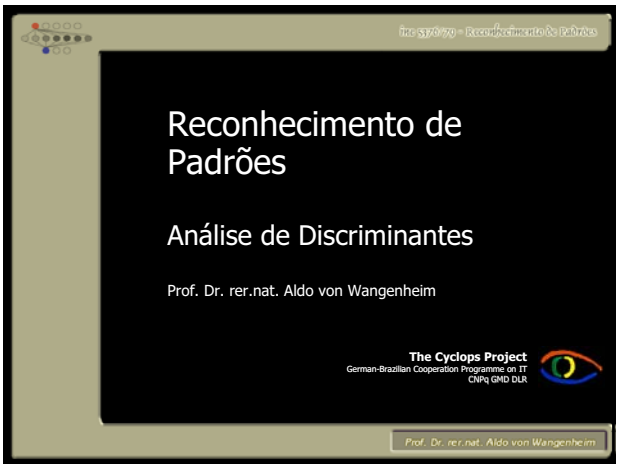
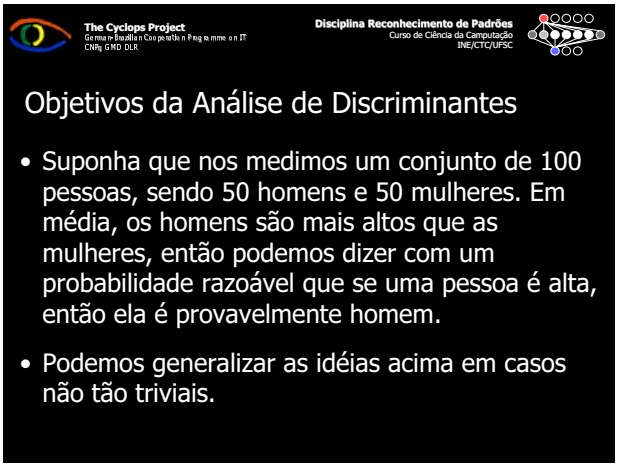

Reconhecimento de Padrões
Análise de Discriminantes
Prof. Dr. rer.nat. Aldo von Wangenheim
The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Prof. Dr. rer.nat. Aldo von Wangenheim


The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC
Análise de Discrimantes
A análise de funções discriminantes é utilizada para determinar quais variáveis discriminam entre dois ou mais grupos que ocorrem naturalmente.
• Clientes de uma empresa:
– Análise de Discriminantes: Como selecionar variáveis que melhor discriminam clientes que permanecem e clientes que abandonam os serviços da empresa?
– Construção de regras de classificação: Conhecidos os valores das variáveis de um novo cliente, classifica-lo no grupo dos que abandonam ou no grupo dos que permanecem na empresa.
• Clientes de um banco:
– Análise de Discriminantes: Como selecionar variáveis que melhor discriminam clientes que pagam e clientes que não pagam seus débitos?
– Construção de regras de classificação: Conhecidos os valores das variáveis de um novo cliente, classifica-lo no grupo dos que pagam ou no grupo dos que não pagam.


The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC
Objetivos da Análise de Discriminantes
• Suponha que nos medimos um conjunto de 100 pessoas, sendo 50 homens e 50 mulheres. Em média, os homens são mais altos que as mulheres, então podemos dizer com um probabilidade razoável que se uma pessoa é alta, então ela é provavelmente homem.
• Podemos generalizar as idéias acima em casos não tão triviais.

Objetivos da Análise de Discriminantes

Estruturando o que foi dito acima, queremos:

- Medir o poder de discriminação de cada variável ou grupo de variáveis;
- Descrever graficamente ou algebricamente diferentes grupos em termos de variáveis discriminadoras;
- Desenvolver regras para classificar novos elementos.

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Objetivos da Análise de Discriminantes

- A idéia por trás da análise de discriminantes é a determinação do quando os grupos diferem em relação as médias das variáveis e então usar estas variáveis para prever novos casos.
- Nosso objetivo será determinar quais variáveis são categóricas, *i.e.,* nos permitem discriminar entre categorias.

