SIN5007

Reconhecimento de Padrões

Professora:
Ariane Machado Lima

Apresentando a disciplina

- Número de créditos: 12
  - Aulas teóricas: 3
  - Horas de estudo: 9

Objetivos

- Apresentar os principais métodos utilizados para análise de dados, compreensão de fenômenos e tomada de decisão em problemas de reconhecimento de padrões. Mais especificamente, serão introduzidos fundamentos e algoritmos para modelagem e classificação de dados, seleção e extração de características, juntamente com os métodos apropriados para avaliação de desempenho dos modelos obtidos.
- As similaridades e diferenças entre os métodos apresentados e suas principais vantagens e desvantagens serão também abordadas ao longo deste curso.

Justificativa

- Técnicas de reconhecimento de padrões podem ser utilizadas em várias aplicações de Inteligência Artificial, como na modelagem de fenômenos naturais e classificação de objetos físicos ou padrões abstratos multidimensionais. A compreensão de padrões em dados de alta dimensão é ainda um componente importante para a descoberta de conhecimento em problemas complexos.

Conteúdo

- Introdução a problemas de reconhecimento de padrões;
- Extração de características e redução de dimensionalidade. Análise de Componentes Principais
- Técnicas para seleção de características;
- Redes Neurais Perceptron Multi-Camadas;
- Máquinas de Vetores Suporte;
- Redes Bayesianas;
- Árvores de Decisão;
- Métodos sintáticos;
- Classificação multiclasse e multirrótulo;
- Classificadores não supervisionados: Técnicas de agrupamento de dados. Aplicações e avaliação de desempenho;

Avaliação

- Chamada oral ;-) – valendo nota! (MO)

  Toda aula envolvendo o entendimento do que já foi ministrado nas aulas anteriore
- 2 Provas Teóricas: MP = média aritmética simples de ambas
- Trabalhos
  - Trabalhos em grupo (4 alunos)
  - Entregas parciais quase todas as semanas: montagem incremental de uma apresentação
  - Apenas um grupo será sorteado por aula para apresentar (vale nota de chamada oral! - MO)
  - Audiência deve perguntar, questionar, sugerir
  - Uma entrega final (TF) – seminário (todos os grupos)

Média Final (MF): (2*MP + 2*MT + MO)/5

Sobre os trabalhos

- Grupos de 4 (um único dataset)
- Recomenda-se que um aluno proponha a utilização de um dataset no qual já esteja trabalhando ou pretende trabalhar
- O aluno nestas condições deve preencher a planilha Google colocando seu nome como Nome1 e assinalar na coluna “Sugere dataset?” como “Sim”. Se todos os 13 já tiverem um Nome1 como sim, preencher Nome2, etc, e então o grupo decide qual dataset utilizar: https://docs.google.com/spreadsheets/d/199Sdnsee3PykqUW8pw3XaQ-vqH4en-eGrt8MqAQ7oOo/edit?usp=sharing
- Os demais alunos devem preencher a mesma planilha com “Sugere dataset?” como “Não”.

Moodle

- Usaremos o sistema Moodle para
  - Repositório do material das aulas
  - Troca de mensagens
  - Submissão de trabalhos

- Basta se cadastrar em https://edisciplinas.usp.br que a disciplina aparece para você (inscrição automática)

Do que vamos precisar?

- Principalmente conceitos de:
  - Cálculo
  - Probabilidade e Estatística
  - Álgebra linear

Do que vamos precisar?

- Principalmente conceitos de:
  - Cálculo
  - Probabilidade e Estatística
  - Álgebra linear

- Não se apavorem...

Do que vamos precisar?

- Principalmente conceitos de:
  - Cálculo
  - Probabilidade e Estatística
  - Álgebra linear

- Não se apavorem...
- ... mas sejam responsáveis (FAÇAM AS REVISÕES NECESSÁRIAS!)

Bibliografia Base

- Duda R.; Hart, P.; Stork, D. **Pattern Classification and Scene Analysis**. John Wiley, 2001.
- Fukunaga, K. **Introduction to Statistical Pattern Recognition**. 2nd Edition. New York: Academic Press, 1990.
- Haykin, S. **Neural Networks: A Comprehensive Foundation**. 2nd Edition. Prentice-Hall, 1999.
- Ripley, B. **Pattern Recognition and Neural Networks.** Cambridge: Cambridge University Press, 1996.

Bibliografia complementar

- Outras dadas ao final de cada aula

R – Livros na EACH

- DALGAAR, P. **Introductory statistics with R**. Springer

- EVERITT, B. ; HOTHORN, T. **A handbook of statistical analysis using R**. Ed. Chapman & Hall/CRC

O que eu espero de vocês?

O que eu espero de vocês?

ESPERO QUE VOCÊS SE DIVIRTAM!

O que eu espero de vocês?

ESPERO QUE VOCÊS SE DIVIRTAM!

E participem das aulas

O que eu espero de vocês?

ESPERO QUE VOCÊS SE DIVIRTAM!

E participem das aulas

SEJAM BEM-VINDOS!

Tema 1

Introdução a Problemas de Reconhecimento de Padrões e Conceitos Básicos

Profa Ariane Machado Lima

O que é reconhecimento ?

Exemplo..

O que é reconhecimento ?

- Quando você procura algo para sentar...

O que é reconhecimento ?

- Quando você procura algo para sentar...
- Voz de um amigo na multidão

O que é reconhecimento ?

- Quando você procura algo para sentar...
- Voz de um amigo na multidão
- Uma velha amiga com aparência diferente

O que é reconhecimento ?

- Quando você procura algo para sentar...
- Voz de um amigo na multidão
- Uma velha amiga com aparência diferente
- Várias letras manuscritas de um alfabeto conhecido (nem todas...)

O que é reconhecimento ?

- Quando você procura algo para sentar...
- Voz de um amigo na multidão
- Uma velha amiga com aparência diferente
- Várias letras manuscritas de um alfabeto conhecido (nem todas...)
- Emoções pela expressão facial

O que é reconhecimento ?

- Quando você procura algo para sentar...
- Voz de um amigo na multidão
- Uma velha amiga com aparência diferente
- Várias letras manuscritas de um alfabeto conhecido (nem todas...)
- Emoções pela expressão facial
- Impressões digitais

O que é reconhecimento ?

- Re-conhecer
- Perceber que você já conhece, mesmo que nunca tenha visto IGUAL (e conseguir associar a quê)
- Realizar uma inferência indutiva com base em conceitos aprendidos no passado (informação *a priori*)

O que é um padrão?

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*
- Um padrão representa um único objeto?

O que é um padrão?

- “A descrição de um objeto”
- “Modelo oficial de pesos e medidas; qualquer objeto que serve de modelo à feitura de outro; desenho decorativo estampado em tecido ou outra superfície” *Aurélio*
- Um padrão representa um único objeto?
- Um padrão representa uma população de objetos, um conceito, uma classe

O que é reconhecimento de padrões?

O que é reconhecimento de padrões?

- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, através da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

O que é reconhecimento de padrões?

- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, através da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

- Todos os atributos de um objeto são relevantes?

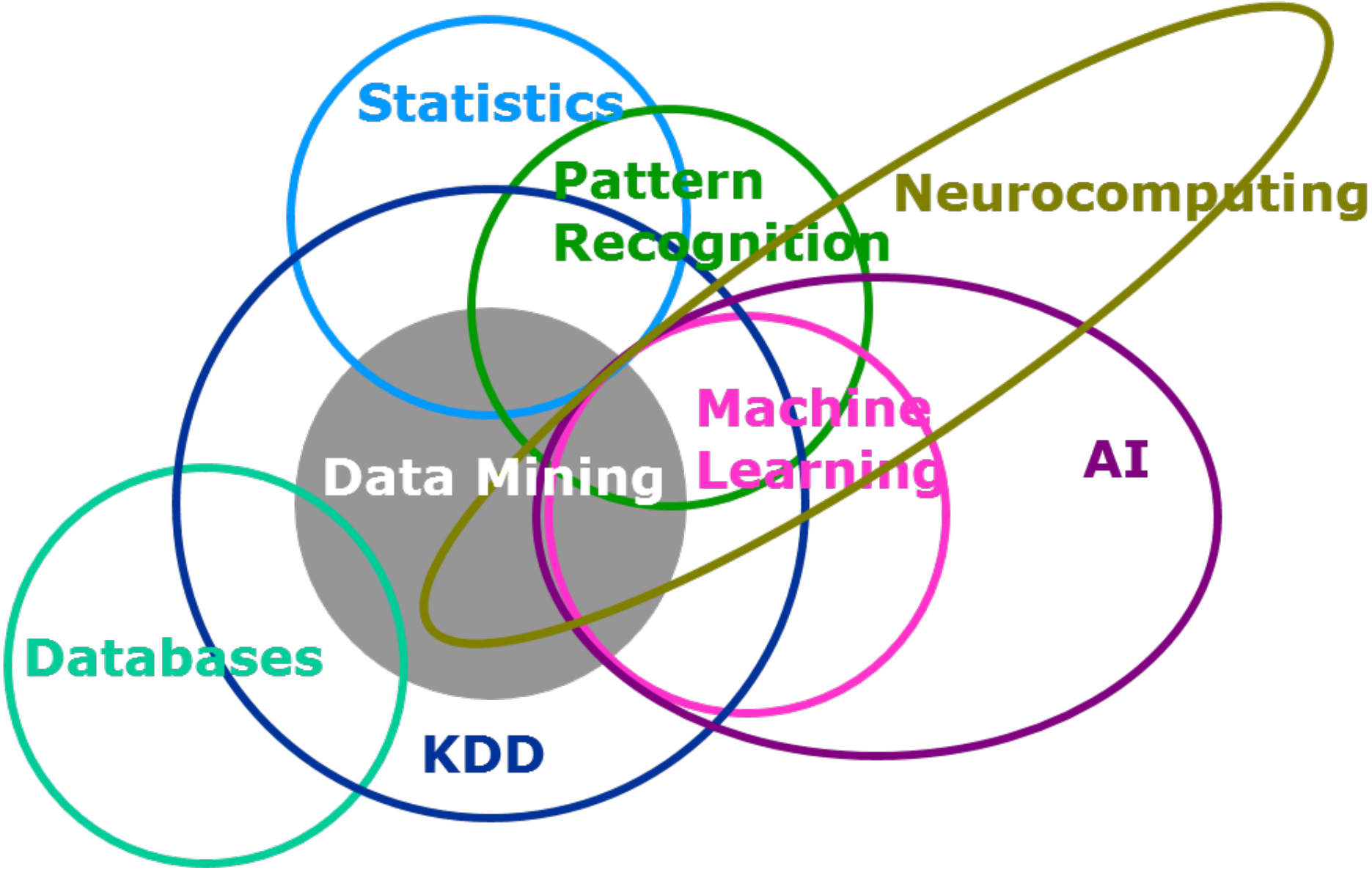
O que é reconhecimento de padrões?

- Problema de discriminar um objeto de entrada não dentre padrões individuais mas dentre populações, através da busca de atributos (mais ou menos) invariantes dentre os membros de uma população.

- Todos os atributos de um objeto são relevantes? Normalmente não...

O que é reconhecimento de padrões?

