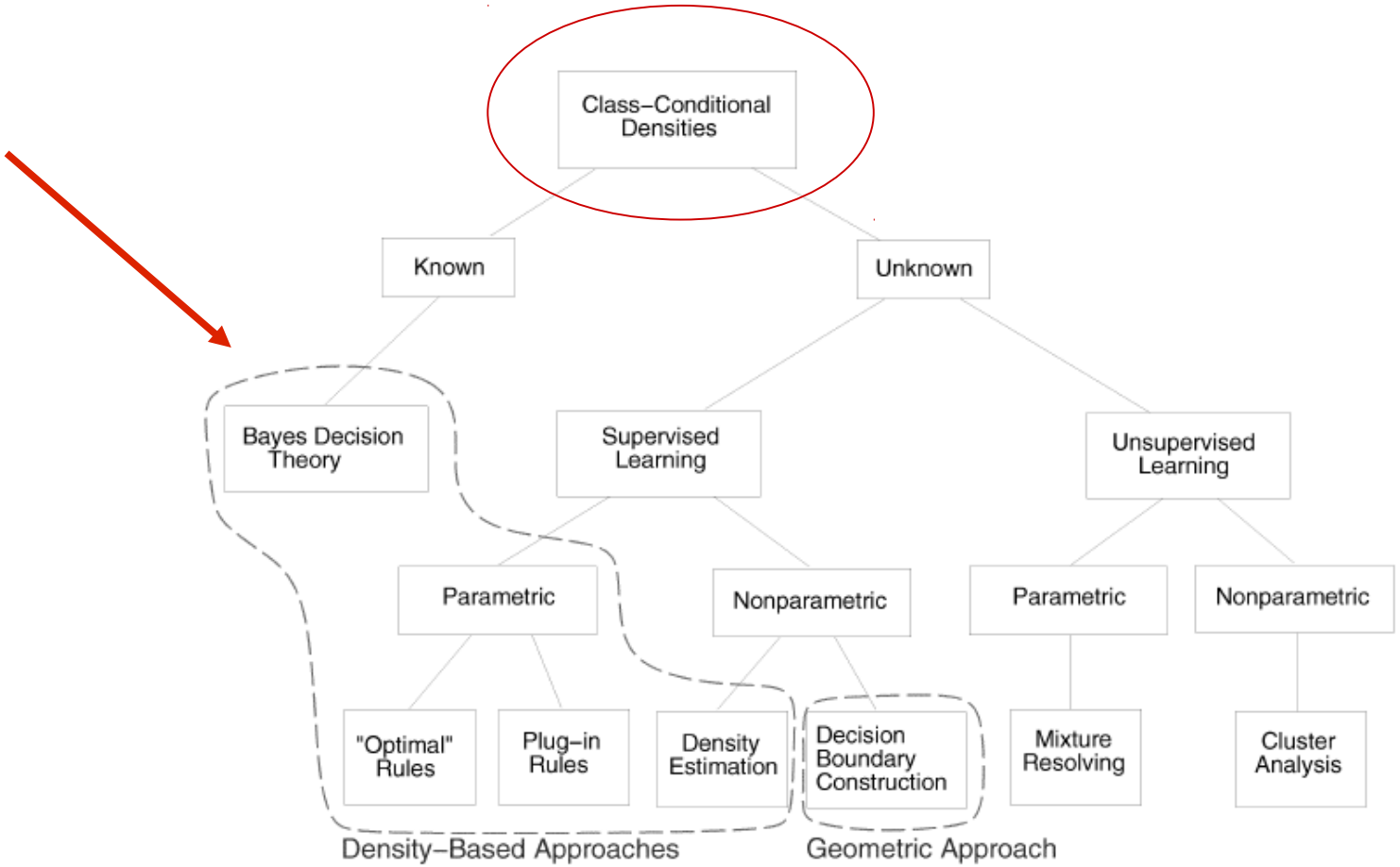
- Ex: densidade condicional da luminosidade dos peixes para as classes robalo e salmão
- P(x|c) : densidade de probabilidade condicional (à classe)
- Ex:



