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 "Avila Data Set" (Data set e artigo relevante em anexo) do site uci machine learning repository (https://archive.ics.uci.edu/ml/datasets/Avila. Fusione os arquivos avila-tr.txt e avila-ts.txt no arquivo avila.txt, DESCONSIDERANDO a variavel de classe (os rotulos). A partir do arquivo avila.txt produza 3 matrizes de dissililaridade usando a distancia Euclidiana (L2), a distancia de city-block (L1) e a distancia de Chebyshev (Linf).
  - Implemente e execute o algoritmo "VFCMddV" 50 vezes para obter uma partição fuzzy em 12 grupos e selecione o melhor resultado segundo a função objetivo.
  - A descrição do algoritmo "VFCMddV" esta no artigo: "Francisco de A.T. de Carvalho, Filipe M de Melo, Yves Lechevallier, A multi-view relational fuzzy c-medoid vectors clustering algorithm. Neurocomputing, v. 163, p. 115-123, 2015".
  - Calcule o Modified partition coefficient e o Partition entropy. Comente.
  - Produza uma partição crisp em 12 grupos e calcule o índice de Rand corrigido, e a F-measure (adaptada para agrupamento). Comente.
  - Observações:
    - Normalize as matrizes de dissimilaridade conforme descrito no artigo que descreve o algoritmo VFCMddV (pagina 119, coluna 1, terceiro paragrafo);
    - Parametros: $k = 12$; $T = 150$; $\epsilon = 10^{-10}$;
    - Para o melhor resultado imprimir: i) os protótipos ii) a matrix de confusão da partição crisp versus a partição a priori; iv) a matrix de pesos de relevância das matrizes de dissimilaridade.

Questão 2

- Considere novamente "Avila Data Set" do site uci machine learning repository (https://archive.ics.uci.edu/ml/datasets/Avila. Fusione os arquivos avila-tr.txt e avila-ts.txt no arquivo avila.txt, CONSIDERANDO a variavel de classe (os rotulos).

a) Use validação cruzada estratificada "30 $\times$ 10-folds" para avaliar e comparar os 5 classificadores seguintes: bayesiano gaussiano, bayesiano baseado em k-vizinhos, bayesiano baseado na janela de Parzen, regressão logística, e voto majoritário. 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 avaliaação 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) Classificador bayesianos gaussiano: considere a seguinte regra de decisão: afetar o exemplo $\mathbf{x}_k$ à classe ω_l se $P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{12} P(\omega_i|\mathbf{x}_k)$ com $P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{c} p(\mathbf{x}_k|\omega_r)P(\omega_r)}$ $(1 \leq l \leq 12)$

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

b) Para cada classe ω_i $(1 \leq i \leq 12)$ 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

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

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

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

Questão 2

ii) Treine um classificador bayesiano baseados em k-vizinhos. Considere as distancias Euclidiana, City-Block e Chebyshev para definir a vizinhança. Use conjunto de validação para fixar o o número de vizinhos k e a distancia.

iii) Treine um classificador bayesiano baseado em janela de Parzen. 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.

iv) Treine um classificador baseado em regressão logistica para cada classe e use a bordagem “um contra todos’ para classificar os exemplos.

v) Treine um classificador usando a regra do voto majoritário à partir dos 4 classificadores bayesiano

Gaussiano, k-vizinhos, janela de parzen e 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: **SEGUNDA-FEIRA 20/06/2022**.
- Colocar no **google classroom**: o programa fonte, o executável (se houver), o relatório do projeto e os slides da apresentação;
- 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.