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Principais Perguntas

- Considerando amostras de ***p*** variáveis, relativas a elementos de ***g*** grupos, como pode-se :
  - verificar as variáveis que discriminam (separam) melhor os grupos;
  - medir analiticamente a separação dos grupos;
  - visualizar graficamente a separação dos grupos.
- Como organizo um esquema de classificação ?
  - A partir de amostras de ***p*** variáveis de elementos de vários grupos, como pode-se criar regras de classificação que permitam classificar novos elementos em um dos grupos?
  - Como avaliar a qualidade do processo de classificação?

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Exemplo

Tomemos para isso o seguinte exemplo:

- Grupo 1: mulheres não portadoras de hemofilia A (normais) (n_1 = 30)
- Grupo 2: mulheres portadoras de hemofilia A (portadoras) (n_2 = 22)
- Suponha que as variáveis discriminadoras são: X_1 e X_2 (duas variáveis contínuas medindo duas substâncias diferentes observadas em exames de sangue).

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

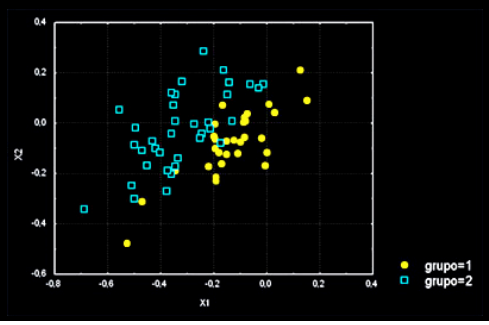
Exemplo: Os dados

ind.	GRUPO	X1	X2
1	1	-0,0056	-0,1657
2	1	-0,1698	-0,1585
3	1	-0,3469	-0,1879
...	...	...	...
31	2	-0,3478	0,1151
32	2	-0,3618	-0,2008
33	2	-0,4986	-0,0860
...	...	...	...

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

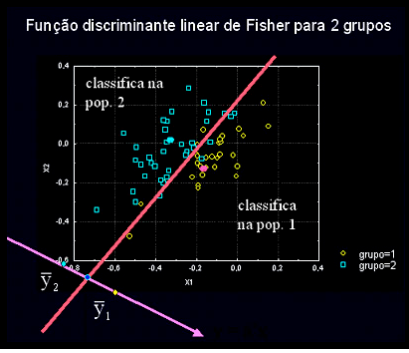
Exemplo:Um *scatter plot* da distribuição dos dados fica assim:



The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Função de Análise de Discriminantes



The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

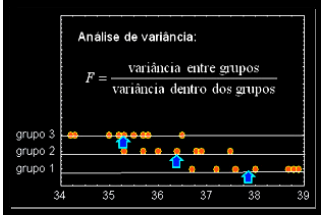
Formas de Análise de Discriminantes

Análise de Discriminates sobre uma única Variável

- Para teste se uma variável discrimina entre grupos usamos o teste F.
- O Teste F é calculado como a razão entre a variância inter grupos e a média da variância intra grupo. Se a variância inter-grupo é significativamente maior, então existe diferença entre as médias.

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC



Análise de variância com 2 agrupamentos

	SS entre	gl	SS dentro	gl	F
comp. sep.	76,7	1	72,31	148	157,0
larg. sép.	54,2	1	94,76	148	84,7
comp. pét.	126,9	1	22,13	148	848,6
larg. pét.	117,3	1	31,68	148	548,1

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Análise de Discriminantes com Múltiplas Variáveis

- Neste caso, teremos uma matriz com variâncias e covariâncias totais, e outra com a média das variâncias e covariâncias intra grupo.
- Comparamos as matrizes via teste F multivariado

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Presupostos Fundamentais

- Dados distribuídos seguindo uma distribuição normal.
  - O que nem sempre ocorre, o mundo não é tão simples
  - Precisamos de um número significativo de dados
- A variância e a covariância entre as variáveis de cada grupo é homogênea, ou seja, não há diferenças absurdas entre elas.

