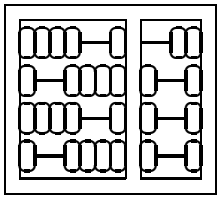


Instituto de Computação
Unicamp

MO 810 - Tópicos em IA - Aprendizado de Máquina

2º Semestre de 2006

Profs. Jacques Wainer e Siome Goldenstein

Projeto 1 - Clusterização

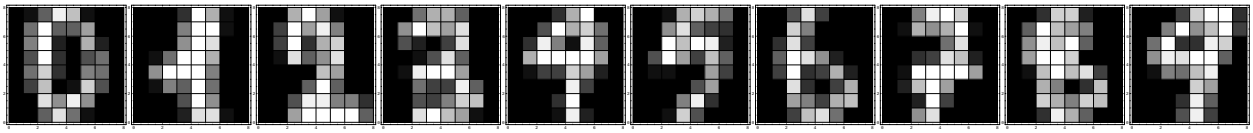
Quinta, 12/09/2009, no início da aula.

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". Os dados estão todos misturados e não sabemos qual digito cada elemento representa.

Neste trabalho, tentaremos fazer a identificação destes grupos de forma automática.

- O trabalho é nos grupos já estabelecidos.
- É permitido o uso de funções e bibliotecas prontas para as técnicas.
- Não é permitido compartilhamento de resultados e funções entre grupos distintos antes da entrega do trabalho.
- 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.
- Os dados estão no arquivo `digits.raw` localizados na página do curso.

Alguns exemplos de elementos do conjunto de dados:



1 Infra Estrutura

Escolha sua plataforma preferida para implementação. Recomendamos o uso de pacotes para manipulação matemática e estatistica, por exemplo: R (ou S-Plus), matlab (ou octave), mathematica e maple. No entanto, para aqueles que preferirem, C/C++/Pascal são também opções.

1. Encontre um método de importar os dados para dentro de seu ambiente.
2. 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

Utilize os três métodos de clusterização vistos em aula (K-Means, K-Means + Mahalanobis, Fuzzy K-Means + Mahalanobis) para separar seus dados em 10 grupos, e para cada um deles faça a análise abaixo:

1. Com o auxílio da função desenvolvida em 1.2, desenhe a representação do centroide de cada grupo.
2. Analise a sensibilidade do resultado do algoritmo para diferentes conjuntos iniciais de sementes.
3. Compare os resultados para clusterização feita com 5, 8, 12 e 15 grupos (ao invés de 10).

3 Análise dos Grupos

1. Para cada técnica de 2, 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 1.2, para cada grupo, desenhe os quatro valores

$$\mu + \sigma_1 v_1, \quad \mu - \sigma_1 v_1, \quad \mu + \sigma_2 v_2, \quad \mu - \sigma_2 v_2,$$

aonde μ é o centroide do grupo, σ_1 e σ_2 o maior e o segundo maior valores singulares e v_1e v_2 são os componentes principais associados a σ_1 e σ_2 respectivamente.

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. Para cada grupo, desenhe o centroide e, ao seu lado, os 5 elementos mais distantes do centroide de acordo com a métrica de mahalanobis induzida pela matriz de covariância deste grupo.
2. Porque é que esta análise é importante?