- Categorização dos dados de entrada em classes identificáveis através da extração de atributos significantes dos dados, dentre muitos outros atributos irrelevantes.



https://www.analyticsvidhya.com/blog/2015/07/difference-machine-learning-statistical-modeling/

Duas classes de problemas de Reconhecimento de Padrões

1. Estudo da capacidade de reconhecimento de padrões de seres humanos e outros seres vivos

2. Desenvolvimento de teoria e técnicas para o desenho de dispositivos capazes de executar uma dada tarefa de reconhecimento para uma dada aplicação

Duas classes de problemas de Reconhecimento de Padrões

1. Estudo da capacidade de reconhecimento de padrões de seres humanos e outros seres vivos

   Psicologia, Fisiologia, Biologia, Neurosciência

2. Desenvolvimento de teoria e técnicas para o desenho de dispositivos capazes de executar uma dada tarefa de reconhecimento para uma dada aplicação

Duas classes de problemas de Reconhecimento de Padrões

1. Estudo da capacidade de reconhecimento de padrões de seres humanos e outros seres vivos

   Psicologia, Fisiologia, Biologia, Neurosciência

2. Desenvolvimento de teoria e técnicas para o desenho de dispositivos capazes de executar uma dada tarefa de reconhecimento para uma dada aplicação

   Engenharia, Computação, Ciência da Informação

Exemplos de problemas de reconhecimento de padrões

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?

Exemplos de problemas de reconhecimento de padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?
  - Cada caracter?
  - Letras x algarismos?
  - Algarismos arábicos?
  - Alfabeto chinês, russo, árabe, etc...

O que são os padrões

- Reconhecimento de caracteres
- O que é o padrão neste caso?
  - Cada caracter?
  - Letras x algarismos?
  - Algarismos arábicos?
  - Alfabeto chinês, russo, árabe, etc...

- Padrões podem ser hierárquicos
- Os padrões dependem da aplicação
- O problema pode envolver 2 ou mais classes

O que mais?

- Teoria da Decisão
  - Tomada de decisões complexas
  - Sistemas de apoio à decisão
  - Tomar uma decisão é semelhante a fazer uma classificação
  - Podemos tomar decisões erradas
  - Há um custo para cada decisão a ser tomada
  - Queremos minimizar o custo total

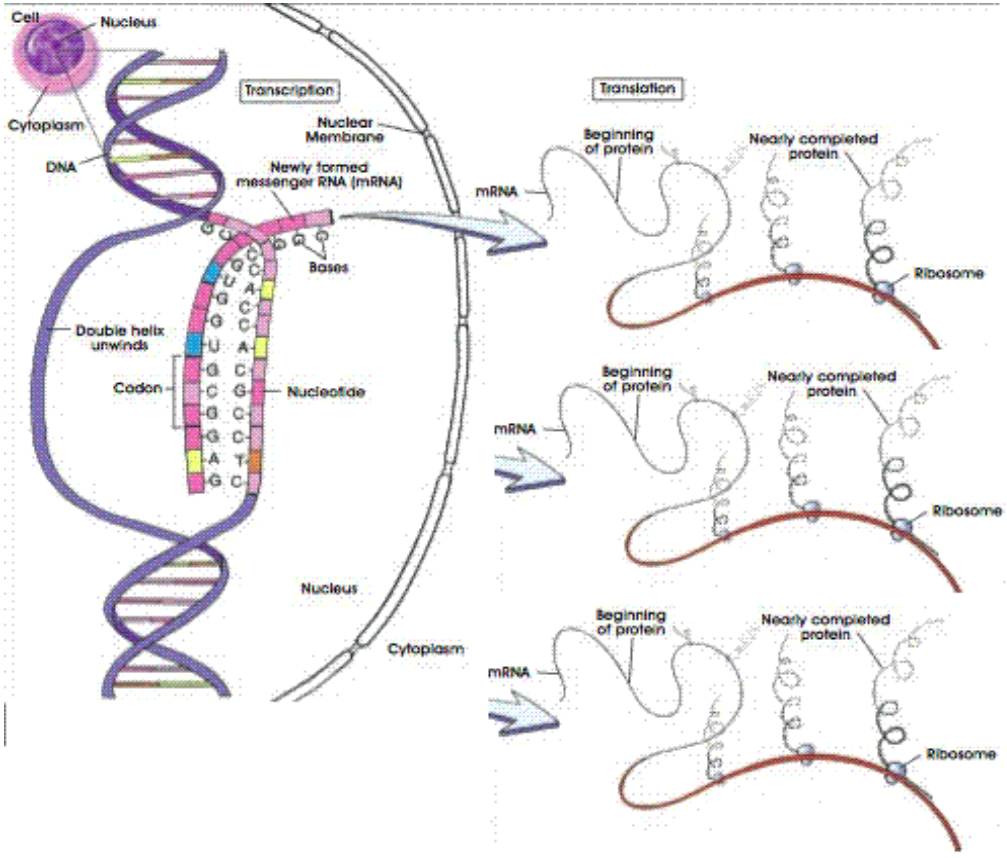
Mais exemplos de problemas de Reconhecimento de Padrões

Mais problemas

- Quais genes estão envolvidos em determinadas doenças?

Mais problemas

- Quais genes estão envolvidos em determinadas doenças?
  - Olhando a expressão

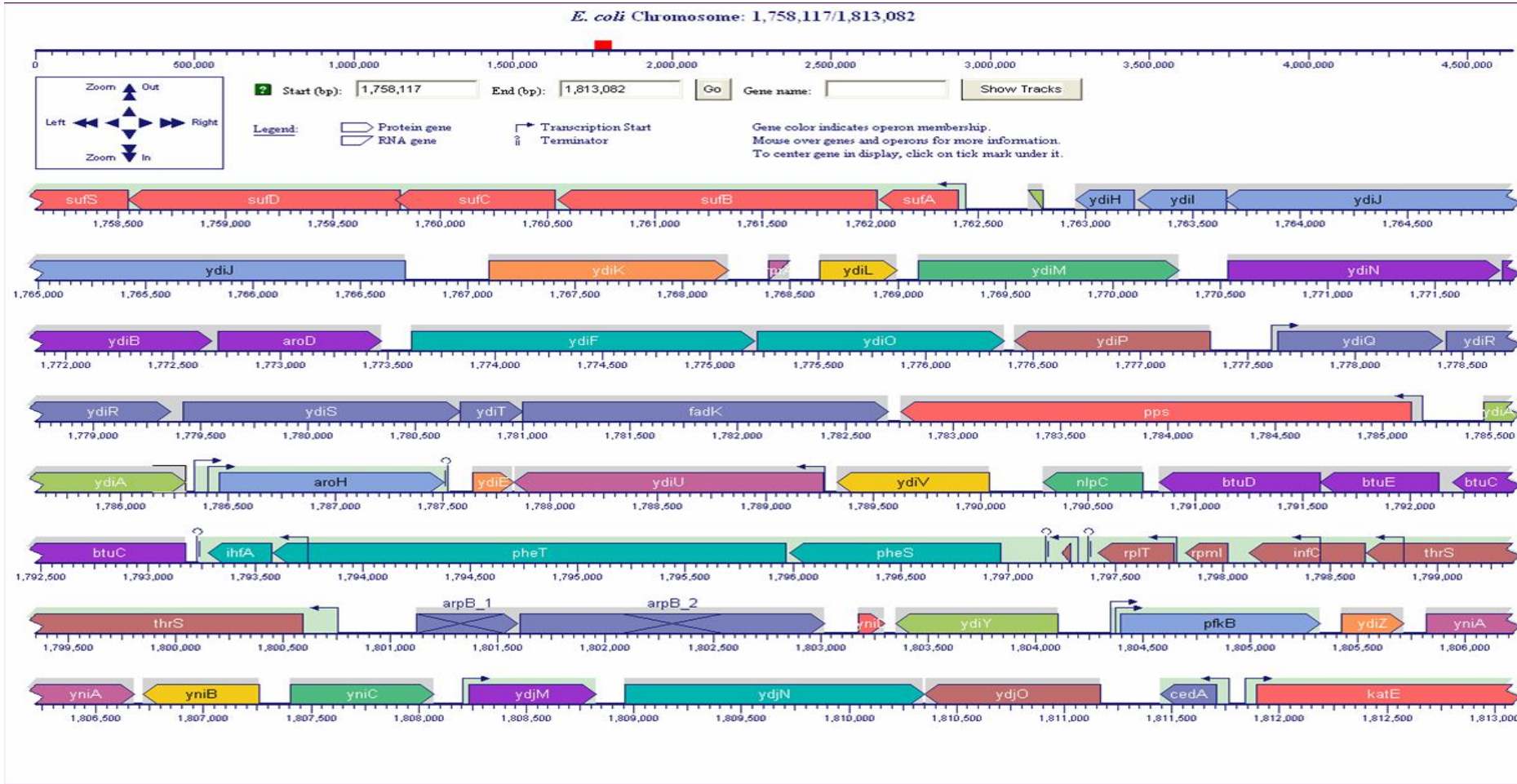


Mais problemas

- Quais genes estão envolvidos em determinadas doenças?
  - Olhando a sequência

Mais problemas

- Dado um genoma recém-sequenciado, como achar onde estão os genes?



Mais problemas

- Como classificar músicas por gênero musical?
- Como distinguir se um instrumento está afinado ou não?

Mais problemas

- Sistemas de reconhecimento de voz

Free Up Your Other Hand With
Voice Recognition TV

Mais problemas

- Sistemas de reconhecimento de imagens:
  - Detecção de “defeitos biológicos” em imagens médicas (traumatismos, aneurisma, tumor, etc)