[DUDA, HART & STORK, 2001]

luminosidade

Técnicas de Classificação



[JAIN et al, 2000]

Teoria da Decisão Bayesiana

- Abordagem estatística para o problema de classificação / tomada de decisão
- O classificador é ótimo dentre todas as opções (minimiza o erro), porém...
- É necessário conhecer todas as probabilidades envolvidas e relevantes

Teorema de Bayes

$$P(X \mid Y) = P(X,Y) / P(Y)$$
$$P(Y \mid X) = P(Y,X) / P(X)$$

- $P(X,Y) = P(X|Y)\ P(Y)$
- $P(Y,X) = P(Y|X)\ P(X)$

- Como $P(X,Y) = P(Y,X)$,

- $P(Y|X) = P(X|Y)\ P(Y)/P(X)$

Veremos que na verdade isso tem muito mais significado...

Teoria da Decisão Bayesiana

- Classes envolvidas (ou estados da natureza)
  - Ex: robalo (c1) e salmão (c2)

- Os objetos pertencem a uma dessas classes
- Dado um objeto, ele pode pertencer aleatoriamente a c1 ou a c2
- Logo, c1 e c2 podem ser vistos como valores de uma variável aleatória c

- P(c) : probabilidade *a priori*

Regra de Decisão

- Você tem que chutar qual será o próximo peixe
- Você apenas tem a informação das probabilidades *a priori* de c1 e c2
- Qual peixe você chutaria (a priori, isto é, sem olhar o peixe)?

Regra de Decisão

- Você tem que chutar qual será o próximo peixe
- Você apenas tem a informação das probabilidades *a priori* de c1 e c2
- Qual peixe você chutaria?
  - c1 se P(c1) > P(c2)
  - c2 caso contrário

- Isto é uma regra de decisão

Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?

Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?
- Escolheríamos sempre a mesma classe, mesmo sabendo que há 2...
- Erro?

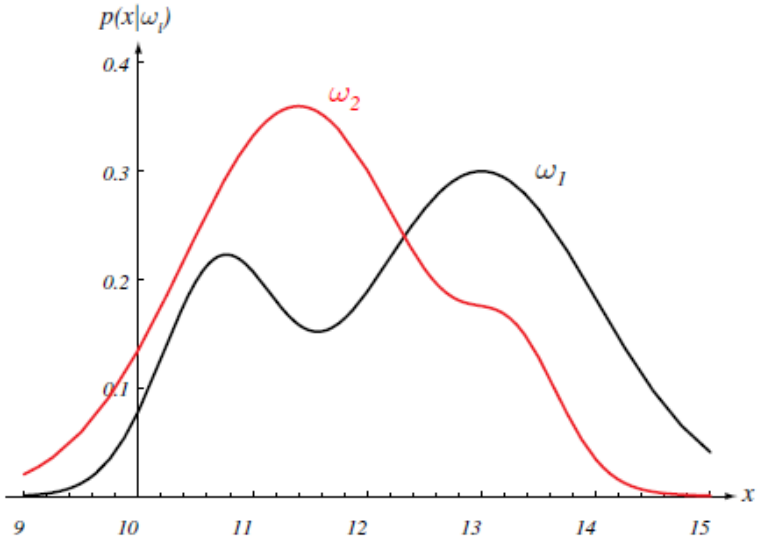
Teoria da Decisão Bayesiana

- E para os 100 próximos peixes?
- Escolheríamos sempre a mesma classe, mesmo sabendo que há 2...
- Erro?
  - Número de peixes classificados errados
  - Aproximadamente P(c2) (se esta *priori* estiver correta)

Probabilidades condicionais

- Se eu sei mais acerca dos peixes (ex: luminosidade, tamanho, etc) – características, para CADA classe.
- P(x|c) : densidade de probabilidade condicional (à classe)
- Ex:

[DUDA, HART & STORK, 2001]



luminosidade

Revisitando Bayes

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- $P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

$P(x) = \sum j\ P(x, cj) = \sum j\ P(x|cj)\ P(cj)$

Revisitando Bayes

- Se você:
  - conhece as prioris e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- $P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

