

Projeto AM 2015-2

Francisco de A. T. de Carvalho¹ Cleber Zenchettin²

1 Centro de Informatica-CIn/UFPE
Av. Prof. Luiz Freire, s/n -Cidade Universitaria, CEP 50740-540, Recife-PE, Brasil,
{fatc,cz}@cin.ufpe.br

1) Considere a tabela de dados "Tic-Tac-Toe Endgame Data Set" do site uci machine learning repository (<http://archive.ics.uci.edu/ml/>).

a) Obtenha a matrix de dissimilaridades entre os objetos usando a seguinte função. Seja $\mathbf{x}_i = (x_{i1}, \dots, x_{ij}, \dots, x_{ip})$. Então: $d(\mathbf{x}_i, \mathbf{x}_l) = \sum_{j=1}^p \delta(x_{ij}, x_{lj})$,

$$\text{onde } \delta(x_{ij}, x_{lj}) = \begin{cases} 1 & \text{if } x_{ij} \neq x_{lj} \\ 0 & \text{if } x_{ij} = x_{lj} \end{cases}$$

b) Execute o algoritmo "Single view fuzzy K-set-medoids clustering algorithm" 100 vezes para obter uma partição fuzzy em 2 grupos e selecione o melhor resultado segundo a função objetivo. A partir da partição fuzzy, obtenha a partição hard em 2 grupos e calcule o índice de Rand corrigido. Para detalhes do algoritmo "Single view fuzzy K-set-medoids clustering algorithm" veja a seção 2.1 do artigo:

"F.A.T. de Carvalho, Y. Lechevalier and F.M. Melo, Relational Partitioning Fuzzy Clustering Algorithms Based on Multiple Dissimilarity Matrices, Fuzzy Sets and Systems, 215, 1-28, 2013".

Observações:

- Parametros: $K = 2, m = 2, T = 150, \epsilon = 10^{-10}, q = 2$
- Para o melhor resultado imprimir: i) a partição fuzzy (matrix U), ii) a partição hard (para cada grupo, a lista de objetos), iii) para cada grupo a lista de medoids, iv) 0 índice de Rand corrigido.

- 2) Considere novamente a tabela de dados "Tic-Tac-Toe Endgame Data Set". Os exemplos são rotulados segundo as classes ω_1 : *positivo* e ω_2 : *negativo*. Os dados são descritos por 9 variáveis categóricas, cada uma delas com 3 categorias "x(1)", "o(0)" e "b(-1)". Cada objeto é descrito pelo par (\mathbf{x}, y) , onde $\mathbf{x} = (x_1, \dots, x_9)$, $x_i \in \{1, 0, -1\}$, $i = 1, \dots, 9$ e $y \in \{\omega_1, \omega_2\}$.

- a) Classificador Bayesiano. Classifique os exemplos segundo a seguinte regra de decisão:

afetar o exemplo \mathbf{x} a classe ω_j se $j = \arg \max P(\omega_j | \mathbf{x})$ com


$$P(\omega_j | \mathbf{x}) = \frac{P(\mathbf{x} | \omega_j) P(\omega_j)}{\sum_{j=1}^c P(\mathbf{x} | \omega_j) P(\omega_j)} \text{ onde } P(\omega_j) \text{ é a probabilidade a priori da}$$

classe ω_j e $P(\mathbf{x} | \omega_j) = \prod_{i=1}^d (p_{ij})^{\frac{x_i(x_i+1)}{2}} (q_{ij})^{(1-x_i^2)} (r_{ij})^{\frac{x_i(x_i-1)}{2}}$ é a probabilidade condicional com $p_{ij} = P(x_i = 1 | \omega_j)$, $q_{ij} = P(x_i = 0 | \omega_j)$ e $r_{ij} = P(x_i = -1 | \omega_j)$, $i = 1, \dots, 9$; $j = 1, 2$.

Estime $P(\omega_j)$ e os parâmetros p_{ij} , q_{ij} , r_{ij} pelo método da máxima verossimilhança. Considerando o conjunto de aprendizagem da classe ω_j

$D = \{\mathbf{x}_1, \dots, \mathbf{x}_k, \dots, \mathbf{x}_{n_j}\}$, use como estimativas desses parâmetros:

$$p_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} \frac{x_{ki}(x_{ki}+1)}{2}; q_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} (1 - x_{ki}^2); r_{ij} = \frac{1}{n_j} \sum_{k=1}^{n_j} \frac{x_{ki}(x_{ki}-1)}{2}$$

- b) Estime diretamente $P(\omega_I|\mathbf{x})$ pelo método dos k-vizinhos mais próximos. Use a distância anterior do item 1. Varie o número de vizinhos.
- c) Use a regra da soma para classificar o exemplo \mathbf{x} a partir do cálculo de $P(\omega_I|\mathbf{x})$ obtido pelos classificadores Bayesiano, e k-vizinhos 
- d) Usar MLP e SVM para fazer a classificação dos dados.

Observações

- a) Use validação cruzada estratificada para avaliar e comparar esses classificadores
- b) Obtenha uma estimativa pontual e um intervalo de confiança para a taxa de erro para cada classificador
- c) usar Friedman test (teste não paramétrico) para comparar os classificadores. Usar também o Nemenyi test (pos teste)

Observações Finais

- No Relatório e na saída da ferramenta devem estar bem claros:
 - a) como foi realizada a combinação dos classificadores;
 - b) 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 dos dados.
- Data de apresentação e entrega do projeto: QUINTA-FEIRA 03/12/2015
- Enviar por email : o programa fonte, o executável, os dados e o relatório do projeto
- PASSAR NA MINHA SALA PARA ASSINAR A ATA DE ENTREGA DO TRABALHO EM 01/12/2015
- ALUNOS DE PÓS-GRADUAÇÃO: o projeto pode ser realizado com no máximo 2 alunos.
- ALUNOS DE GRADUAÇÃO: o projeto pode ser realizado com no máximo 4 alunos.