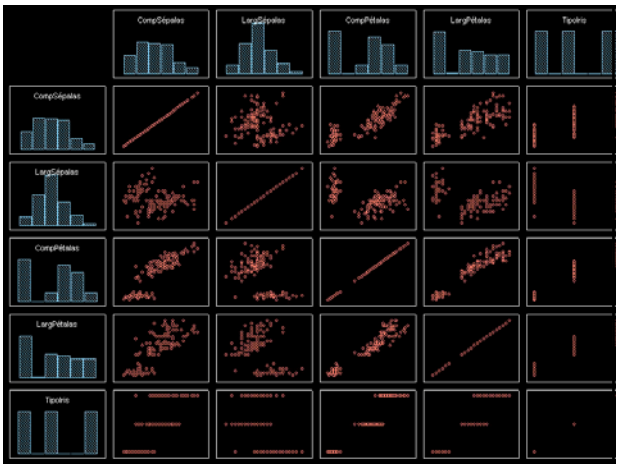
The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Modos de usar

Análise de Discriminantes do Tipo Passo a Passo

- A cada passo, revisamos todas a variáveis e avaliamos qual contribui mais na discriminação entre os grupos. Está variável então será incluída em um "modelo".
- Guiada pelo teste F para a inclusão o remoção da variável.



The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Análise de Discrimantes para Determinação de Funções Discriminantes entre Vários Grupos

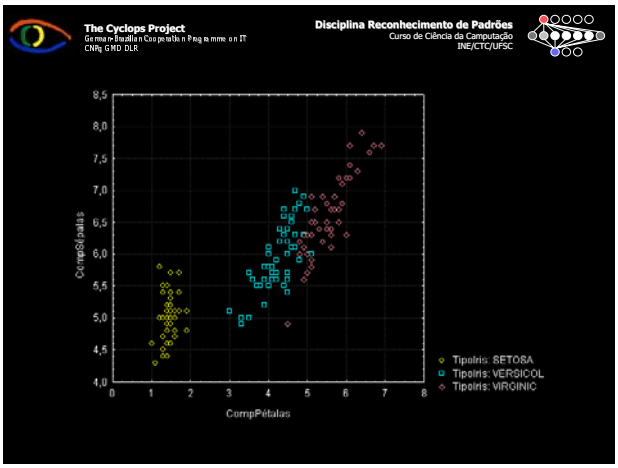
- Calculamos mais que uma função discriminante.
- Temos uma função que discrimina um grupo de cada outro grupo do modelo.

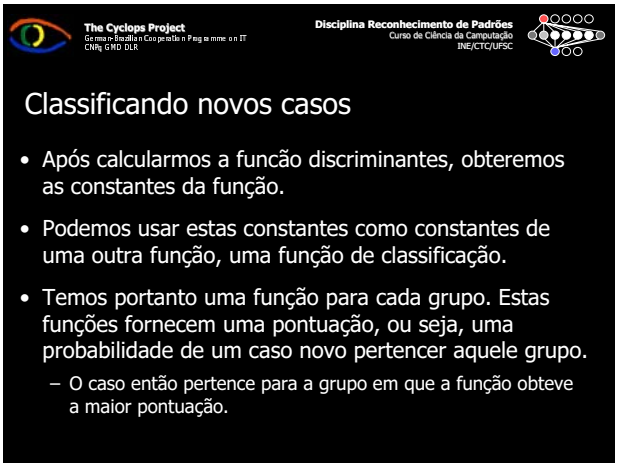
The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

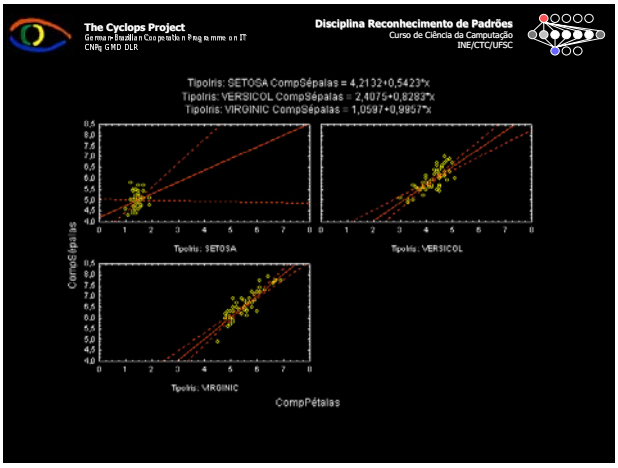
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Exemplo