$P(x) = \sum j\ P(x, cj) = \sum j\ P(x|cj)\ P(cj)$

- posteriori = verossimilhança * *priori* / evidência

Revisitando Bayes

- Se você:
  - conhece as prioris e as condicionais
  - vê um novo peixe (x)

  como você classificaria?
- Pela fórmula de Bayes:
- P(ci | x) = P(x|ci) P(ci) / P(x)

Atualização da sua crença (*priori*) após ver os dados = *posteriori*

$$P(x) = \sum_j P(x, c_j) = \sum_j P(x|c_j) P(c_j)$$

- *posteriori* = verossimilhança * *priori* / evidência

Nova regra de decisão

- Se você:
  - conhece as *prioris* e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

Nova regra de decisão

- Se você:
  - conhece as prioris e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

Nova regra de decisão

- Se você:
  - conhece as prioris e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

- Erro: P(erro | x) =
  - P(c1| x) se decidirmos por c2
  - P(c2| x) se decidirmos por c1

Nova regra de decisão

- Se você:
  - conhece as prioris e as condicionais
  - vê um novo peixe (x)

  como você classificaria?

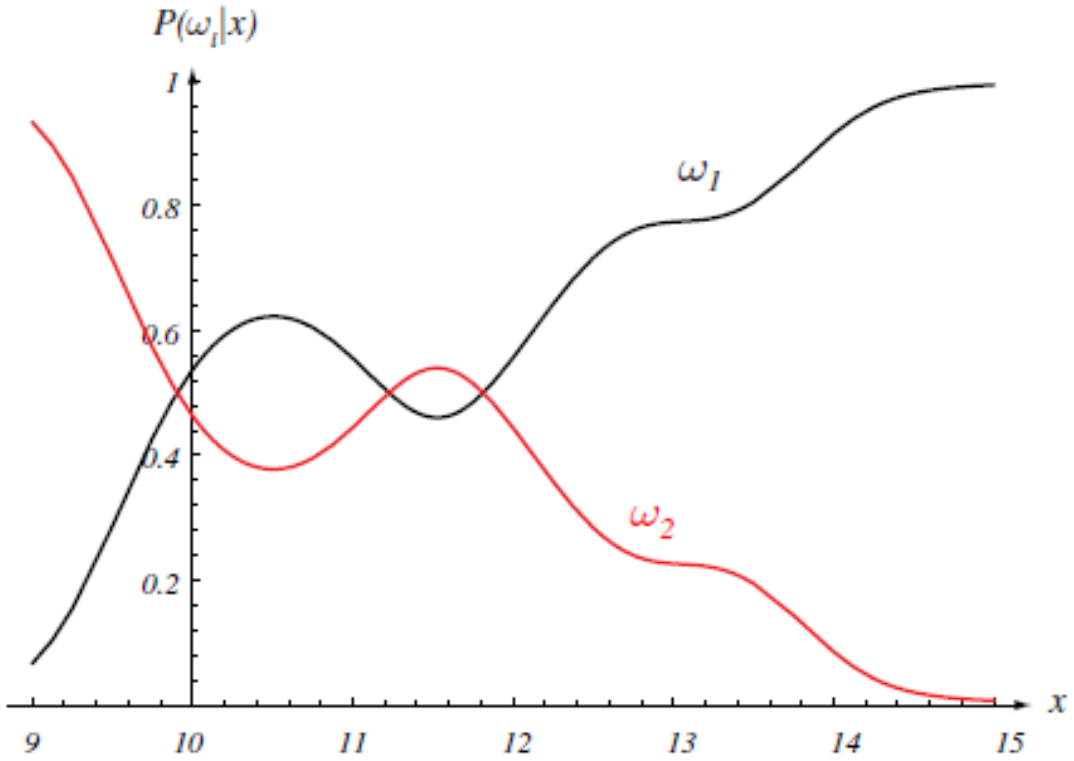
- c1 se P(c1|x) > P(c2|x)
- c2 caso contrário

Veremos que a Regra de decisão Bayesiana tem erro mínimo

- Erro: P(erro | x) =
  - P(c1| x) se decidirmos por c2
  - P(c2| x) se decidirmos por c1

