Projeto AM 2024-2

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 "SPECTF Heart" do site UCI (http://archive.ics.uci.edu/dataset/96/spectf+heart). Concatene os datasets SPECTF.test e SPECTF.train para formar o data set SPECTF com 267 individuos (linhas) descritos por 45 variaveis (colunas), sendo a primeira coluna a variavel de classe.
  - Execute os algoritmos KFCM-K e KFCM-K-W.1 50 vezes para obter cada um uma partição fuzzy com $c \in \{2, 3, 4, 5\}$. Para cada c selecione o melhor resultado segundo a função objetivo. Para cada c obtenha a correspondente partição crisp a partir da melhor partição fuzzy. Para cada c e partição crisp calcule a silhueta (Sil). Faça o plot Sil $\times c$ para $c \in \{2, 3, 4, 5\}$ e escolha o numero de clusters: $c^* = arg \max_c Sil(c)$.
  - Para cada algoritmo e melhor partição fuzzy com c^*, calcule o Modified partition coefficient. Comente.
  - Para cada algoritmo e partição crisp correspondente a melhor partição fuzzy com c^* , calcule o índice de Rand corrigido. Comente.
  - Parametros: $T = 100$; $\epsilon = 10^{-6}$; $m = 1.1$;
  - Para cada algoritmo e melhor resultado segundo a função objetivo com c^* mostrar: i) os protótipos de cada grupo $(\mathbf{g}_1, \ldots, \mathbf{g}_c)$; ii) o vetor de parametros de largura de cada grupo $(\mathbf{s}_1, \ldots, \mathbf{s}_c)$ iii) a matrix de confusão da partição crisp versus a partição a priori; iv) o plot da função objetivo versus as iterações;
  - Referencia para os algoritmos KFCM-K e KFCM-K-W.1: Gaussian Kernel Fuzzy C-Means with Width Parameter Computation and Regularization, https://doi.org/10.1016/j.patcog.2023.109749

Questão 2

- Considere novamente o dataset "SPECTF" com duas classes a priori.

a) Use validação cruzada estratificada “30 $\times$ 10-folds” para avaliar e comparar os 5 classificadores: i) bayesiano gaussiano, ii) bayesiano baseado em k-vizinhos, iii) bayesiano baseado na janela de Parzen, iv) regressão logística, v) usando a regra do voto majoritário a partir dos 4 primeiros classificadores. Quando necessario, faça validação cruzada 5-folds nos 9 folds restantes para fazer ajuste de hiper-parametros e depois treine o modelo novamente com o conjunto aprendizagem de 9-folds usando os valores selecionados para os hiper-parametros. Use amostragem estratificada.

b) Obtenha uma estimativa pontual e um intervalo de confiança para cada metrica de avaliação do classificadores (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), usando cada uma das métricas

d) Para cada metrica de avaliação, plot a curva de aprendizagem para o classificador bayesiano Gaussiano. Usando amostragem estratificada, use 70% dos dados para treinamento e 30% para teste. Treine o algoritmo com conjuntos de treinamento de 5% a 100% do conjunto original de treinamento, com passo de 5% (usando amostragem estratificada). Comente.

Questão 2

- Considere os seguintes classificadores:
  - i) Treine um classificador bayesiano gaussiano no dataset SPECTF. Considere a seguinte regra de decisão: afetar o exemplo $\mathbf{x}_k$ à classe ω_l $(1 \leq l \leq 2)$ se $P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{2} P(\omega_i|\mathbf{x}_k)$ com $P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{2} p(\mathbf{x}_k|\omega_r)P(\omega_r)}$
  - a) Use a **estimativa de maxima verossimilhança** para $P(\omega_i)$
  - b) Para cada classe ω_i $(1 \leq i \leq 2)$ 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:
    $p(\mathbf{x}_k|\omega_i, \boldsymbol{\theta}_i) = (2\pi)^{-\frac{d}{2}}(|\hat{\boldsymbol{\Sigma}}_i^{-1}|)^{\frac{1}{2}} \exp\left\{-\frac{1}{2}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)^\top \hat{\boldsymbol{\Sigma}}_i^{-1}(\mathbf{x}_k - \hat{\boldsymbol{\mu}}_i)\right\}$,
    onde
    $\hat{\boldsymbol{\Sigma}}_i$ eh a estimativa de MV de $\boldsymbol{\Sigma}_i$; $\hat{\boldsymbol{\mu}}_i$ eh a estimativa de MV de $\boldsymbol{\mu}_i$
  - ii) Treine um classificador bayesiano baseados em k-vizinhos no dataset SPECTF. Considere as distâncias Euclidiana, City-Block e Chebishev para definir a vizinhança. Use validação cruzada para fixar o o número de vizinhos k e a distância.
  - iv) Treine um classificador baseado em regressão logistica no dataset SPECTF. Use validação cruzada para fixar os hiper-parâmetros.
  - v) Treine um classificador a partir dos classificadores, bayesiano, k-vizinhos, janela de parzen, e regressão logistica, baseado na regra do voto majoritário.

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: **SEGUNDA-FEIRA 18/11/2024**.
- Colocar no **google classroom**: o programa fonte, o executável (se houver), os slides da apresentação e o relatório do projeto
- NÃO COLOCAR NO **google classroom** ARQUIVO ZIP: COLOCAR OS ARQUIVOS INDIVIDUAIS
- Tempo de apresentação: **15 minutos** para cada equipe (rigoroso), incluindo discussão.
- A **PRESENCA** e **PARTICIPACAO** de todos os membros de cada equipe é **OBRIGATORIA** durante a apresentação;
- Os horarios de apresentação de cada equipe serão divulgados posteriormente.