

Instituto de Computação
Unicamp



MO 906 - Introdução à Inteligência Artificial

1º Semestre de 2010

Prof. Siome Goldenstein

Projeto

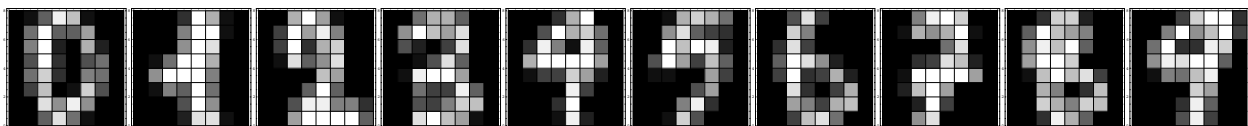
Quarta-feira, 14/06/2010 12:00.

Este trabalho tem como objetivo extrair informações a partir de dados não anotados. Cada elemento do conjunto de dados é uma linha com 64 inteiros [0-16] separados por vírgulas. Cada elemento descreve uma matriz 8x8, com 16 tons de intensidade, que representa um dígito manuscrito [0,9], capturado através de algum mecanismo de “scanning” ótico ou por “tablet”. Infelizmente, não sabemos a que dígito cada elemento representa.

Neste trabalho, tentaremos fazer a separação automática destes grupos.

- O trabalho é em grupos de dois alunos.
- Não é permitido compartilhamento de funções entre grupos distintos antes da entrega do trabalho - isto será considerado fraude (vide critérios na ementa do curso).
- Após a entrega, posso escolher um elemento de cada grupo para uma avaliação oral sobre o trabalho. O desempenho do escolhido determinará a nota do grupo inteiro.
- A qualidade do relatório é importante, e como envolve gráficos, deve ser feita no computador. Justifique tudo o que fizerem, e acrescentem o código documentado de todas as implementações realizadas.
- É necessário enviar por e-mail um arquivo R que quando executado gera um arquivo texto e algumas imagens gravadas no mesmo diretório com os principais resultados descritos no relatório.
- Os dados estão em
<http://www.ic.unicamp.br/~siome/teaching/2010/mc906-0110/material/digits.raw>

Alguns exemplos de elementos do conjunto de dados:



1 Infra Estrutura

Neste trabalho vocês devem aprender a utilizar o R¹, uma ferramenta poderosa para manipulação de dados.

1. Aprenda a utilizar o ambiente e a linguagem com um dos vários tutoriais disponíveis na rede.
2. Descubra como importar os dados.
3. Crie a funcionalidade de desenhar a representação gráfica, imagem 2D, de um elemento qualquer, permitindo que essas intensidades sejam números reais no intervalo [0,16].

2 Clusterização

Implemente (não use uma rotina pronta) o método K-Means de clusterização para separar seus dados em 10 grupos.

1. Com o auxílio da função desenvolvida em 1.3, desenhe a representação do centroide de cada grupo.
2. Analise a sensibilidade do resultado do algoritmo para 20 diferentes conjuntos iniciais de sementes e determine, usando métricas claras, qual a melhor clusterização.
3. (Extra) Implemente outros métodos para clusterização e compare seus resultados.
4. (Extra) Compare os resultados para clusterização feita com 5, 8, 12 e 15 grupos (ao invés de 10).

3 Análise dos Grupos

1. Calcule a matriz de covariância de cada grupo encontrado em 2.1.
2. Utilizando 3.1, faça a Análise de Componentes Principais (PCA) de cada grupo. Com o auxílio da função de 1.3, para cada grupo, desenhe os quatro valores

$$\mu \pm \sigma_1 v_1 \pm \sigma_2 v_2,$$

aonde μ é o centroide do grupo, v_1 e v_2 são os primeiro e segundo componentes principais e σ_1 e σ_2 o primeiro e segundo valores principais.

3. (Extra) Repita 3.1 e 3.2 para as outras técnicas em 2.3.

4 Consistência dentro dos Grupos

Utilizando 2.1 e 3.1, calcule a distância de Mahalanobis de cada elemento para o centroide do grupo ao qual ele pertence (utilizando a matriz de covariância do grupo).

1. Desenhe o histograma de distâncias de Mahalanobis ao centroide de cada grupo, e mostre (com 1.3) quem são os três elementos mais distantes do centroide.
2. Porque é que esta análise é importante?
3. (Extra) Faça o mesmo com as demais técnicas implementadas em 2.3.

¹<http://www.r-project.org/>