Posterioris

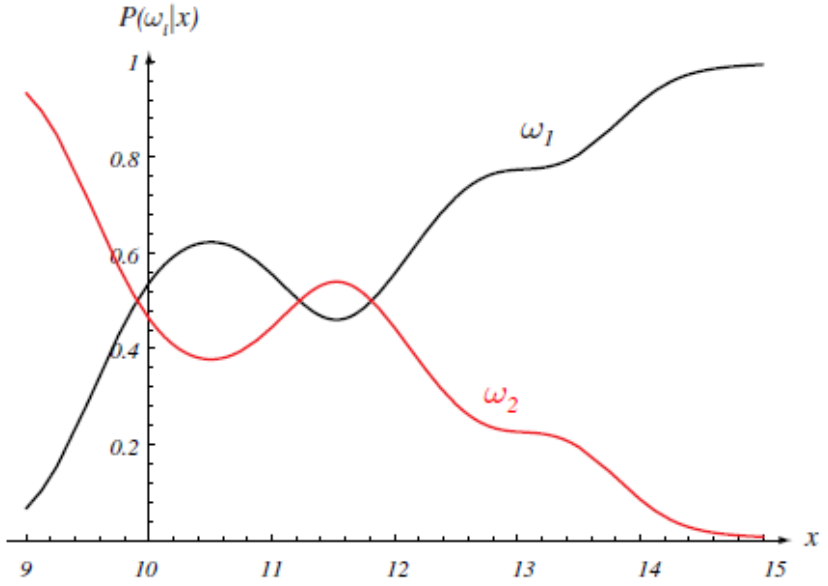
Ex: $P(\omega_1 = 2/3)$ e $P(\omega_2 = 1/3)$



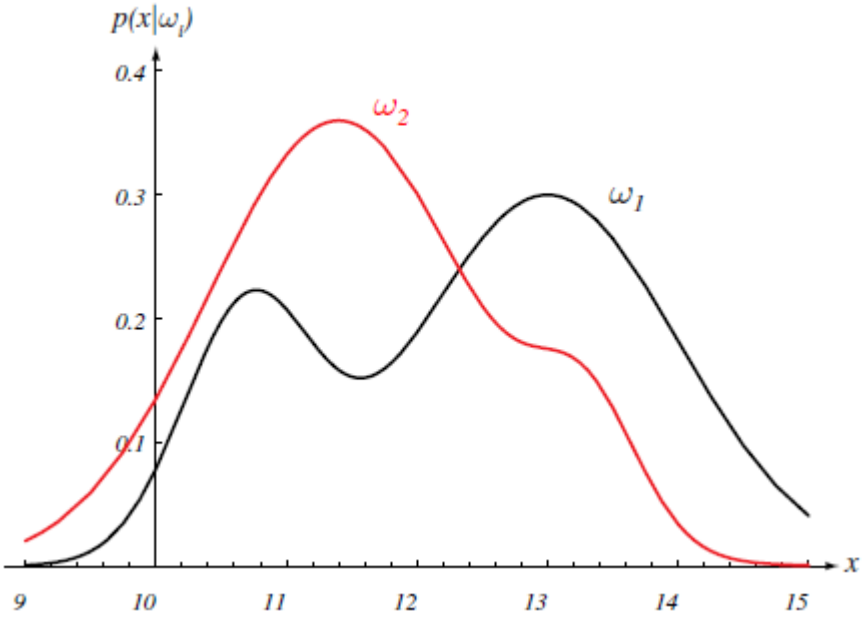
[DUDA, HART & STORK, 2001]

Ex: $P(\omega_1 = 2/3)$ e $P(\omega_2 = 1/3)$

Posterioris



Condicionais (verossimilhanças)



Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

- Se P(c1) = P(c2)?

Regra de decisão bayesiana

$P(ci \mid x) = P(x|ci)\ P(ci) / P(x)$

- Preciso calcular P(x) para decidir?
- Não! Como p(x) é uma constante (em relação às classes), a regra é:
  - c1 se $P(x|c1)\ P(c1) > P(x|c2)\ P(c2)$
  - c2 caso contrário

- Se P(c1) = P(c2), a decisão se resume à verossimilhança

Generalizando...