- Para exemplificar a utilização de análise de discriminates vamos nos basear em um conjunto de dados bastante utilizado para demonstrar Análise de Disciminantes: o conjunto de dados sobre três espécies de flores do gênero *Iris*, *Iris setosa* (comum nos jardins da nossa ilha), *Iris versicolor* e *Iris virginica*. Estes dados foram colhidos por Fisher em 1936 e até hoje servem de exemplo de como se pode escolher funções discriminantes para um conjunto de dados composto por três classes. Os dados descrevem 150 espécimes de Iris de acordo com 4 características: comprimento das sépalas, comprimento das pétalas, largura das sépalas e largura das pétalas. A quinta váriável é a ***variável de grupo*** ou ***variável categórica***, que associa a classificação a cada espécime ou caso observado. Apresentamos uma parte desse conjunto de dados abaixo:


The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC
CompSépalas
CompPétalas
Tipoíris: SETOSA
Tipoíris: VERSICOL
Tipoíris: VIRGINIC


The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC
Classificando novos casos
• Após calcularmos a funcão discriminantes, obteremos as constantes da função.
• Podemos usar estas constantes como constantes de uma outra função, uma função de classificação.
• Temos portanto uma função para cada grupo. Estas funções fornecem uma pontuação, ou seja, uma probabilidade de um caso novo pertencer aquele grupo.
– O caso então pertence para a grupo em que a função obteve a maior pontuação.

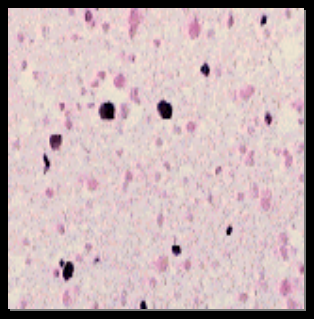

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR
Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC
Tipoíris: SETOSA CompSépalas = 4,2132+0,5423*x
Tipoíris: VERSICOL CompSépalas = 2,4075+0,8283*x
Tipoíris: VIRGINIC CompSépalas = 1,0597+0,9907*x
CompSépalas
CompPétalas

The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Outro exemplo: Câncer Cerebral

- No câncer cerebral mais comum, o que ocorre nas células gliais ou de suporte do cérebro, o fator mais importante a ser determinado é a taxa de crescimento do tumor, o que vai determinar a forma de tratamento.
- Determinar esta taxa depende de uma série de fatores que um patologista experiente é capaz de identificar em uma lâmina de uma biópsia cerebral.



The Cyclops Project
German-Brazilian Cooperation Programme on IT
CNPq GMD DLR

Disciplina Reconhecimento de Padrões
Curso de Ciência da Computação
INE/CTC/UFSC

Exercício

- Procure nos "Links Úteis" da página por fontes de software livre para Análise de Discriminantes.
  - http://www.inf.ufsc.br/~awangenh/RP/estatisticas.html#4.5. Links
- Tome um conjunto de quatro sets de dados, dentre estes:
  - Os dados da flor do Gênero Iris disponíveis na página
  - Os dados de câncer cerebral (gliomas) que serão fornecidos
  - Outros dois conjuntos quaisquer que você deverá procurar também nos "Links Úteis".
- Realize dois conjuntos de Análises de Discriminantes sobre estes sets de dados:
  - Um deles buscando uma variável discriminatória para divisão em apenas dois grupos
  - Outra multigrupos, buscando o conjunto de funções discriminatórias.
  - No último caso utilize apenas metade dos dados para a A.D., utilizando então as funções geradas para classificar os dados restantes. Verifique a acurácia de sua classificação.