

Projeto AM 2017-1

Francisco de A. T. de Carvalho¹

1 Centro de Informatica-CIn/UFPE
Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE, Brasil,
fatc@cin.ufpe.br

- 1) No conjunto de dados "Image Segmentation" do site uci machine learning repository considere a tabela de dados segmentation.test (<http://archive.ics.uci.edu/ml/machine-learning-databases/image>). Essa tabela de dados contém 2100 objetos e 7 classes. Os objetos são descritos por 19 variáveis que podem ser divididas em 2 views:
 - Shape view: as primeiras 9 variáveis
 - RGB view: as 10 ultimas variáveis

Produza duas matrizes de dissimilaridade com a distancia Euclidiana, uma para cada view, e execute o algoritmo da seção 2.2.2 do artigo "F.A.T. de Carvalho, Y. Lechevallier and F.M. melo, Relational Partitioning Fuzzy Clustering Algorithms Based on Multiple Dissimilarity Matrices, Fuzzy Sets and Systems, 215, 1-18, 2013" nessa tabela de dados 100 vezes para obter uma partição fuzzy em 7 grupos e selecione o melhor resultado segundo a função objetivo. A partir da partição fuzzy, obtenha a partição hard em 7 grupos e calcule o índice de Rand corrigido.

Observações:

- No algoritmo da seção 2.2.2, os representantes dos grupos são calculados segundo a proposição 2.3, os pesos de relevância das views são calculados com a equação (9), os graus de pertinencia dos objetos nos grupos são calculados com a equação (11);
- Parametros: número de grupos $K = 7$; parametro $m = 1.6$; parametro $q = 3$;
- Para o melhor resultado imprimir: i) a lista de representantes de cada grupo, ii) o matrix de pesos de relevancia das views, iii) a partição exclusiva (para cada grupo, a lista de objetos), iv) 0 índice de Rand corrigido.

- 2) Considere novamente a tabela de dados "Image Segmentation". Os exemplos são rotulados segundo as classes "brickface", "sky", "foliage", "cement", "window", "path", "grass".
- a) Use validação cruzada estratificada "30 times ten fold" para avaliar e comparar classificadores descritos abaixo. Se necessário, retire do conjunto de aprendizagem, um conjunto de validação para fazer ajuste de parametros e depois treine o modelo novamente com os conjuntos aprendizagem + validação.
 - b) Obtenha uma estimativa pontual e um intervalo de confiança para a taxa de erro de cada classificador;
 - c) Usar Friedman test (teste não parametrico) para comparar os classificadores. Se necessário, usar também o Nemenyi test (pos teste);

Considere os seguintes classificadores:

- i)) Classificador bayesiano. Considere a seguinte regra de decisão: afetar o exemplo \mathbf{x}_k à classe ω_i se $P(\omega_i|\mathbf{x}_k) = \max_{i=1}^7 P(\omega_i|\mathbf{x}_k)$ com $P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^C p(\mathbf{x}_k|\omega_r)P(\omega_r)}$

- a) Estime $P(\omega_i)$ pelo metodo de maxima verossimilhança.
- b) Para cada classe ω_i ($i = 1, 2, 3$) estime $p(\mathbf{x}_k|\omega_i) = p(\mathbf{x}_k|\omega_i, \theta_i)$ pelo método da máxima verossimilhança, onde $\theta_i = \begin{pmatrix} \mu_i \\ \Sigma_i \end{pmatrix}$,

$$\mu_i = \begin{pmatrix} \mu_{i1} \\ \dots \\ \mu_{ij} \\ \dots \\ \mu_{id} \end{pmatrix}, \text{ e } \Sigma_i = \begin{pmatrix} \lambda_{i1} & \dots & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & \lambda_{ij} & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & \lambda_{id} \end{pmatrix},$$

i)) Classificador bayesiano.

b) supondo uma normal multivariada:

$$p(\mathbf{x}_k | \omega_i, \theta_i) = (2\pi)^{-\frac{d}{2}} (\prod_{j=1}^d \lambda_{ij})^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} \sum_{j=1}^d \frac{(x_{kj} - \mu_{ij})^2}{\lambda_{ij}} \right\}. \text{ Use como estimativas}$$

$$\hat{\boldsymbol{\mu}}_i = \begin{pmatrix} \hat{\mu}_{i1} \\ \vdots \\ \hat{\mu}_{ij} \\ \vdots \\ \hat{\mu}_{id} \end{pmatrix}, \text{ e } \hat{\boldsymbol{\Sigma}}_i = \begin{pmatrix} \hat{\lambda}_{i1} & \dots & 0 & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & \hat{\lambda}_{ij} & \dots & 0 \\ \vdots & \dots & \vdots & \dots & \vdots \\ 0 & \dots & 0 & \dots & \hat{\lambda}_{id} \end{pmatrix}, \text{ onde } \hat{\mu}_{ij} = \frac{1}{n} \sum_{k=1}^n x_{kj} \text{ e}$$

$$\hat{\lambda}_{ij} = \frac{1}{n} \sum_{k=1}^n (x_{kj} - \hat{\mu}_{ij})^2$$

• Treine dois classificadores bayesianos, um em cada view.

ii) Usar o k-NN com distancia Euclidiana para fazer a classificação dos dados. Treine dois k-NNs, um para cada view. Use conjunto de validação para fixar o numero de vizinhos.

iii) Voto majoritário: afetar o exemplo \mathbf{x}_k a classe ω_j se

$$\Delta_{jkB1} + \Delta_{jkB2} + \Delta_{jkKNN1} + \Delta_{jkKNN2} = \max_{r=1}^7 (\Delta_{rkB1} + \Delta_{rkB2} + \Delta_{rkKNN1} + \Delta_{rkKNN2}) \text{ onde}$$

$\Delta_{rkB1} = 1$ se o objeto \mathbf{x}_k é classificado na classe ω_r ; $\Delta_{rkB1} = 0$ senão;

$\Delta_{rkB2} = 1$ se o objeto \mathbf{x}_k é classificado na classe ω_r ; $\Delta_{rkB2} = 0$ senão;

$\Delta_{rkKNN1} = 1$ se o objeto \mathbf{x}_k é classificado na classe ω_r ; $\Delta_{rkKNN1} = 0$ senão.

$\Delta_{rkKNN2} = 1$ se o objeto \mathbf{x}_k é classificado na classe ω_r ; $\Delta_{rkKNN2} = 0$ senão.

Observações Finais

- No Relatório e na saída da ferramenta devem estar bem claros:
 - a) como foram organizados os experimentos de tal forma a realizar corretamente a avaliação dos modelos e a comparação entre os mesmos.
Fornecer também uma descrição dos dados.
- Data de apresentação e entrega do projeto: QUINTA-FEIRA 15/06/2017
- Enviar por email : o programa fonte, o executável (se houver), os dados e o relatório do projeto
- Tempo de apresentação: 10 minutos (rigoroso).
- PASSAR NA MINHA SALA PARA ASSINAR A ATA DE ENTREGA DO TRABALHO EM 15/06/2017
- o projeto pode ser realizado com no máximo 3 alunos.