- **x** pode ser um vetor de características
- Pode haver várias classes (c1, ..., ck)
  - Decido pela classe i se P(ci | **x**) > P(cj| **x**), i ≠ j
- Posso realizar *a* ações (ao invés de apenas classificar), α1, ..., αa
  - Por ex, ficar indeciso e não fazer nada
- Cada ação αi tem um custo para cada cj: função perda λ(αi | cj)
- Perda esperada de se tomar uma ação αi ao observar x:

$$R(\alpha i|\mathbf{x}) = \sum_{j=1..c} \lambda(\alpha i \mid cj)\, P(cj|\mathbf{x})$$

Risco condicional

Regra de Decisão Bayesiana

- $\alpha(\mathbf{x})$ é uma função que escolhe αi com base em $\mathbf{x}$
- Risco total: $R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, dx$
- Como minimizamos R?

Regra de Decisão Bayesiana

- $\alpha(\mathbf{x})$ é uma função que escolhe αi com base em $\mathbf{x}$
- Risco total: $R = \int R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, dx$
- Como minimizamos R?
- $\alpha(\mathbf{x})$ deve sempre escolher a ação αi que tem o menor risco condicional $R(\alpha i|\mathbf{x})$
  - Regra de decisão Bayesiana
- Este R mínimo (R*) é o risco de Bayes

  melhor resultado possível

Vejamos o exemplo de classificação binária

Classificação Binária

- $\lambda_{ij} = \lambda(\alpha i \mid cj)$, αi = classificar como classe ci
- $R(\alpha 1|\mathbf{x}) = \lambda_{11} P(c1|\mathbf{x}) + \lambda_{12} P(c2|\mathbf{x})$
- $R(\alpha 2|\mathbf{x}) = \lambda_{21} P(c1|\mathbf{x}) + \lambda_{22} P(c2|\mathbf{x})$

Tomamos a ação $\alpha 1$ (isto é, classificamos como pertencente à classe c1) se $R(\alpha 2|\mathbf{x}) > R(\alpha 1|\mathbf{x})$