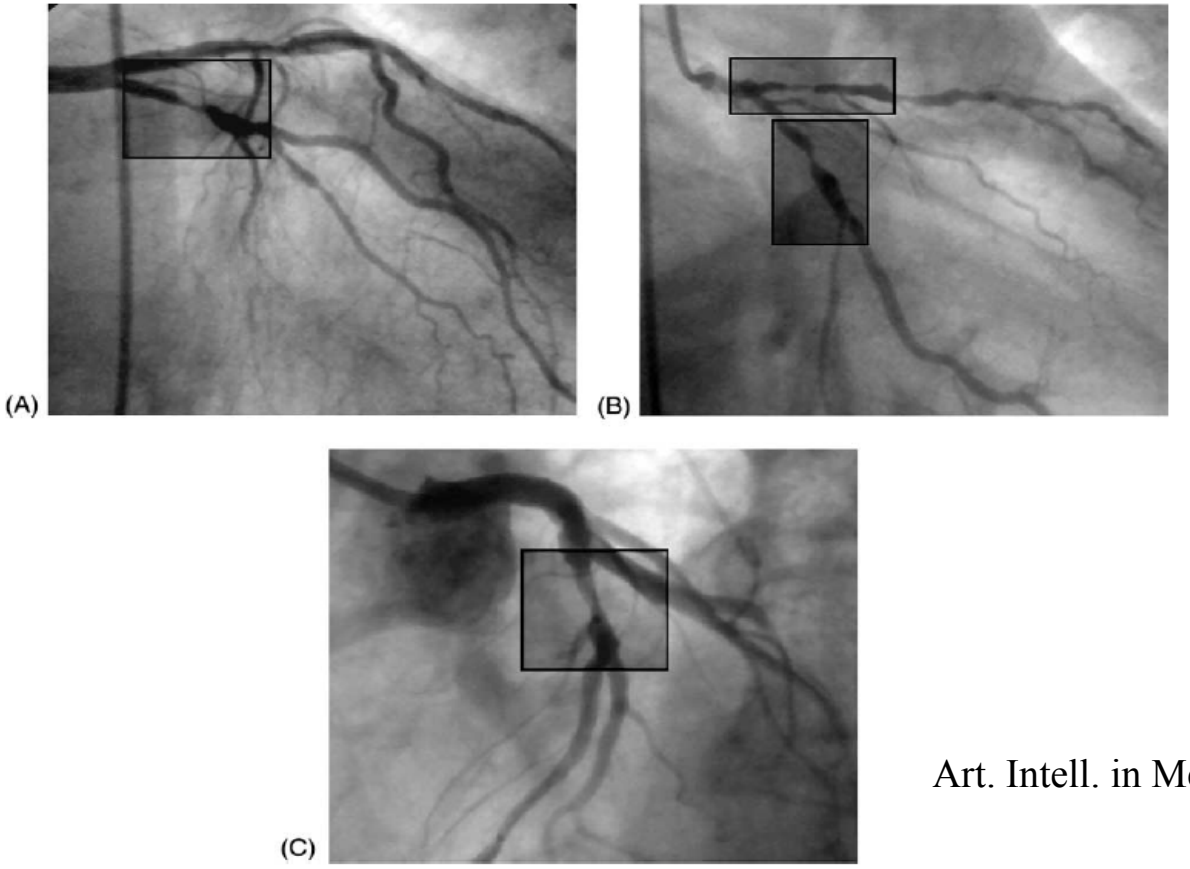
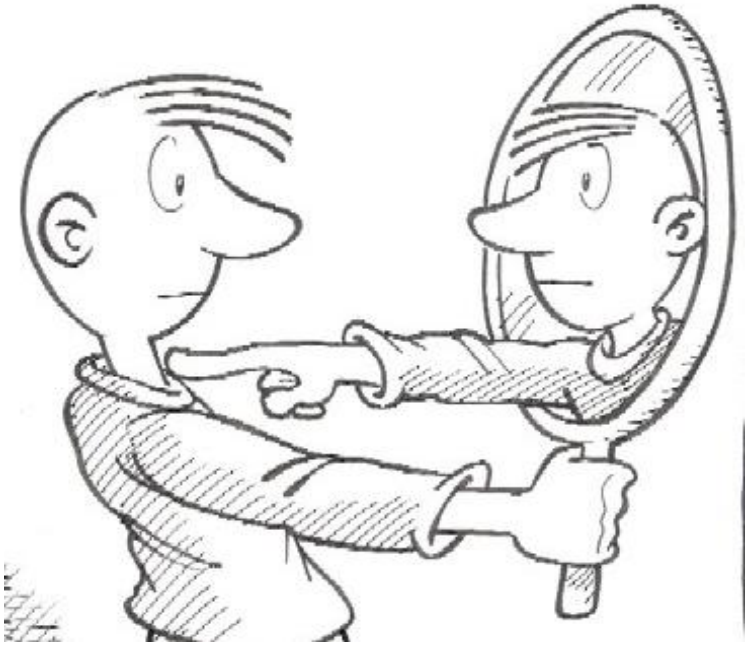


Fig. 1. (A–C) Coronographic images of coronary arteries with stenoses.

Art. Intell. in Med 26 (2002)

Mais problemas

- Sistemas de reconhecimento de imagens:
  - Reconhecimento de indivíduos em cenas criminais, retrato falado, etc



Mais problemas

- Sistemas de reconhecimento de imagens:
  - Busca por imagens

Mais problemas

- Detecção de fraude em cartões de crédito
- Comportamento de clientes (bons e maus)

R

- Ambiente para análise de dados
  - Funções estatísticas, de reconhecimento de padrões, de visualização de dados
- Código aberto
- Distribuição + pacotes (incluindo datasets)
- Materiais de apoio (manuais, tutoriais, help oline)
- http://www.R-project.org
- Publicações em

  http://CRAN.R-project.org/doc/Rnews

R - observações

- R não é orientado a objetos!
  - read.table()
  - help.search()
- Dinâmica (dados tipados dinamicamente)
- Uso do prompt ou via script
- Facilidade em baixar pacotes
- Facilidade de manipulação de dados
- Gráficos com qualidade de publicação

R – Livros na EACH

- DALGAAR, P. **Introductory statistics with R**. Springer

- EVERITT, B. ; HOTHORN, T. **A handbook of statistical analysis using R**. Ed. Chapman & Hall/CRC

R

- install.packages("HSAUR")
- library("HSAUR")
- vignette(package = "HSAUR")
- vignette("Ch_introduction_to_R", package = "HSAUR")
- edit(vignette("Ch_introduction_to_R", package = "HSAUR"))

R

1. R: instalação
   - *R Installation and Administration* no site
2. R: estudo
   - Cap. 1 de "*A handbook of statistical analysis using R*"
   - Manuais do site do R, em particular:
     - *An Introduction to R*
     - *The R language definition*
     - *R Data Import/Export*

Trabalhos em R ou **Python**

- Python:
  - Orientada a objetos, linguagem de programação de propósito mais geral que o R
  - Várias bibliotecas para vários propósitos, inclusive de reconhecimento padrões
  - Gratuita, também com uma grande comunidade
  - Acredito que seja mais rápida que R
  - Site oficial: ww.python.org

HOMEWORK!!!

- Instalação e estudo de R ou Python

Para aprendermos mais, mais exemplos...

- Robalos e salmões: precisam ser processados e embalados separadamente
- Fotografias

[DUDA, HART & STORK, 2001]

Para aprendermos mais, mais exemplos...

- Objetivo?

Para aprendermos mais, mais exemplos...

- Objetivo?
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)

Para aprendermos mais, mais exemplos...

- Objetivo?
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)
- Como podemos fazer isso?

Para aprendermos mais, mais exemplos...

- Objetivo?
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)
- Como podemos fazer isso?
  - Escolhemos algum atributo (ou característica) e vemos se ele consegue diferenciar (discriminar) robalos e salmões
  - Queremos descrever um modelo para robalos e salmões, no caso f(x) <= l é um e f(x) > l é outro, x sendo a característica e l um limiar

Para aprendermos mais, mais exemplos...

- Objetivo?
  - Dada uma foto de um peixe, classificá-lo em 1 de 2 categorias possíveis: robalo ou salmão (classificação binária)
- Como podemos fazer isso?
  - Escolhemos algum atributo (ou característica) e vemos se ele consegue diferenciar (discriminar) robalos e salmões
  - Queremos descrever um modelo para robalos e salmões, no caso f(x) <= l é um e f(x) > l é outro, x sendo a característica e l um limiar

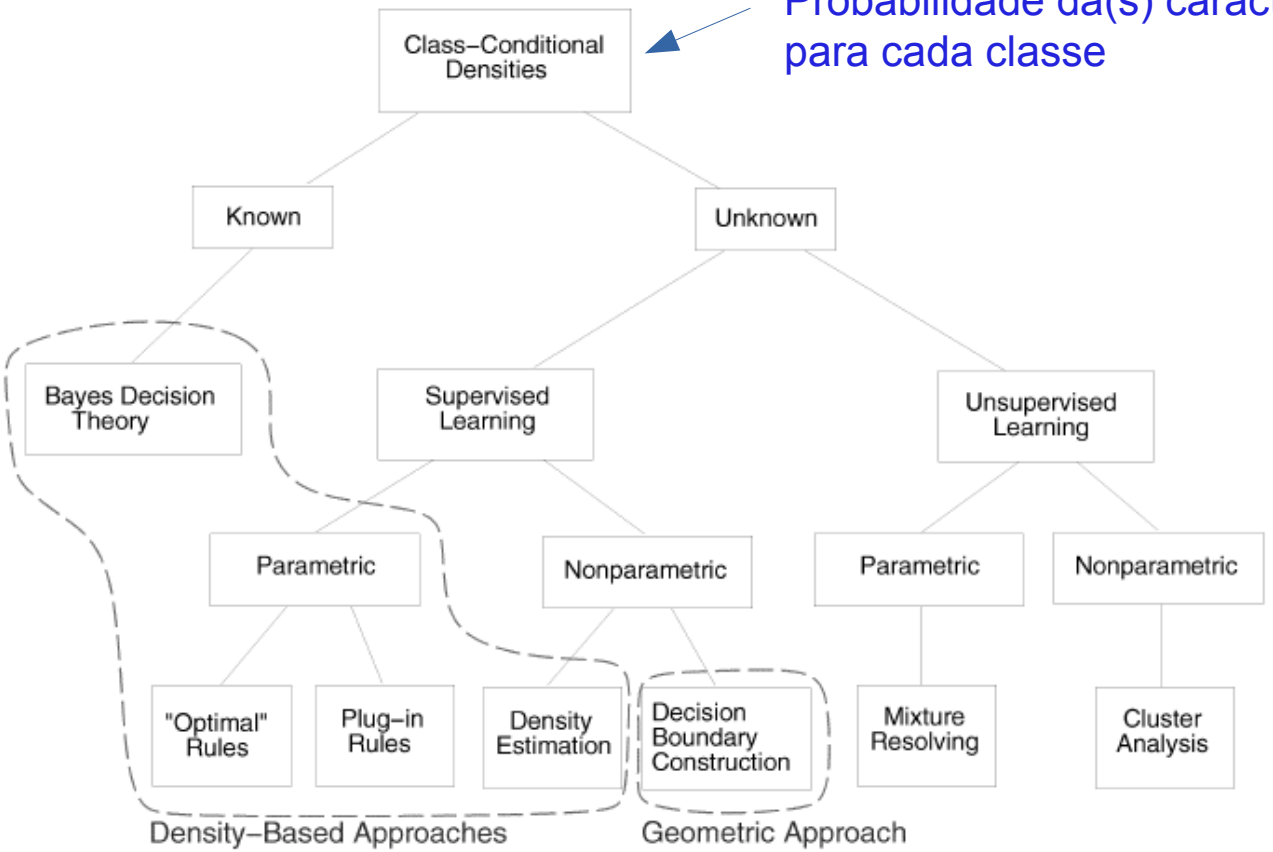
x, por exemplo, comprimento

Aprendizado computacional

- O processo de criar um modelo, uma hipótese acerca do conceito real, é conhecido como aprendizado computacional
- Como aprendemos algo?
- Como poderíamos aprender o que é robalo e o que é salmão?

Métodos de Classificação

Probabilidade da(s) característica(s) para cada classe



[JAIN et al, 2000]

Aprendizado computacional

- O processo de criar um modelo, uma hipótese acerca do conceito real, é conhecido como aprendizado computacional
- Como aprendemos algo?
- Como poderíamos aprender o que é robalo e o que é salmão?
- Uma das formas é através de exemplos
- Aprendizado supervisionado:
  - Amostra de treinamento: exemplos rotulados de cada classe
- Aprendizado não supervisionado: não tenho exemplos, mas apenas os objetos a serem classificados

Escolhemos 1 característica

- Tamanho (comprimento)
- Como medimos?

