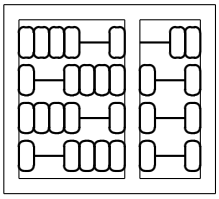

MC906 – Trabalho .:. Clusterização de Imagens

Instituto de Computação — UNICAMP

Prof.: Anderson Rocha anderson.rocha@ic.unicamp.br

Deadline: 23/10/2011 – Entrega por e-mail
(PDF do relatório + arquivo .R)

Objetivos

Este trabalho tem como objetivo extrair informações a partir de dados não anotados.

No arquivo `digitos.zip`, temos imagens de números. Cada imagem do conjunto possui 64×64 pixels e está representada no formato PGM onde cada pixel tem um valor 0 ou 1. O trabalho consiste em encontrar uma representação para esses dígitos que seja razoável para agrupá-los e encontrar os grupos mais significativos e seus representantes.

Suponha que os dados estão todos misturados e não sabemos qual dígito cada elemento representa.

Atividades

1. Usando a técnica de agrupamento K-Médias (K-Means) vista em aula, agrupe os dados em 10 grupos. Tem como descobrir o valor $K = 10$ automaticamente? Demonstre.

2. Crie uma funcionalidade de impressão e imprima o centróide de cada um dos $K = 10$ grupos.

3. Analise a sensibilidade dos grupos modificando as sementes iniciais.

4. Divida seus dados aleatoriamente entre treinamento e teste (50%-50%) e selecione "apenas" 80% dos componentes que mais variam após transformar seus dados de treinamento com PCA como visto em aula. Projete os dados de treinamento no sub-espaço resultante e faça a clusterização com $K = 10$.

   Para cada exemplo do teste, projete esse exemplo no sub-espaço calculado anteriormente e descubra a qual grupo o exemplo deve ser atribuído usando distância Euclidiana. Calcule o acerto do algoritmo. O algoritmo acerta mais ou menos com 50% das componentes selecionadas?

Requisitos

O trabalho é individual e deve ser feito em R. Um relatório (max. 6 páginas) deve ser entregue ao final.