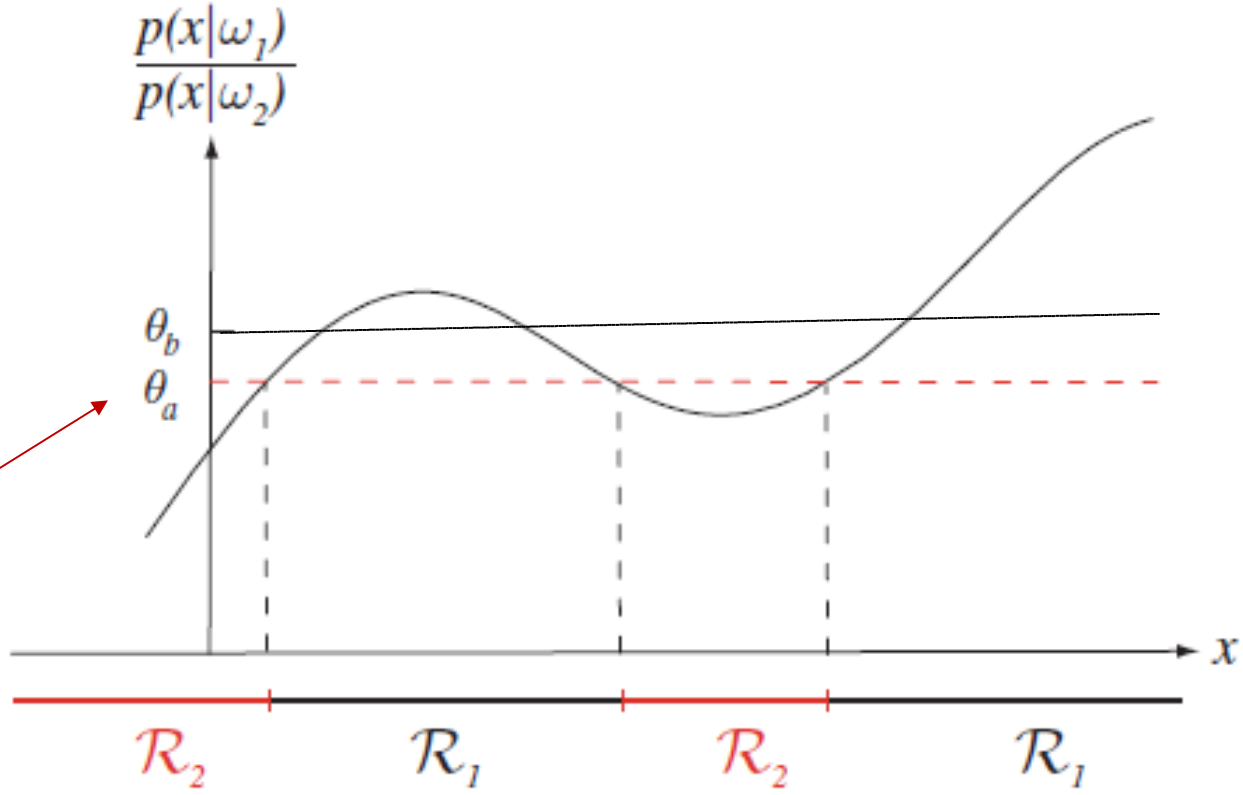
- $\lambda_{21} P(c1|\mathbf{x}) + \lambda_{22} P(c2|\mathbf{x}) > \lambda_{11} P(c1|\mathbf{x}) + \lambda_{12} P(c2|\mathbf{x})$
- $(\lambda_{21} - \lambda_{11}) P(c1|\mathbf{x}) > (\lambda_{12} - \lambda_{22}) P(c2|\mathbf{x})$
- $(\lambda_{21} - \lambda_{11}) P(\mathbf{x}|c1) P(c1) > (\lambda_{12} - \lambda_{22}) P(\mathbf{x}|c2) P(c2)$
- $$\frac{P(\mathbf{x}|c1)}{P(\mathbf{x}|c2)} > \frac{(\lambda_{12} - \lambda_{22}) P(c2)}{(\lambda_{21} - \lambda_{11}) P(c1)} = \theta$$ Likehood ratio (Razão de verossimilhança)

Função perda zero-um

Função perda zero-um:

$$\lambda(\alpha i \mid cj) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

Razão de verossimilhança



Perda 0-1

prioris iguais

Ri: Região de decisão por i

[DUDA, HART & STORK, 2001]

