

Projeto

- 1) Considere no Tolbox do MatLab, um conjunto de 300 padrões formando 3 classes, uma classe com 150, uma outra com 100 e a ultima com 50 padrões. Esses padrões são descritos por 2 variaveis quantitativas. Os padrões de cada classe são gerados a partir de distribuições normais bi-variadas segundo os seguintes parâmetros:
 - a. Classe 1: $\mu_1 = 0, \mu_2 = 0, \sigma_1 = 2, \sigma_2 = 1, \sigma_{12} = 1.7, \rho_{12} = 0.85$
 - b. Classe 2: $\mu_1 = 0, \mu_2 = 3, \sigma_1 = 0.5, \sigma_2 = 0.5, \sigma_{12} = 0.0, \rho_{12} = 0.00$
 - c. Classe 3: $\mu_1 = 4, \mu_2 = 3, \sigma_1 = 2, \sigma_2 = 1, \sigma_{12} = -1.7, \rho_{12} = -0.85$

Aplique a esses padrões o algoritmo de particionamento de tipo “nuvens dinâmicas” para dados relacionais (tabela de dissimilaridades) com 3 clusters 100 vezes e selecione o melhor resultado segundo o critério de adequação entre os clusters e seus representantes. Calcule a taxa de erro de classificação global e por classe para o melhor resultado. Esses padrões deverão ser comparados segundo a distancia Euclidiana.

Esse algoritmo tem os seguintes etapas:

- a. Inicialização. Selecionar aleatoriamente os protótipos $g_1, \dots, g_K \in \{x_1, \dots, x_n\}$ dos clusters C_1, \dots, C_K . Formar a partição $P = \{C_1, \dots, C_K\}$ afetando x_i ($i=1, \dots, n$) a classe C_k ($k=1, \dots, K$) se $k = \operatorname{argmin}(1 \leq s \leq K) d(x_i, g_s)$
 - b. Etapa de representação. Para cada classe C_k ($k=1, \dots, K$) encontrar $g_k \in \{x_1, \dots, x_n\}$: $k = \operatorname{argmin}(1 \leq s \leq K) \text{Soma}(i \in C_s) d(x_i, g_s)$
 - c. Etapa de Alocação. Para cada x_i ($i=1, \dots, n$), afetar x_i a classe C_k se $k = \operatorname{argmin}(1 \leq s \leq K) d(x_i, g_s)$
 - d. Critério de parada: se os padrões não mudarem de classe STOP, senão voltar para a etapa b)
- 2) Considere o resultado da clusterização obtida no item 1 através do algoritmo de particionamento. Os padrões estão agrupados em 3 clusters. Considere no TolBox do MatLab (se não existir, implemente) os seguintes classificadores:
 - a. O discriminante amostral supondo mesma matriz de covariâncias;
 - b. Janela de Parzen. Parâmetros a serem considerados: funções de kernel (use a normal e uma outra de sua escolha) e o parâmetro de suavização h .
 - c. Todas as possíveis combinações desses classificadores: produto, soma, mediana.
 - 3) Use validação cruzada estratificada (ou hold-out estratificado) para avaliar e comparar esses classificadores (e as possíveis combinações dos mesmos).

Observações:

- No Relatório e na saída da ferramenta devem estar bem claros: i) como foi realizada a combinação dos classificadores; ii) 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 Entrega do relatório: 21/06/2011
- Enviar por email: o programa fonte, o executável, os dados e o relatório do projeto.
- O projeto deve ser realizado sozinho ou em duplas
- Nota: o projeto vale 60% e a prova 40%