[DUDA, HART & STORK, 2001]

Escolhemos 1 característica

- Tamanho (comprimento)
- Como medimos?
- Pré-processamento (no caso segmentação para “achar” o peixe na imagem) e medição da característica (comparação com um objeto controle de tamanho conhecido)

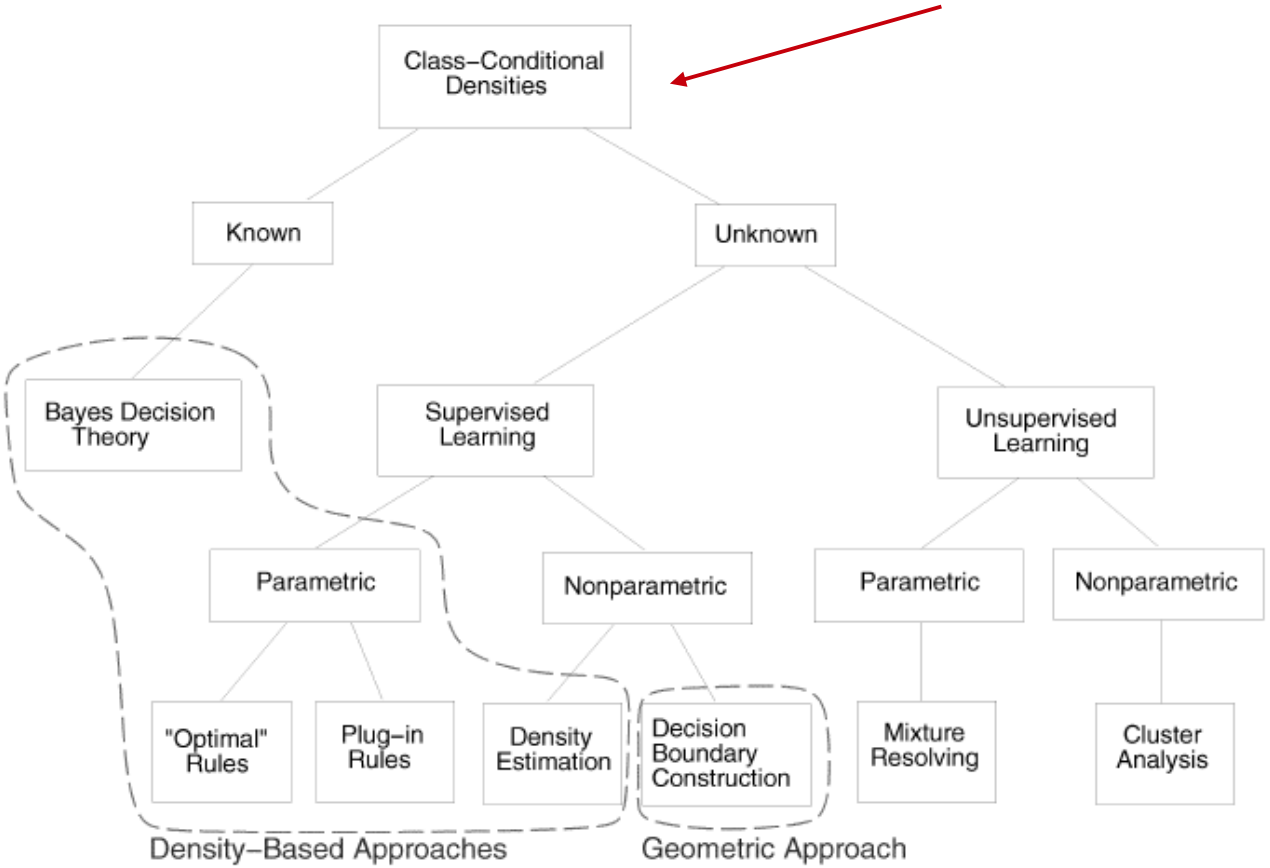
[DUDA, HART & STORK, 2001]

Tamanhos dos peixes

- Note que x (cada objeto, representado pelo seu tamanho) é uma variável aleatória

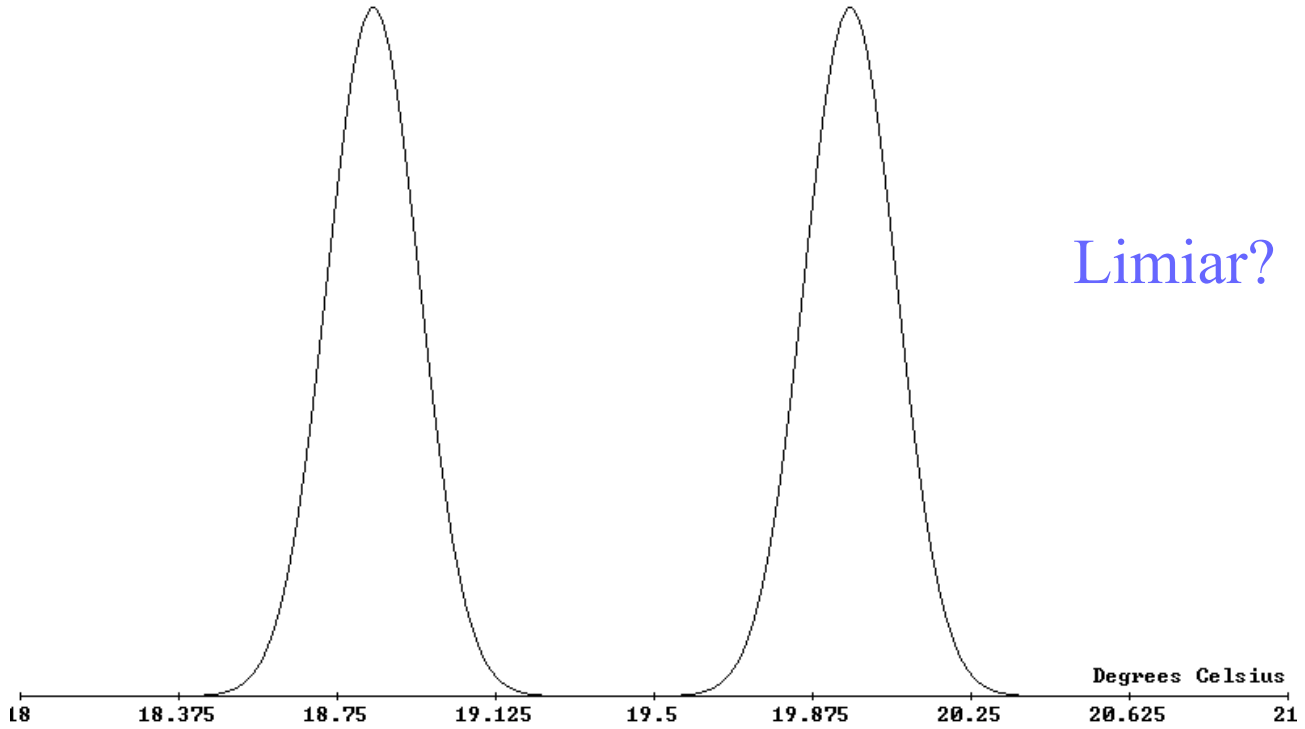
- Gostaríamos de conhecer a distribuição de x para robalos e salmões para podermos fazer a classificação (acharmos um limiar)
  - Distribuições condicionais: P(x | robalo) e P(x | salmão)

Métodos de Classificação



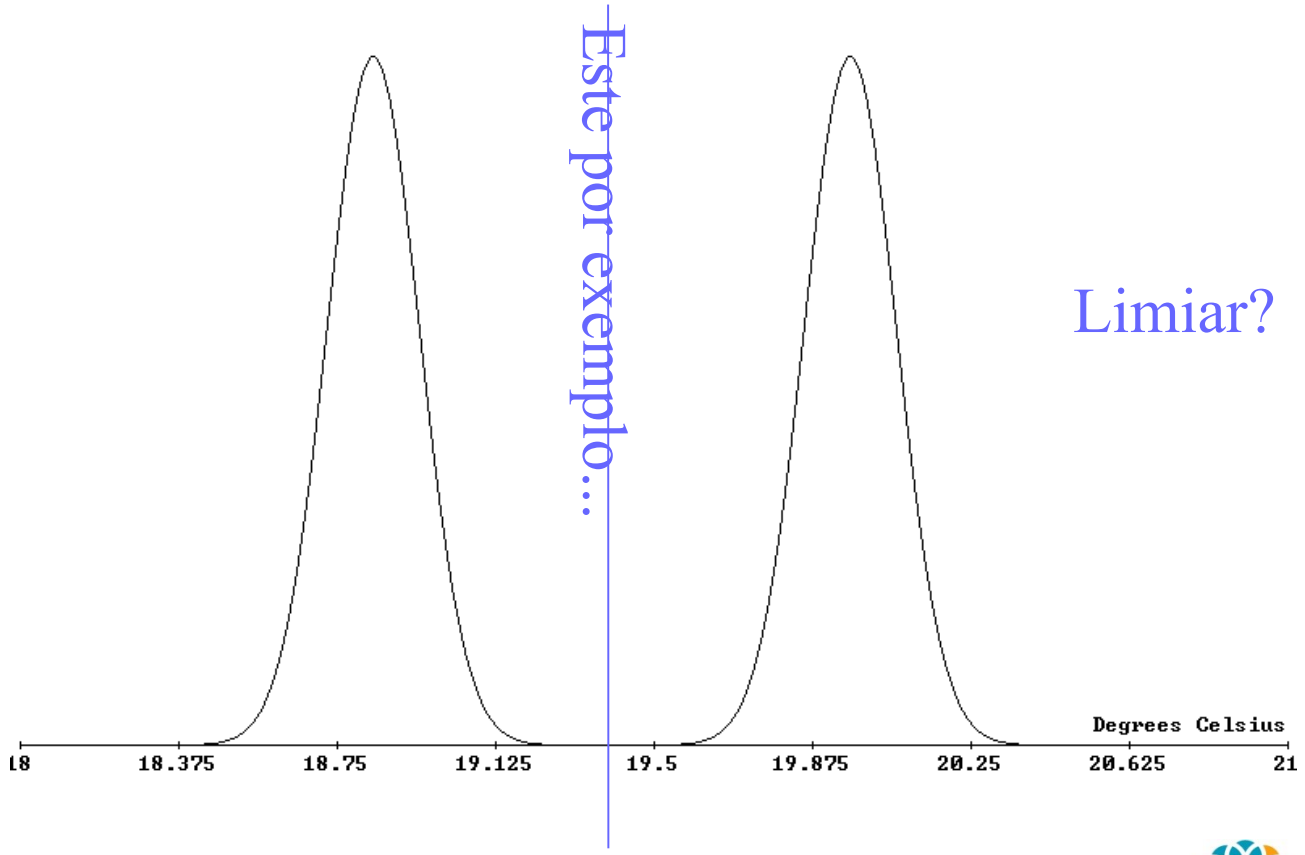
[JAIN et al, 2000]

Caso Ideal: temos as condicionais reais (ex: Deus nos deu)
Caso ideal dos sonhos! E elas são totalmente separáveis!!!!



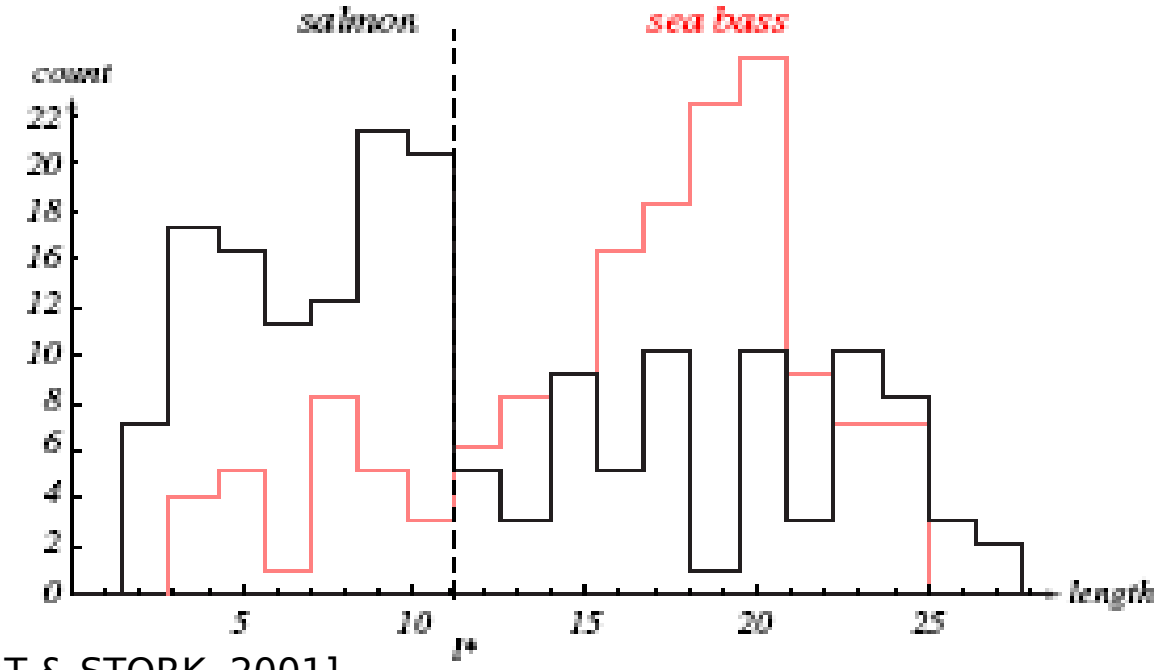
Limiar?

