Projeto AM 2021-2

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

Questão 1

- Considere os dados "Image Segmentation" do site uci machine learning repository (https://archive.ics.uci.edu/ml/datasets/Image+Segmentation). Considere 3 datasets, o primeiro considerando as variáveis 4 a 9 (shape), o segundo considerando as variaveis 10 a 19 (rgb), o terceiro considerando as variaveis 4 a 19 (shape + rgb).
  - Em cada dataset execute o algoritmo "EFCM-LS1" 50 vezes para obter uma partição fuzzy em 7 grupos e selecione o melhor resultado segundo a função objetivo.
  - A descrição do algoritmo "EFCM-LS1" está no artigo: "Sara I.R. Rodríguez, Francisco de A.T.de Carvalho, Fuzzy clustering algorithms with distance metric learning and entropy regularization, Applied Soft Computing, Volume 113, Part A, December 2021, 107922". Implemente a seguinte variante desse algoritmo:
    - Função objetivo: equação (12); Cálculo dos prototipos: conforme Algoritmo 1;
    - Cálculo dos pesos de relevância das variaveis: equação (26); Cálculo do grau de pertinência de um objeto em um grupo: conforme a ultima linha da Tabela 1
  - Para cada dataset e partição fuzzy, calcule o Modified partition coefficient e o Partition entropy. Comente.
  - Para cada dataset e partição fuzzy, produza uma partição crisp em 7 grupos e calcule o índice de Rand corrigido, e a F-measure (adaptada para agrupamento). Comente.
  - Compare as partições crisp em 7 grupos duas a duas com o índice de Rand corrigido, e a F-measure. Comente.
  - Observações:
    - Parametros: $c = 7$; $T = 150$; $\epsilon = 10^{-10}$;
    - Para o melhor resultado imprimir: i) os protótipos ii) a matrix de confusão da partição crisp versus a partição a priori; iii) a matrix de confusão de uma partição crisp versus a outra; iv) a matrix de pesos de relevância das variáveis

Questão 2

- Considere novamente os 3 datasets dos dados "Image Segmentation".

a) Use validação cruzada estratificada “30 $\times$ 10-folds” para avaliar e comparar os 4 classificadores usando a regra do voto majoritário baseados, respectivamente, nos classificadores bayesiano gaussiano, bayesiano baseado em k-vizinhos, bayesiano baseado na janela de Parzen e regressão logística. Quando necessario, retire do conjunto de aprendizagem, um conjunto de validação (20%) para fazer ajuste de hiper-parametros e depois treine o modelo novamente com o conjunto aprendizagem + validação. Use amostragem estratificada.

b) Obtenha uma estimativa pontual e um intervalo de confiança para cada metrica de avaliaąão do classificador (Taxa de erro, precisão, cobertura, F-measure);

c) Usar o Friedman test (teste não parametrico) para comparar os classificadores, e o pós teste (Nemenyi test)

- Considere os seguintes classificadores:

i) Treine um classificador bayesiano gaussiano em cada um dos 3 datasets. Em seguida, treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos gaussianos. Considere a seguinte regra de decisão: afetar o exemplo $\mathbf{x}_k$ à classe ω_l se

$$P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{7} P(\omega_i|\mathbf{x}_k) \text{ com } P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{c} p(\mathbf{x}_k|\omega_r)P(\omega_r)} \; (1 \leq l \leq 7)$$

a) Use a **estimativa de maxima verossimilhança** para $P(\omega_i)$

b) Para cada classe ω_i $(i = 1, 2)$ use a seguinte estimativa de máxima verossimilhança de $p(\mathbf{x}_k|\omega_i) = p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i)$, supondo uma normal multivariada:

$$p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = (2\pi)^{-\frac{d}{2}} (|\boldsymbol{\Sigma}_i^{-1}|)^{\frac{1}{2}} \exp\left\{-\tfrac{1}{2}(\mathbf{x}_k - \boldsymbol{\mu}_i)^{tr}\boldsymbol{\Sigma}_i^{-1}(\mathbf{x}_k - \boldsymbol{\mu}_i)\right\}, \text{ onde}$$

$$\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\Sigma}_i \end{pmatrix}, \boldsymbol{\Sigma}_i = diag(\sigma^2, \ldots, \sigma^2)$$

$$\boldsymbol{\mu}_i = \tfrac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k, \mu_{ij} = \tfrac{1}{n}\sum_{k=1}^{n} x_{kj}$$

$$\sigma^2 = \tfrac{1}{d\times n}\sum_{k=1}^{n} ||\mathbf{x}_k - \boldsymbol{\mu}_i||^2 = \tfrac{1}{d\times n}\sum_{k=1}^{n}\sum_{j=1}^{d}(x_{kj} - \mu_{ij})^2 \; (1 \leq j \leq d)$$

Questão 2

ii) Treine um classificador bayesiano baseados em k-vizinhos em cada um dos 3 datasets. Use a distância Euclidiana para definir a vizinhança. Use conjunto de validação para fixar o o número de vizinhos k. Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos baseados em k-vizinhos.

iii) Treine um classificador bayesiano baseado em janela de Parzen em cada um dos 3 datasets. Use a função de kernel multivariada produto com o mesmo h para todas as dimensões e a função de kernel Gaussiana unidimensional. Use conjunto de validação para fixar o parâmetro h. Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos baseados em janela de Parzen.

iv) Para cada um dos 3 datasets, treine um classificador baseado em regressão logistica para cada classe e use a bordagem “um contra todos’ para classificar os exemplos. Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores baseados em regressão logistica

Observações Finais

- No Relatório deve estar bem claro 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 sucinta dos dados. No relatorio mostrar os detalhes da obtenção dos hiper-parametros do modelo, se houver.
- Data de apresentação e entrega do projeto: **QUARTA-FEIRA 02/02/2022**.
- Colocar no **google classroom**: o programa fonte, o executável (se houver), os dados e o relatório do projeto
- Tempo de apresentação: **15 minutos** para cada equipe (rigoroso), incluindo discussão.
- Apresença de todos os membros de cada equipe é **obrigatória** durante a apresentação;
- Os horarios de apresentação de cada equipe serão divulgados posteriormente.