Projeto AM 2018-1

Francisco de A. T. de Carvalho[1]

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

Execute a variante KCM-F-GH do algoritmo KCM-F-H descrito na seção 3.2 do artigo “FAT de Carvalho, EC Simões, LVC Santana, MRP Ferreira, Gaussian Kernel C-Means Hard Clustering Algorithms with Automated Computation of the Width Hyper-Parameters, Pattern Recognition, 79, 370-386, 2018" na tabela de dados completa (complet view, 2100 objetos e 19 variáveis), na tabela shape view (2100 objetos e 9 variaveis) e na tabela RGB view (2100 objetos e 10 variaveis), 100 vezes para obter uma partição em 7 grupos. Em cada caso selecione o melhor resultado segundo a função objetivo. Em cada caso, calcule o índice de Rand corrigido em relação à partição à priori em 7 classes.

Observações:

- No algoritmo 2, pagina 376 da seção 3.2, as distancias entre os objetos e os representantes dos grupos são calculados segundo a equação (21), o vetor de hyperparametros é calculado com a equação (24), a afetação dos objetos os grupos é realizada segundo a equação (21);
- Parametros: numero de grupos $c = 7$; parametro $\gamma = (\frac{1}{\sigma^2})^p$ onde, p é o numero de variaveis e σ^2 é a media entre o 0.1 e o 0.9 quantil de $||\mathbf{x}_l - \mathbf{x}_k||$ $l \neq k$;
- Para o melhor resultado obtido para cada conjunto de dados imprimir: i) o numero de objetos de cada grupo, ii) o vetor de hyperparametros, iii) a partição (para cada grupo, a lista de objetos), iv) 0 indice 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 os 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 acerto 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 gaussiano. Considere a seguinte regra de decisão: afetar o exemplo $\mathbf{x}_k$ à classe ω_l se $P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{7} P(\omega_i|\mathbf{x}_k)$ 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 $(i = 1, \ldots, 7)$ estime $p(\mathbf{x}_k|\omega_i) = p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)$ pelo método da máxima verossimilhança, supondo uma normal multivariada, onde:

- $\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\Sigma}_i \end{pmatrix}, \boldsymbol{\Sigma}_i = diag(\sigma_{i1}^2, \ldots, \sigma_{i2}^2)$
- $p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = (2\pi)^{-\frac{d}{2}} (|\boldsymbol{\Sigma}^{-1}|)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^{tr} \boldsymbol{\Sigma}^{-1} (\mathbf{x}_k - \boldsymbol{\mu}_i)\right\}$
- $\boldsymbol{\mu}_i = \frac{1}{n} \sum_{k=1}^{n} \mathbf{x}_k,$
- $\sigma_{il}^2 = \frac{1}{n} \sum_{k=1}^{K} (x_{kl} - \mu_l)^2 \; (1 \leq l \leq d)$

ii) Usar um classificador bayesiano baseado em k-vizinhos com ponderação local para fazer a classificação dos dados. Treine três classificadores bayesianos baseados em k-vizinhos com ponderação local, um para cada view. Use a distância Euclidiana para definir a vizinhança. Use conjunto de validação para fixar o o número de vizinhos k.

iii) Regra da maximo: afetar o exemplo $\mathbf{x}_k$ a classe ω_j se

$$(1-L)P(\omega_j) + L\max\Big(P_{GAUSS,VIEW1}(\omega_j|\mathbf{x}_k), P_{GAUSS,VIEW2}(\omega_j|\mathbf{x}_k), P_{GAUSS,VIEW3}(\omega_j|\mathbf{x}_k),$$
$$P_{KVIZ,VIEW1}(\omega_j|\mathbf{x}_k), P_{KVIZ,VIEW2}(\omega_j|\mathbf{x}_k), P_{KVIZ,VIEW3}(\omega_j|\mathbf{x}_k)\Big) =$$
$$\max_{r=1}^{7}\Big[(1-L)P(\omega_r) + L\max\Big(P_{GAUSS,VIEW1}(\omega_r|\mathbf{x}_k), P_{GAUSS,VIEW2}(\omega_r|\mathbf{x}_k), P_{GAUSS,VIEW3}(\omega_r|\mathbf{x}_k),$$
$$P_{KVIZ,VIEW1}(\omega_r|\mathbf{x}_k), P_{KVIZ,VIEW2}(\omega_r|\mathbf{x}_k), P_{KVIZ,VIEW3}(\omega_r|\mathbf{x}_k))\Big]$$

com $L = 3$ (três views: complete view, shape view, RGB view)

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: SEXTA-FEIRA 23/11/2018
- 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 31/11/2018
- ALUNOS DE PÓS-GRADUAÇÃO: O PROJETO DEVE SER REALIZADO COM 2 ALUNOS.
- ALUNOS DE GRADUAÇÃO: O PROJETO DEVE SER REALIZADO COM 4 ALUNOS