Projeto AM 2021-1

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

Questão 1

- Considere os dados "Yeast Data Set" do site uci machine learning repository

  (http://archive.ics.uci.edu/ml/datasets/Yeast).
  - Execute o algoritmo "FCM-DFCV" 100 vezes para obter uma partição fuzzy em 10 grupos e selecione o melhor resultado segundo a função objetivo.
  - Para detalhes do algoritmo "FCM-DFCV" veja o artigo: "Franciso de A.T. de Carvalho, Camilo P. Tenório, Nicomedes L. Cavalcanti Junior, Partitional fuzzy clustering methods based on adaptive quadratic distances, Fuzzy Sets and Systems, 157 (2006), 2833-2857". Implemente a seguinte variante desse algoritmo:
    - Função objetivo: equação (17),
    - Cálculo do prototipo: equação (03)
    - Cálculo dos pesos de relevância das variaveis: equação (19)
    - Cálculo do grau de pertinência de um objeto em um grupo: equação (20)
  - Para cada partição fuzzy, calcule o Modified partition coefficient e o Partition entropy. Comente.
  - Para cada partição fuzzy, produza uma partição crisp em 10 grupos e calcule o índice de Rand corrigido, e a F-measure. Comente.
  - Observações:
    - Parametros: $c = 10$; $m = \{1.1, 1.6, 2.0\}$; $T = 150$; $\epsilon = 10^{-10}$;
    - Para o melhor resultado imprimir: i) os protótipos ii) a matrix de confusão da partição crisp versus a partição a pripri; iii)o Modified partition coefficient e o Partition entropy v) 0 indice de Rand corrigido, a F-measure e erro de classificação.

Questão 2

- Considere novamente os dados "Yeast Data Set".
  a) Use validação cruzada estratificada “5-folds” para avaliar e comparar os classificadores descritos abaixo. 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) Classificador bayesiano gaussiano: considere a seguinte regra de decisão: afetar o exemplo $\mathbf{x}_k$ à classe ω_l se $P(\omega_l|\mathbf{x}_k) = \max_{i=1}^{10} P(\omega_i|\mathbf{x}_k)$ com $P(\omega_i|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_i)P(\omega_i)}{\sum_{r=1}^{c} p(\mathbf{x}_k|\omega_r)P(\omega_r)}$ $(1 \leq l \leq 10)$

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

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

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

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

Questão 2

ii) Treine um classificador bayesiano baseados em k-vizinhos. 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 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 majoritario a partir dos classificadores i) a iv).

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 04/08/2021**.
- 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), 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.