Caso Ideal: temos as condicionais reais (ex: Deus nos deu)
Caso ideal dos sonhos! E elas são totalmente separáveis!!!!



Se não temos as condicionais, vamos estimá-las... Veja esse histograma dos tamanhos dos peixes de exemplo...

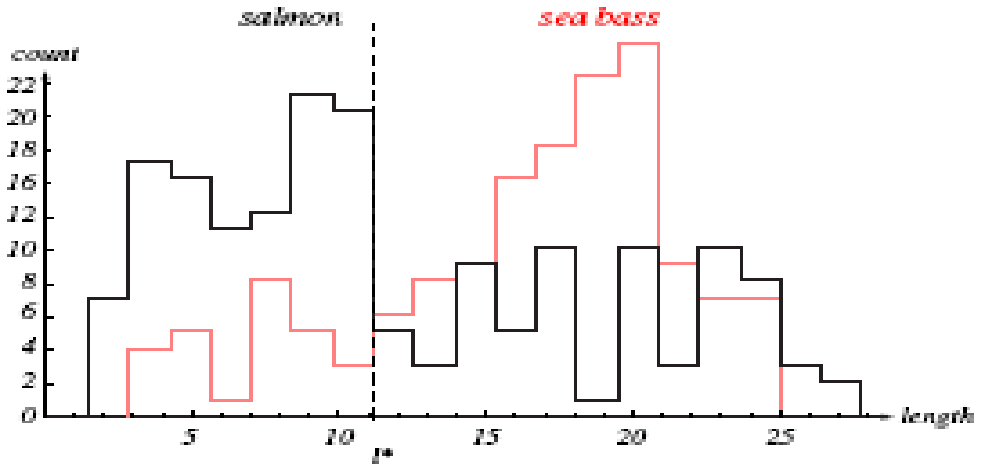
- Qual limiar escolhemos?



[DUDA, HART & STORK, 2001]

Histograma dos tamanhos dos peixes

- Qual limiar escolhemos?
- A cada limiar há um erro de classificação associado
- E os custos de uma classificação errada? São simétricos ou assimétricos
  - Erro tipo I: taxa de falso positivo (1-especificidade)
  - Erro tipo II: taxa de falso negativo (1-sensibilidade ou 1-recall)
  - teoria da decisão...



[DUDA, HART & STORK, 2001]

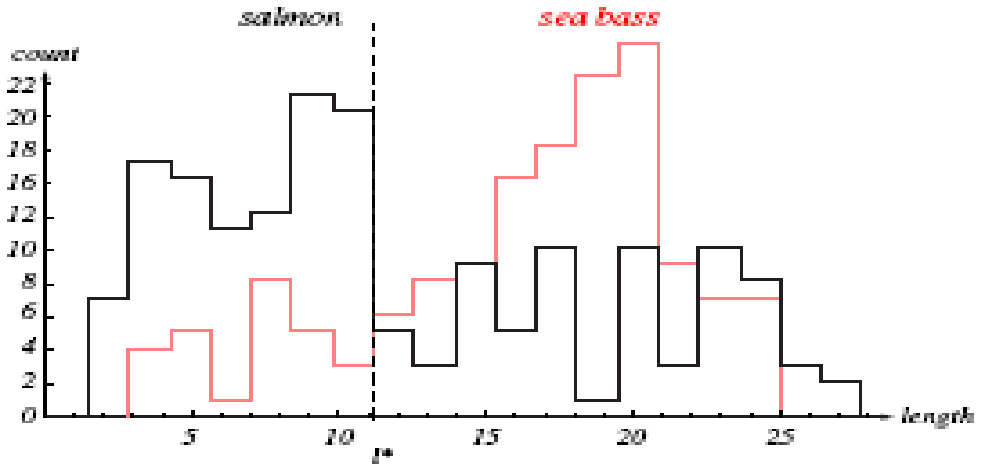
Histograma dos tamanhos dos peixes

- Qual limiar escolhemos?
- A cada limiar há um erro de classificação associado
- E os custos de uma classificação errada? São simétricos ou assimétricos
  - Erro tipo I: taxa de falso positivo (1-especificidade)
  - Erro tipo II: taxa de falso negativo (1-sensibilidade ou 1-recall)
  - teoria da decisão...

Tamanho não parece ser uma boa característica...

Vamos escolher outra!

[DUDA, HART & STORK, 2001]

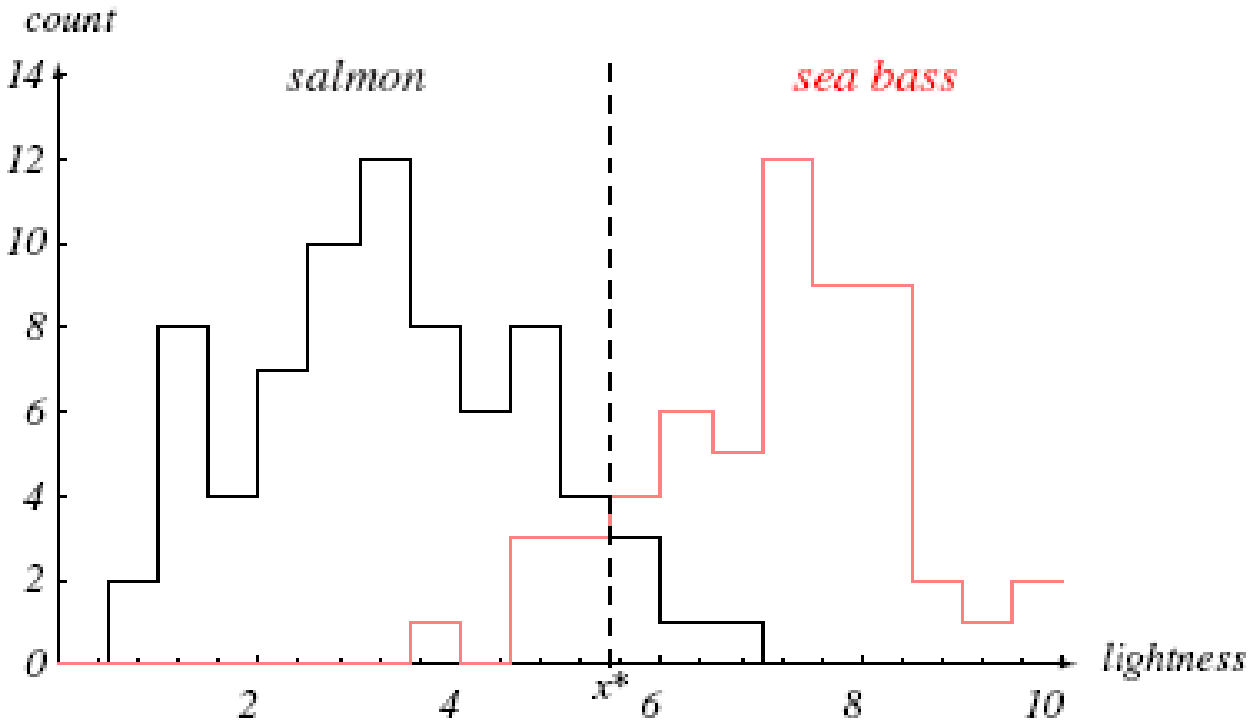


Brilho

Brilho

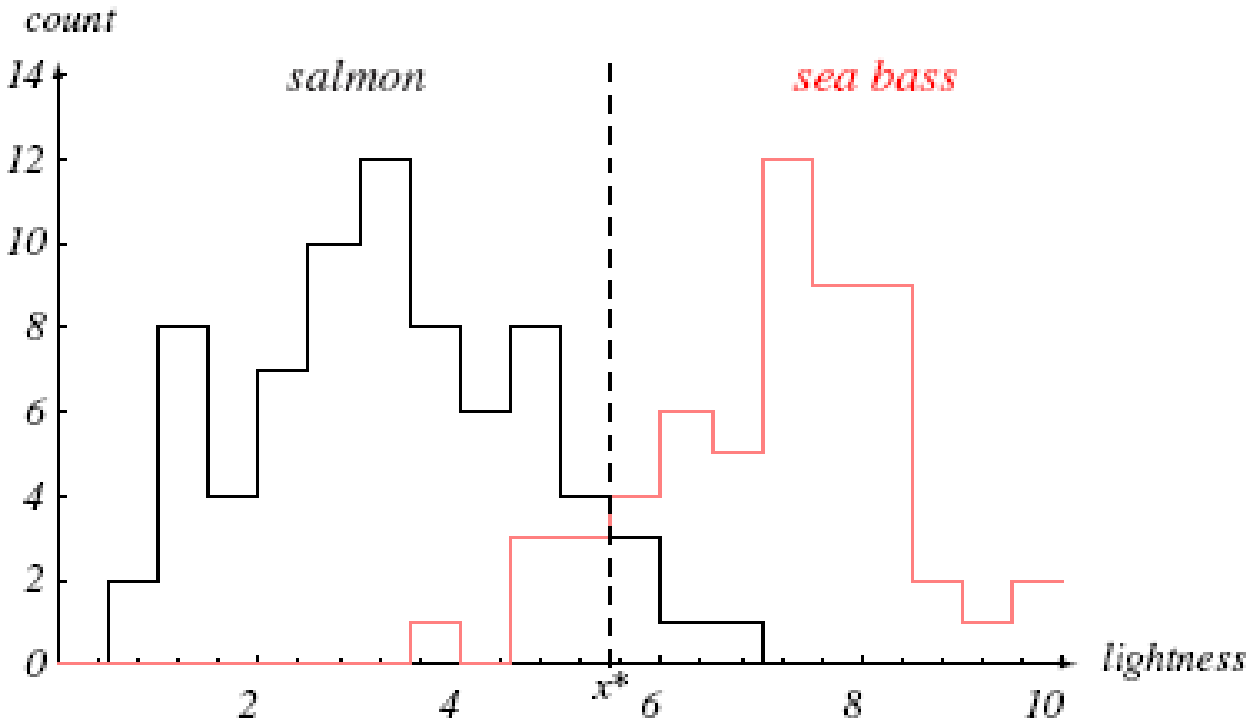
- Outros problemas a serem tratados:
  - Ruído
  - Normalização

Histograma do Brilho



[DUDA, HART & STORK, 2001]

Histograma do Brilho



MELHOROU!

[DUDA, HART & STORK, 2001]

Podemos melhorar?

Podemos melhorar?

- E se combinássemos ambas?

Podemos melhorar?

