

Projeto AM 2023-1

Francisco de A. T. de Carvalho¹

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 "Image Segmentation" do site uci machine learning repository (<https://archive.ics.uci.edu/ml/datasets/Image+Segmentation>). Considere 3 datasets, o primeiro considerando as variáveis 4 a 9 (shape), o segundo considerando as variáveis 10 a 19 (rgb), o terceiro considerando as variáveis 4 a 19 (shape + rgb). ✓
- Em cada dataset execute o algoritmo FCM com a distância de City-Block 50 vezes para obter uma partição fuzzy em 7 grupos e selecione o melhor resultado segundo a função objetivo. —
- Para cada dataset e partição fuzzy, calcule o Modified partition coefficient e o Partition entropy. Comente. —
- Para cada dataset e partição fuzzy, produza uma partição crisp em 7 grupos e calcule o índice de Rand corrigido, e a F-measure (adaptada para agrupamento). Comente. — ✓
- Compare as partições crisp em 7 grupos duas a duas com o índice de Rand corrigido, e a F-measure. Comente. ●
- Parametros: $c = 7$; $T = 100$; $\epsilon = 10^{-6}$; $m \in \{1.1, 1.6, 2.0\}$; ✓
- Para o melhor resultado imprimir: i) os protótipos ii) a matrix de confusão da partição crisp versus a partição a priori; iii) a matrix de confusão de uma partição crisp versus a outra; quem é a outra?

Questão 2

- Considere novamente os 3 datasets dos dados "Image Segmentation".
- a) Use validação cruzada estratificada "30 × 10-folds" para avaliar e comparar os 4 classificadores usando a regra do voto majoritário baseados, respectivamente, nos classificadores bayesiano gaussiano, bayesiano baseado em k-vizinhos, bayesiano baseado na janela de Parzen e regressão logística. 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 avaliação 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), usando cada uma das métricas
- Considere os seguintes classificadores:
 - i) Treine um classificador bayesiano gaussiano em cada um dos 3 datasets. Em seguida, treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos gaussianos. Considere a seguinte regra de decisão: afetar o exemplo \mathbf{x}_k à classe ω_l se
$$P(\omega_l|\mathbf{x}_k) = \max_{i=1}^7 P(\omega_i|\mathbf{x}_k) \text{ com } P(\omega_l|\mathbf{x}_k) = \frac{p(\mathbf{x}_k|\omega_l)P(\omega_l)}{\sum_{r=1}^c p(\mathbf{x}_k|\omega_r)P(\omega_r)} \quad (1 \leq l \leq 7)$$
 - a) Use a **estimativa de maxima verossimilhança** para $P(\omega_i)$

Questão 2

- b) Para cada classe ω_i ($i = 1, 2$) use a seguinte estimativa de máxima verossimilhança de $p(\mathbf{x}_k|\omega_i) = p(\mathbf{x}_k|\omega_i, \theta_i)$, supondo uma normal multivariada:

$$p(\mathbf{x}_k|\omega_i, \theta_i) = (2\pi)^{-\frac{d}{2}} (|\Sigma_i^{-1}|)^{\frac{1}{2}} \exp \left\{ -\frac{1}{2} (\mathbf{x}_k - \mu_i)^T \Sigma_i^{-1} (\mathbf{x}_k - \mu_i) \right\}, \text{ onde}$$

$$\theta_i = \left(\begin{matrix} \mu_i \\ \Sigma_i \end{matrix} \right), \Sigma_i = \text{diag}(\sigma^2, \dots, \sigma^2); \mu_i = \frac{1}{n} \sum_{k=1}^n \mathbf{x}_k, \mu_{ij} = \frac{1}{n} \sum_{k=1}^n x_{kj}$$

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

- ii) Treine um classificador bayesiano baseados em k-vizinhos em cada um dos 3 datasets. Use a distância Euclidiana para definir a vizinhança. Use conjunto de validação para fixar o o número de vizinhos k . Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos baseados em k-vizinhos.
- iii) Treine um classificador bayesiano baseado em janela de Parzen em cada um dos 3 datasets. 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 . Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores bayesianos baseados em janela de Parzen.
- iv) Para cada um dos 3 datasets, treine um classificador baseado em regressão logística para cada classe e use a bordagem “um contra todos” para classificar os exemplos. Treine um classificador usando a regra do voto majoritário à partir dos 3 classificadores baseados em regressão logística

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 relatório mostrar os detalhes da obtenção dos hiper-parâmetros do modelo, se houver.
- Data de apresentação e entrega do projeto: **QUARTA-FEIRA 14/06/2023**.
- Colocar no **google classroom**: o programa fonte, o executável (se houver), os slides da apresentação e o relatório do projeto
- Tempo de apresentação: **15 minutos** para cada equipe (rigoroso), incluindo discussão.
- Presença e participação de todos os membros de cada equipe é **obrigatória** durante a apresentação;
- Os horários de apresentação de cada equipe serão divulgados posteriormente.