Projeto AM 2020-2

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 "banknote authentication Data Set" do site uci machine learning repository (https://archive.ics.uci.edu/ml/datasets/banknote+authentication#).
  - Normalize os dados.
  - Execute o algoritmo "Variable-wise kernel fuzzyc-means clustering algorithms with kernelization of the metric" 100 vezes para obter uma partição fuzzy em 2 grupos e selecione o melhor resultado segundo a função objetivo.
  - Para detalhes do algoritmo "Partitioning fuzzy K-medoids clustering algorithms with relevance weight for each dissimilarity matrix estimated locally" veja o artigo: "Marcelo R.P. Ferreira and Franciso de A.T. de Carvalho, Kernel fuzzy c-means with automatic variable weighting, Fuzzy Sets and Systems, 237 (2014), 1-46". Implemente a seguinte variante desse algoritmo:
    - Função objetivo: equações (15), (22) e (23)
    - Cálculo do prototipo: equação (27)
    - Cálculo dos pesos de relevância das variaveis: equaç ao (30)
    - Cálculo do grau de pertinência de um objeto em um grupo: equaç ao (32)
  - Para cada partição fuzzy, calcule o Modified partition coefficient e o Partition entropy. Comente.
  - Para cada partição fuzzy, produza uma partição crisp em 2 grupos e calcule o índice de Rand corrigido, e a F-measure.
  - Observações:
    - Parametros: $c = 2$; $m = \{1.1, 1.6, 2.0\}$; $T = 150$; $\epsilon = 10^{-10}$;
    - Para o melhor resultado imprimir: i) o protótipo ii) a partição crisp (para cada grupo, a lista de objetos), iii) o numero de objetos de cada grupo crisp, iv)o Modified partition coefficient e o Partition entropy v) 0 indice de Rand corrigido, a F-measure e erro de classificação (atribuição).

Questão 2

- Considere novamente os dados "banknote authentication Data Set".
  - a) Use validação cruzada estratificada “10-folds” para avaliar e comparar os classificadores combinados descritos abaixo. Quando necessario, retire do conjunto de aprendizagem, um conjunto de validação (20%) para fazer ajuste de parametros e depois treine o modelo novamente com os conjuntos 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;

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 $P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{2} P(\omega_i|\mathbf{x}_k)$ com $P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{c} p(\mathbf{x}_k|\omega_r)P(\omega_r)}$ $(1 \leq l \leq 2)$
  - 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:

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

    $\boldsymbol{\theta}_i = \begin{pmatrix} \boldsymbol{\mu}_i \\ \boldsymbol{\Sigma}_i \end{pmatrix}$, $\boldsymbol{\Sigma}_i = diag(\sigma_{i1}^2, \ldots, \sigma_{id}^2)$

    $\boldsymbol{\mu}_i = \frac{1}{n}\sum_{k=1}^{n} \mathbf{x}_k$, $\sigma_{ij}^2 = \frac{1}{n}\sum_{k=1}^{n}(x_{kj} - \mu_{ij})^2$ $(1 \leq j \leq d)$

Questão 2

ii) Treine um classificador bayesiano baseados em k-vizinhos. Normalize os dados e use a distância Euclidiana para definir a vizinhança. Use conjunto de validação para fixar o o número de vizinhos k.

iii) Treine um classificador bayesiano baseado em janela de Parzen. Use a função de kernel multivariada produto com um mesmo h para todas as dimensões e a função de kernel Gaussiana unidimensional. Use conjunto de validação para fixar o parâmetro h.

iv) Treine um classificador baseado em regressão logistica com os 4 atributos numericos x_1, x_2, x_3, x_4.

v) Treine um classificador baseado em regressão logistica com os 4 atributos numericos x_1, x_2, x_3, x_4 originais e mais 4 atributos suplementares $x_5 = x_1^2$, $x_6 = \sqrt{x_2}$, $x_7 = \log(x_3)$, $x_8 = \frac{1}{x_4}$. Use regularização. Use conjunto de validação para fixar o parâmetro de regularização.

vi) Treine um classificador usando a regra do voto majoritario usando os classificadores i) a v).

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: TERÇA-FEIRA 09/03/2021.
- 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).
- Apresença de todos os membros de cada equipe é obrigatória durante a apresentação;
- Os horarios de apresentação de cada equipe serão divulgados posteriormente.