- E se combinássemos ambas?
- $\mathbf{x} = (x_1, x_2)^T$ onde:

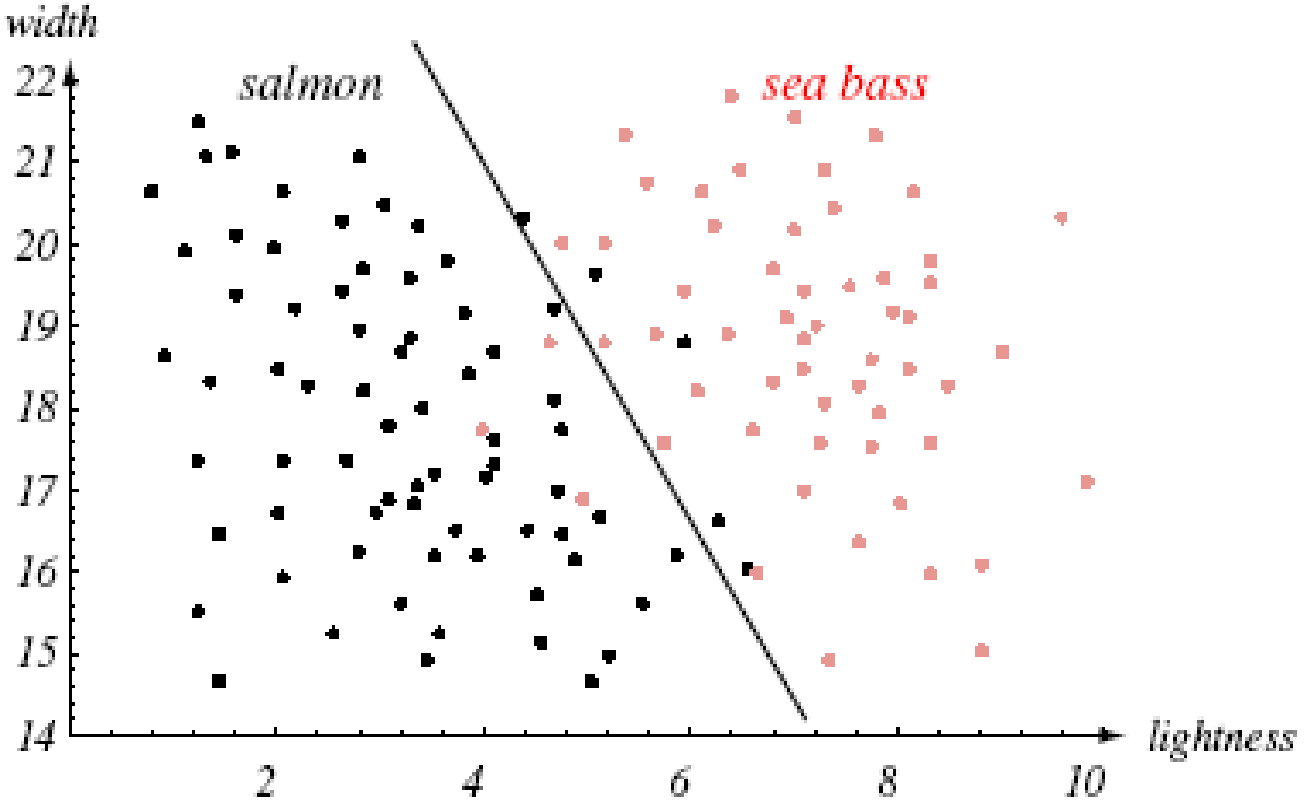
  x_1 é o comprimento (em cm)

  x_2 é a intensidade de brilho
- **x** é um vetor de características de um objeto (em um espaço de características bidimensional) – note o negrito para vetor
- x_1 e x_2 são variáveis aleatórias
- **x** é um vetor aleatório

Tamanho e brilho

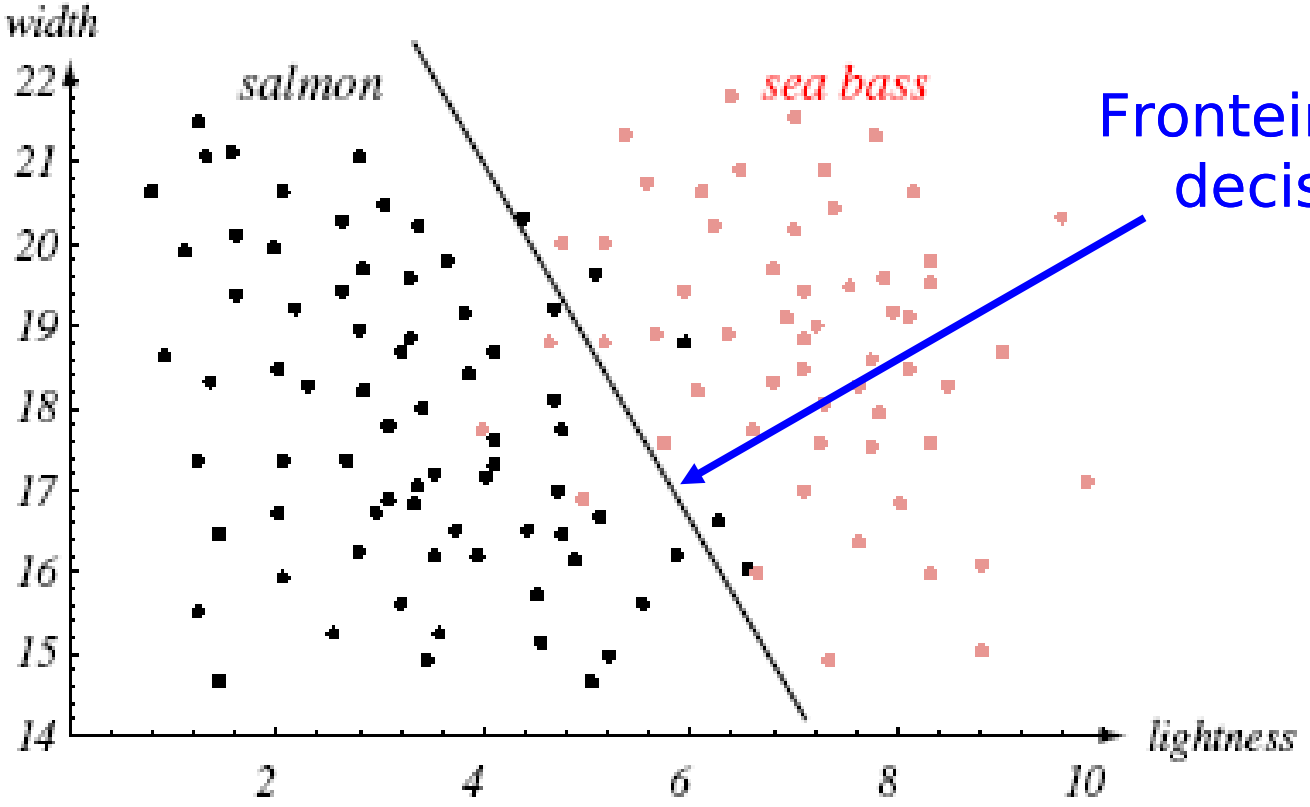
- Como representar?

Tamanho e brilho



[DUDA, HART & STORK, 2001]

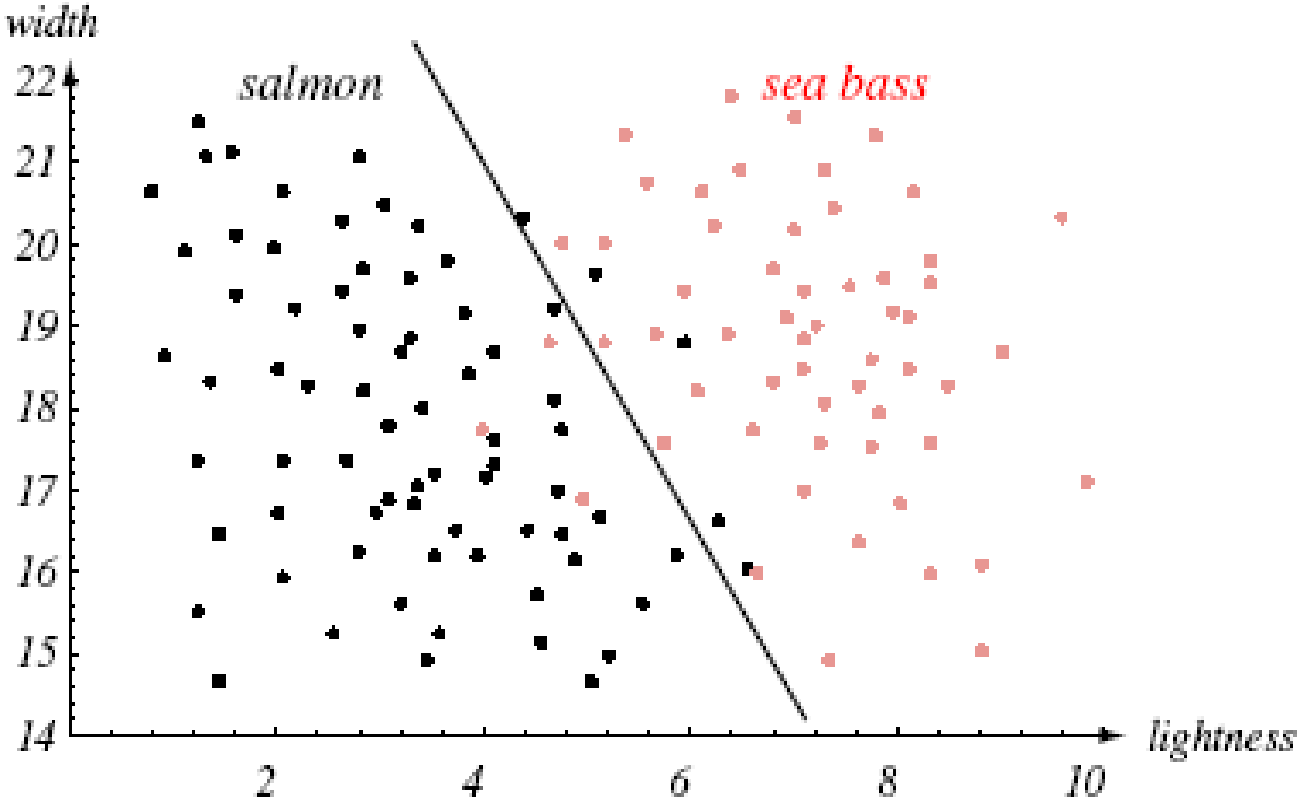
Tamanho e brilho



Classificador linear: a fronteira de decisão é um hiperplano (curvatura nula no espaço n-dimensional)

[DUDA, HART & STORK, 2001]