Razão de verossimilhança

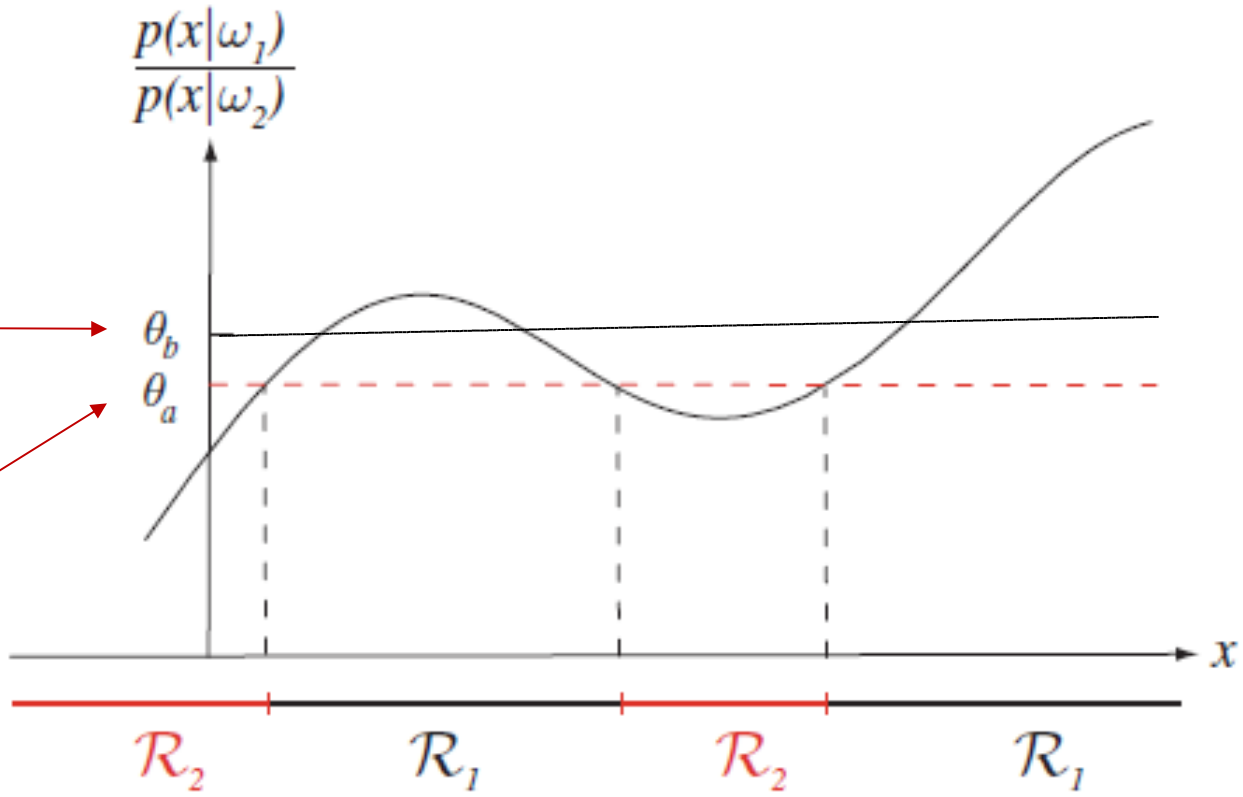
$$\frac{P(\mathbf{x}|c1)}{P(\mathbf{x}|c2)} > \frac{(\lambda_{12} - \lambda_{22})\, P(c2)}{(\lambda_{21} - \lambda_{11})\, P(c1)}$$

$\lambda_{12} > \lambda_{21}$

(R1 se torna menor)

Perda 0-1

prioris iguais



Ri: Região de decisão por i

[DUDA, HART & STORK, 2001]

Classificação de múltiplas classes – taxa de erro mínima

- Função perda zero-um:

$$\lambda(\alpha i \mid cj) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$$

- $R(\alpha i|\mathbf{x}) = \sum_{j=1..c} \lambda(\alpha i \mid cj)\, P(cj|\mathbf{x})$

$$= \sum_{j \neq i} P(cj|\mathbf{x})$$

$$= 1 - P(ci|\mathbf{x})$$

- Decida por ci se $P(ci|\mathbf{x}) > P(cj|\mathbf{x})$, para todo $j \neq i$

Resumindo

- Para uma função perda genérica, não há classificador melhor que o teste bayesiano da razão de verossimilhança no sentido de minimizar o custo
- Se a sua função perda é a zero-um, isso se resumo a escolher a classe com maior *posteriori*

Problema

- Normalmente não conhecemos as probabilidades condicionais P(. |ci)
- Neste caso temos que estimá-las a partir de dados conhecidos, o que pode ser complexo
- Mesmo conhecendo todas as probabilidades necessárias, o teste pode ser complexo em termos de tempo e memória

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  -
-

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
-

Alternativas

- Nestes casos, temos que optar por um classificador mais simples
- Por ex: escolher uma família de densidades conhecida, ou seja, com uma forma matemática conhecida e um número finito de parâmetros. Daí basta estimar os parâmetros (ex. Normal, Poisson, ...)
  - CLASSIFICADOR PARAMÉTRICO
- Quando não é possível escolher uma forma matemática então opta-se por um CLASSIFICADOR NÃO PARAMÉTRICO

