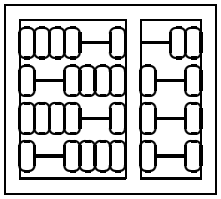


Instituto de Computação
Unicamp

MO 416 - Introdução à Inteligência Artificial

1º Semestre de 2010 - **Lista 2**

Prof. Siome Goldenstein

Entrega: Terça, 13/04/2010, **no início da aula.**

Questão 1 - Conceitos Matemáticos

1. Seja $x_1, \cdots, x_n$ amostras d-dimensionais e Σ qualquer matriz quadrada $d \times d$ não singular. Mostre que o vetor x que minimiza

$$\sum_{k=1}^{n} (x_k - x)^t \Sigma^{-1} (x_k - x)$$

é a média das amostras

$$\bar{x} = \frac{1}{n} \sum_{k=1}^{n} x_k.$$

2. Considere um conjunto de $n = 2k+1$ amostras, k delas coincidem em $x = -2$, k em $x = 0$ e uma delas em $x = a > 0$.

   (a) Mostre que a partição de dois grupos que minimiza

$$J_e(2) = \sum_{i=1}^{2} \sum_{x_i \in D_i} \|x - \mu\|^2$$

   agrupa as k amostras em $x = 0$ com a amostra $x = a$ se $a^2 < 2(k+1)$.

   (b) Qual é o agrupamento ideal quando $a^2 > 2(k+1)$?

3. Assuma que temos um número n (n grande) de amostras de uma Gaussiana d-dimensional – $p(x) \sim N(\mu, \Sigma)$, onde Σ é uma matriz de covariância simétrica positiva definida arbitrária.

   (a) Prove que a distribuição de $J_e(1) = \sum_{x \in D} \|x - \mu\|^2$ é normal com média $nd\sigma^2$. Expresse σ em função de Σ.

   (b) Prove que a variância desta distribuição é $2nd\sigma^4$.

Questão 2 - K-Means

1. **Geração de dados sintéticos**. Crie um dataset sintético em 2D com:

   (a) 30 amostras de uma distribuição Gaussiana com $\mu = [2.5, 2.5]^\top$ e com $\Lambda = I$.

   (b) 20 amostras de uma distribuição uniforme em $x_1 \in [-5, -3], x_2 \in [-1, 0]$.

   (c) 20 amostras de uma distribuição Gaussiana com $\mu = [2.5, -2.5]^\top$ e onde Λ tem autovalor 2 associado ao autovetor $[1, 1]^\top$ e autovalor 0.5 associado ao outro autovetor.

   (d) 30 amostras de uma distribuição Gaussiana com $\mu = [-2.5, 2.5]^\top$ e onde Λ tem autovalor 1.2 associado ao autovetor $[1.5, 1]^\top$ e autovalor 0.8 associado ao outro autovetor.

2. Implemente o K-Means e a métrica de avaliação de clusterizações

$$J_e \sum_{i=1}^{k} \sum_{x \in D_i} \|x - \mu_i\|^2 .$$

3. Rode sua implementação de K-Means 10 vezes diferentes, com valores inicias distintos, para K = 2, 3, 4, 5, 6, 7 e 8.

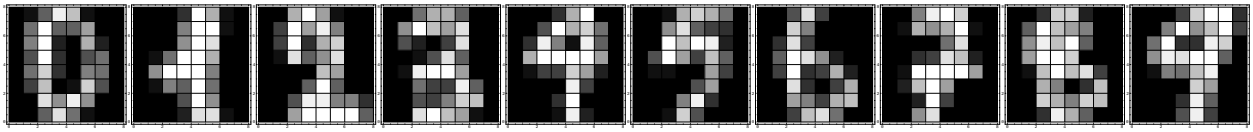
4. Crie uma forma gráfica e compacta para visualizar estes resultados e mostre ao seu lado o valor de J_e.

Questão 3 - Digitos Manuscritos

Os dados necessários para esta questão estão disponíveis em:

`http://www.ic.unicamp.br/~siome/teaching/2010/mo416-0110/data/digits.raw`

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 digito cada elemento representa. Alguns exemplos de elementos do conjunto de dados:



1. Encontre um método de importar os dados para dentro de seu ambiente de desenvolvimento e 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. Implemente e utilize o algoritmo de K-Means para clusterização deste dado em 10 grupos. Nesta questão não quero o uso de uma biblioteca pronta de clusterização, e é necessário apresentar em anexo a listagem da implementação.

   (a) Com o auxílio da função desenvolvida em 3.2, desenhe a representação do centroide de cada grupo.

   (b) Analise a sensibilidade do resultado do algoritmo para diferentes conjuntos iniciais de sementes.

3. Calcule a matriz de covariância de cada grupo encontrado.

4. Com o auxílio da função de visualização, para cada grupo, desenhe

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

onde μ é o centroide do grupo, σ_1 e σ_2 são o maior e o segundo maior autovalores da matriz de covariância calculada em 3.3 e v_1 e v_2 os autovetores associados a σ_1 e σ_2.

5. Calcule a distância de Mahalonobis de cada elemento para o centroide do grupo ao qual ele pertence (utilizando a matriz de covariância do grupo). Calcule e mostre o histograma destas distâncias para cada grupo.

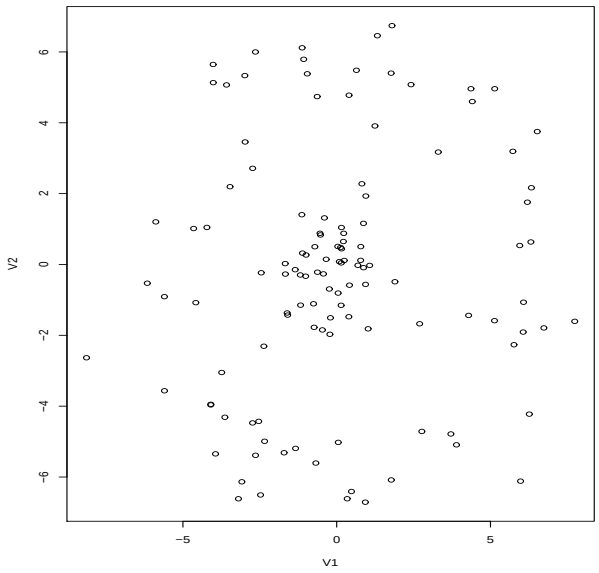
6. Para cada grupo desenhe o seu centroide e os 3 elementos mais distantes, mostre o valor da distância calculada em 3.5 de cada um deles.

Questão 4 - Normalized-Cuts

Os dados necessários para esta questão estão disponíveis em:

`http://www.ic.unicamp.br/~siome/teaching/2010/mo416-0110/data/toy.raw`

Este é um exemplo artificial com dados 2D, onde cada linha possui coordenadas x e y.



1. Implemente o algoritmo Normalized Cuts e aplique sobre estes dados. Mostre graficamente as 10 clusterizações geradas pelos diferentes candidatos a threshold (com pontos de grupos diferentes com cores ou marcações distintas), indicando o valor do Normalized-Cut de cada um. Indique também qual destes é o resultado final.

2. Utilize este conceito para criar uma "ordenação" dos dados, em termos de pertinência entre os dois grupos. Crie uma imagem para cada uma destas possibilidades, e faça um vídeo que anime o efeito de transição entre os grupos visualmente.

3. Como este conceito de ordenação pode ser feito com os algoritmos de clusterização por Expectation-Maximization, como por exemplo o K-Means?