Tamanho e brilho



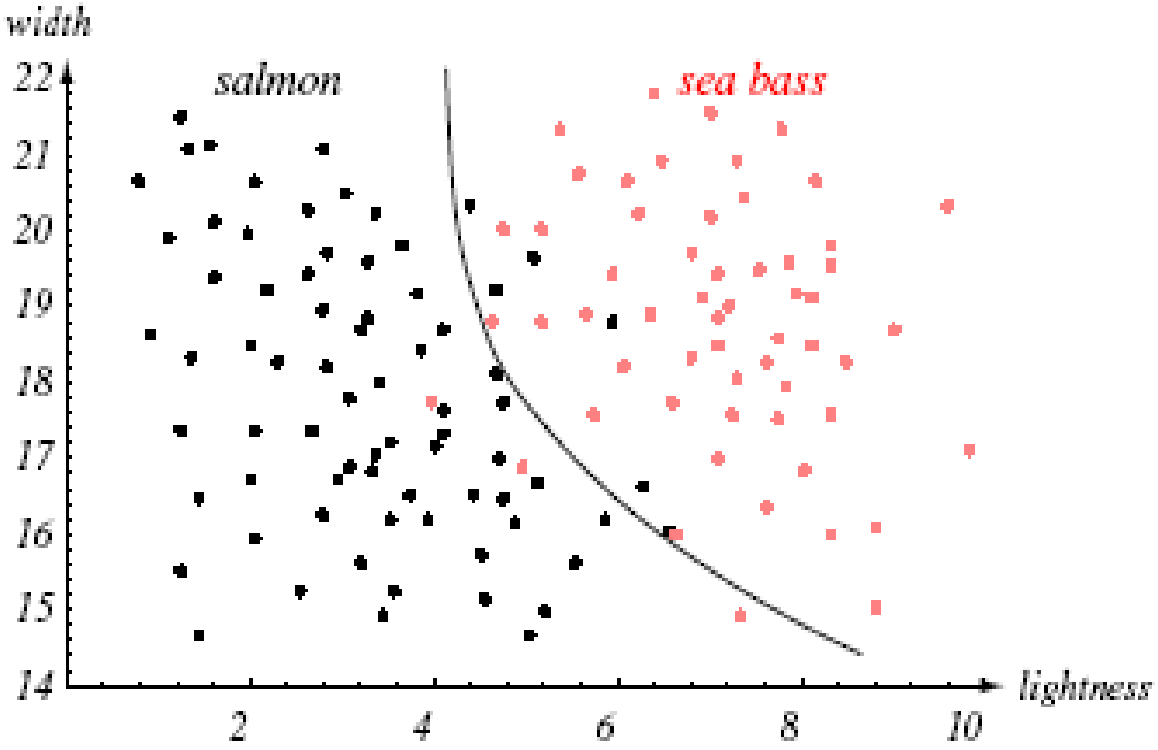
Podemos “desenhar” essa fronteira de uma forma diferente?

[DUDA, HART & STORK, 2001]

Tamanho e brilho

- Que tal assim?

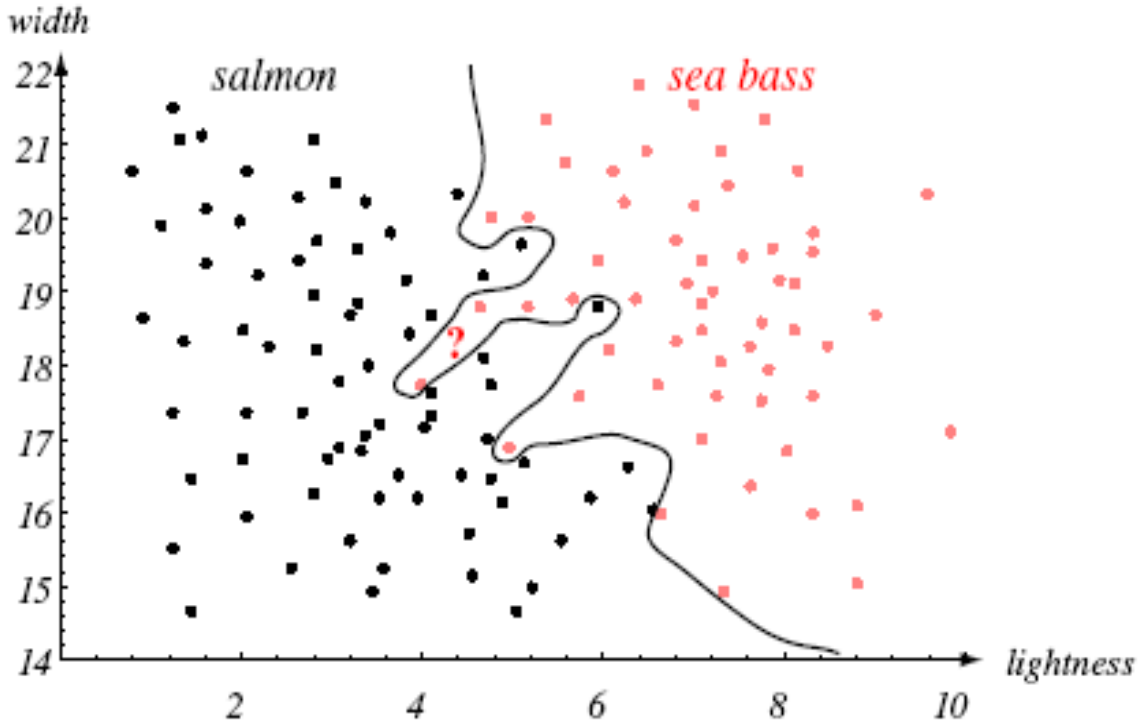
Classificador não é mais linear



[DUDA, HART & STORK, 2001]

Tamanho e brilho

- Ou assim?



[DUDA, HART & STORK, 2001]

Generalização x overfitting

1. Complexidade do classificador vale a pena?
2. Como será o desempenho (sucesso da classificação) para objetos fora da amostra de treinamento?

Generalização x overfitting

1. Complexidade do classificador vale a pena?
2. Como será o desempenho (sucesso da classificação) para objetos fora da amostra de treinamento?

- Não queremos decorar a amostra de treinamento (overfitting), mas aprender um conceito, criar um modelo a partir da generalização dos exemplos, sem generalizar demais
- Maior sucesso na amostra de treinamento não garante maior sucesso nos dados “novos”

Generalização x overfitting

- Generalização: permite aprender além do que está na amostra de treinamento. Mas se generalizar demais vai errar demais.
  - Ex: ao se querer aprender um conceito, pode erroneamente classificar tudo como sendo daquela classe alvo
- Overfitting: ótimos resultados na amostra de treinamento, mas ruim na amostra de teste.
  - No pior caso, só conhece (todos) os elementos da amostra de treinamento

Dilema na escolha do modelo

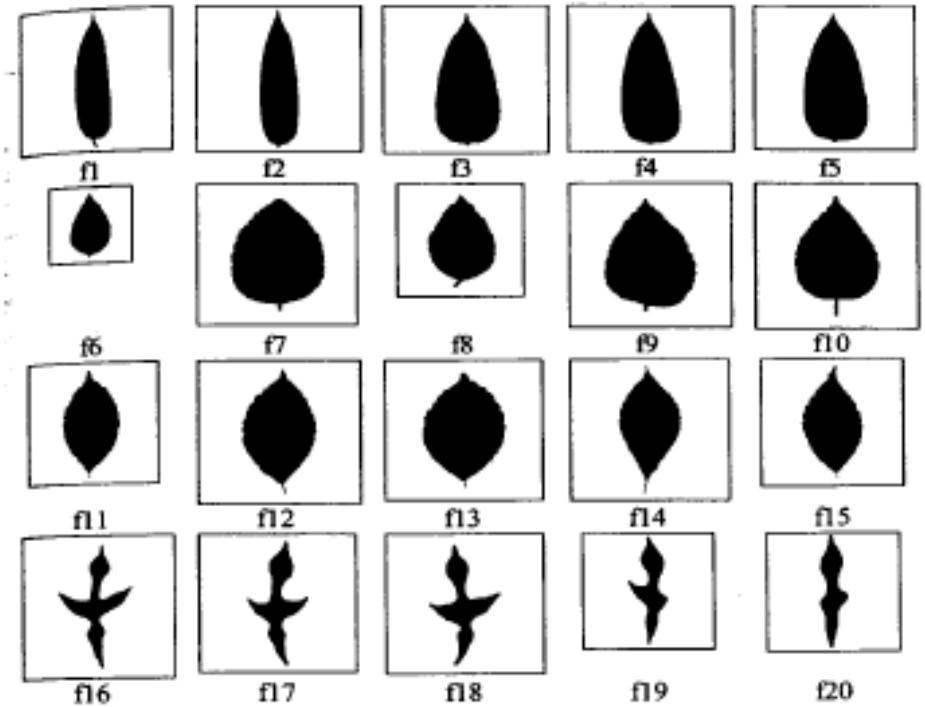
- Há um gradiente de modelos dos mais simples aos mais complexos: qual escolher?
- Tempo de processamento, memória, erros, custos, overfitting, ...
- Como predizer quão boa será a generalização do classificador?
- Discutiremos essas questões durante o curso.

Podemos aumentar o número de características?

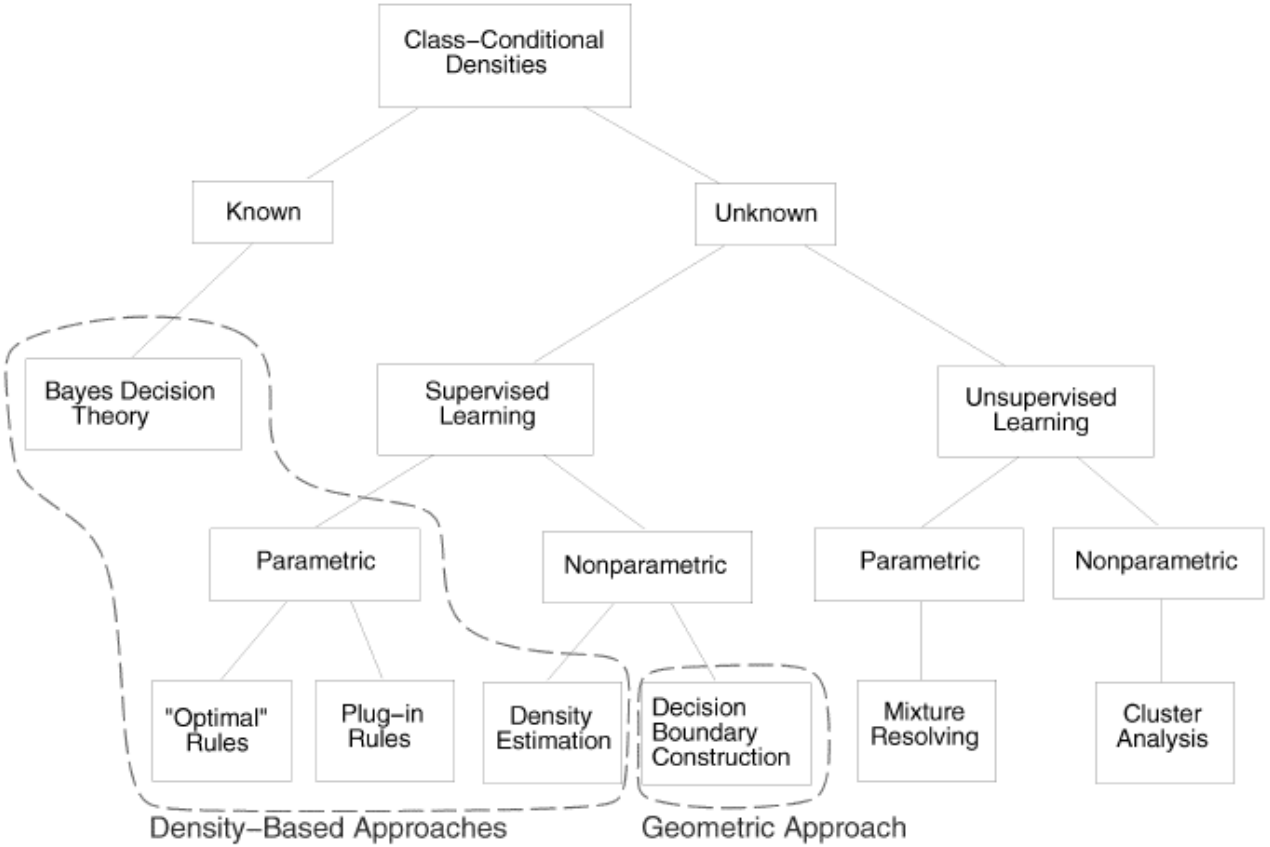
- Se usar duas características foi melhor que usar uma, podemos usar 3, 4 , 5, ...?
- Até onde ir?
- Sempre melhora?
- Aula que vem...