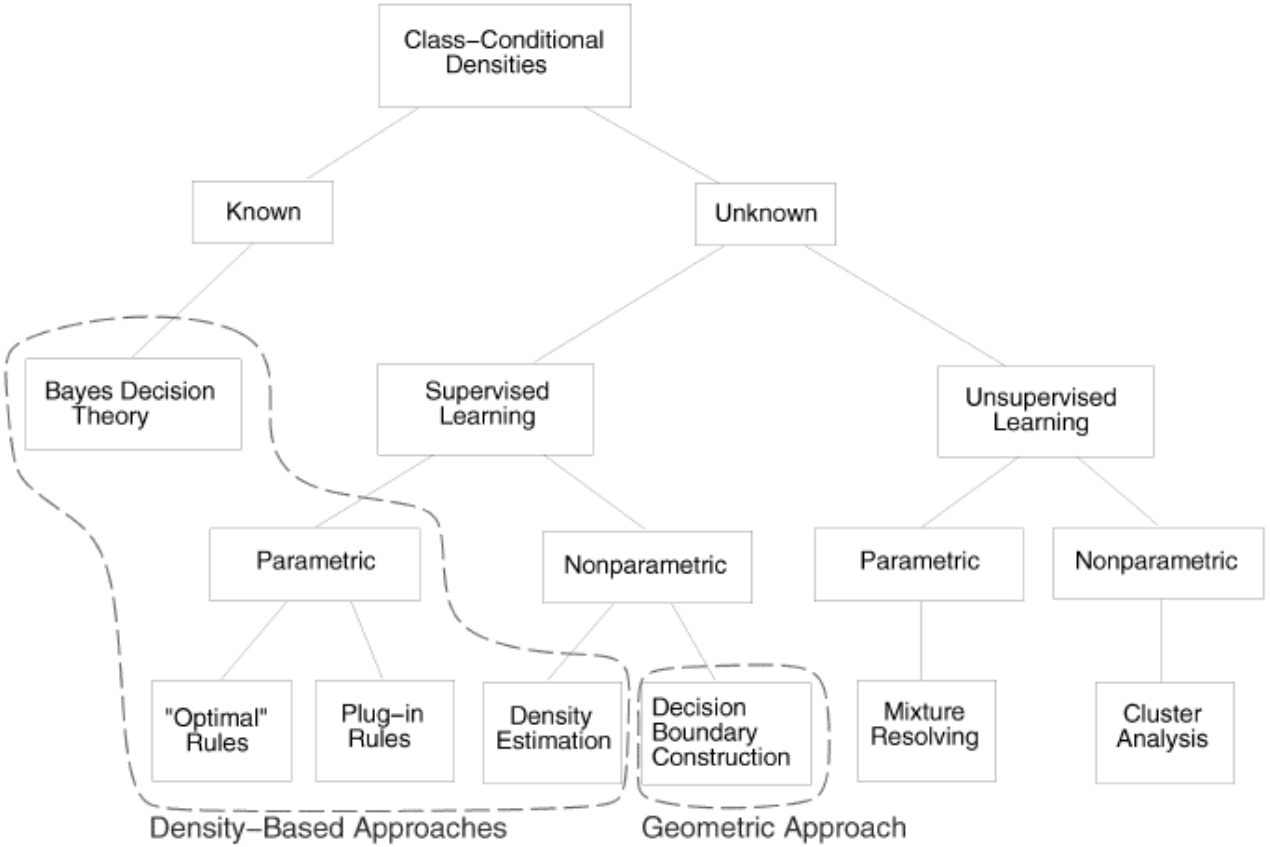
Métodos de construção de um classificador

- Aprendizado supervisionado: o modelo é aprendido a partir de amostras rotuladas (com sua classificação)
  - Amostra de treinamento:

  X = {(x,c)| x é um exemplo e c é sua classe}

- Aprendizado não supervisionado: a classificação é feita a partir de dados não rotulados

Métodos de Classificação



[JAIN et al, 2000]

Referências

- DUDA, R.; HART, P.; STORK, D**. Pattern Classification**. John Willey, 2001 (Cap. 2.1 a 2.3)

- JAIN, A.K.; DUIN, R.P.W.; MAO, J. Statistical Pattern Recognition : A Review. **IEEE Transactions on Pattern Analysis and Machine Intelligence**, v. 22, n. 1, p. 4-37, 2000 (seções 2 e 7)