Classificação binária ou multiclasse

- Binária: 2 classes. Ex: robalos e salmões
- Multiclasse: mais de 2 classes. Ex: folhas de diferentes espécies



Classificação supervisionada x não-supervisionada



[JAIN et al, 2000]

Classificação

- Classificação supervisionada (por exemplos – amostra de treinamento)
  - Número de classes definido
  - Duas etapas:
    - Aprendizado (ou treinamento)
    - Reconhecimento
- Classificação não supervisionada
  - Número de classes não necessariamente definido
  - Objetivo: encontrar grupos de objetos
    - Alta similaridade entre objetos do mesmo grupo
    - Baixa similaridade entre objetos de grupos diferentes
    - Noção de distância

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Por quê?

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros

Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros
- Qual é o melhor?

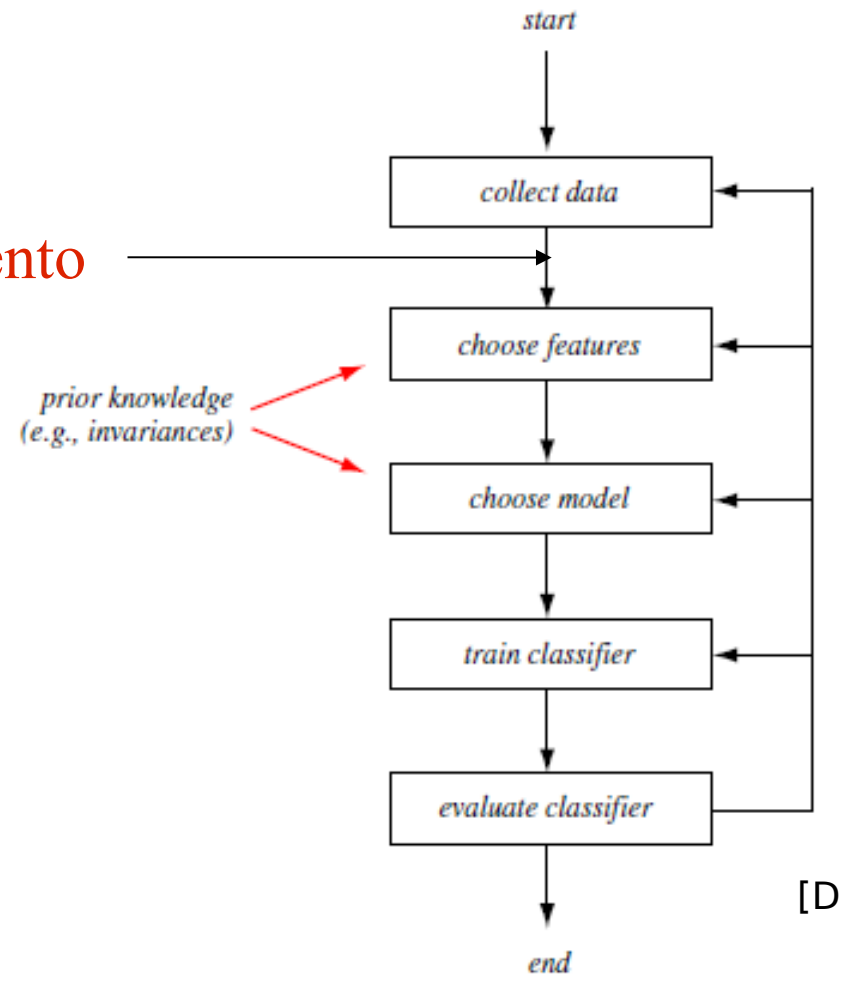
Classificação - observações

- Um mesmo conjunto de dados ⇒ diferentes classificadores
  - Diferentes características
  - Diferentes técnicas de construção de classificadores
  - Diferentes parâmetros

- Qual é o melhor?
  - Necessidade de avaliação dos classificadores
  - Comparação de desempenho entre opções

Ciclo de construção de classificadores baseados em aprendizado supervisionado

Pré-processamento



[DUDA, HART & STORK, 2001]

Exemplos de tarefas pré-processamento

- Eliminação (ou não) de instâncias com *missing values*
- Eliminação de ruídos
- Normalização
- Discretização

Outros fatores envolvidos em classificação

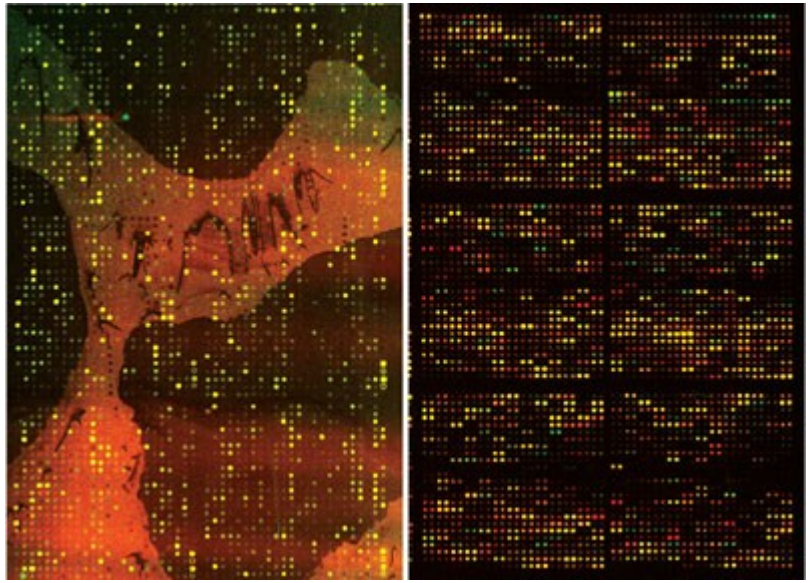
- Missing values:
  - pode-se não conhecer alguns atributos de algumas instâncias
  - Opções: usar ou não essas instâncias (dependendo do tamanho da amostra)
  - Se usar, o método utilizado tem que lidar com isso

Outros fatores envolvidos em classificação

- Eliminação de ruídos (exemplos):
  - Erros de sequenciamento (Bioinformática)
  - Imagens

Outros fatores envolvidos em classificação

- Normalização:
  - Tornar os elementos comparáveis (mesma escala)
  - Ex: variação de luz



Microarray:
Ruído e normalização

Outros fatores envolvidos em classificação

- Características invariantes ou não:
  - Ex (imagens): invariantes a escala, rotação, translação, oclusão, distorção
  - Ex (reconhecimento de fala): amplitude, voz, velocidade (você pode querer ou não detectar emoções)

Outros fatores envolvidos em classificação

- Dados potencialmente incorretos:
  - Duplicações
  - Erros de digitação
  - Informação não atualizada
  - Informação não confiável (ex: usa drogas?)

Outros fatores envolvidos em classificação

Tipos das variáveis:

- Contínuas (valores reais) x discretas (valores inteiros)
- Categóricas (sempre discretas): nominais ou ordinais
  - Nominais: sem relação de ordem matemática (sexo (0=masculino e 1 = feminino), cor)
  - Ordinais: possui ordem (faixa de renda, nível sócio-econômico, tamanho (P/M/G), ...)
- Intervalar: intervalo exato (e constante) entre os valores da escala (temperatura em graus Celsius)
- Racional: Presença de um zero absoluto permite estabalecer proporções entre os valores da escala, ex: falar que o valor y = 2*x corresponde ao dobro (ex: temperatura em Kelvin : 100K representa o dobro de calor de 50K)

Tipos de variáveis

Diferentes análises possíveis para cada tipo

Ex: normalmente não faz sentido:

- distância entre tipos de dores
- média de sexo
- etc.

(não são intervalares, cujos intervalos deveriam ser constantes)

Quadro 9.2 Níveis de mensuração e estatísticas possíveis

Escala	Exemplos	Operações empíricas básicas	Estatísticas possíveis
Nominal	Sexo, cor dos olhos, partido político	Determinação de igualdade	Número de casos % moda
Ordinal	Classificação em concursos, escalas tipo Likert	Anteriores, determinação >, <	Anteriores, mediana, percentis
Intervalar	Escore de QI, temperatura medida em graus Celsius	Anteriores, determinação dos intervalos das diferenças	Anteriores, média, desvio-padrão, correlação de postos (Spearman [estatística não paramétrica, dados não normais]), correlação produto-momento (Pearson)[estatística paramétrica, dados normalizados])
Racional	Tempo, temperatura medida em graus Kelvin, número de filhos	Anteriores, determinação da igualdade de razões	Todas, por exemplo, coeficiente de variação

[Appolinário, 2012]

Classificação não é trivial

Não basta simplesmente aplicar uma técnica

Tem que entender o problema específico, identificar possíveis fatores que atrapalhem a classificação

Atividades e trabalhos para as próximas aulas

- Aula que vem (15/08) não tem aula (Semana de SI e Workshop do PPGSI)
  - Presença obrigatória para alunos regulares do PPGSI e alunos especiais (lista de presença passará no Workshop no horário da aula)
  - Mas o “trabalho” deve ser feito para esta data

Atividade 1 – Definição dos grupos e escolha de um dataset para o trabalho (para 15/08)

- Dataset deve possuir dados rotulados de 2 classes (ex: compostos para drogas: ativos/inativos)
  - Iremos durante o curso aplicar técnicas de aprendizado supervisionado para realizar classificação binária
- Pode ser dados de pesquisa de vocês (preferencialmente) ou um dataset público (ex: UCI)
- Preencher na planilha Google (slide 7) :
  - Nomes dos integrantes
  - Dataset escolhido

Atividade 2 – Descrição do dataset (para 22/08)

- Entregar slides iniciais do trabalho final, contendo:
  - Descrição do dataset: o que é, quais são as classes, número de instâncias, número de características (tirando campos id e classificação), número de instâncias sem nenhum *missing value*
  - Tipos de cada uma das características (nominais, intervalares, etc.) (se muitas, quantas características de cada tipo)
  - Gerar o pdf dos slides e:
    - Postar no Moodle (Atividade 2 – Descrição do dataset)
    - Trazer para a aula de 22/08 para apresentar (caso seja sorteado)

Revisão necessária (tarefa para vocês!!!)

- Variável aleatória
  - Discreta e contínua
- Vetor aleatório
- Função de probabilidade
- Função de distribuição de probabilidade,

Função densidade de probabilidade,

Distribuição conjunta

Referências

- TOU, J. T.; GONZALEZ, R. C. **Pattern Recognition Principles**. Addison Wesley, 1974.
- DUDA, R.; HART, P.; STORK, D**. Pattern Classification and Scene Analysis**. John Willey, 2001 (Cap. 1)
- COSTA, L. F.; CESAR, R. M. Jr. **Shape Classification and Analysis: Theory and Practice**. CRC Press, 2009
- EVERITT, B. ; HOTHORN, T. **A handbook of statistical analysis using R**. Ed. Chapman & Hall/CRC
- MAGALHÃES, M. N.; LIMA, A. C. P. de: **Noções de Probabilidade e Estatística**. Edusp, 2002
- APPOLINARIO, F. **Metodologia da Ciência**: Filosofia e Prática de Pesquisa. Cap 9: Variáveis e Níveis de Mensuração. Cengage Learning.

Aula 1

Introdução a Problemas de Reconhecimento de Padrões e Conceitos Básicos

Profa Ariane Machado Lima